A Method to Detect and Represent Temporal Patterns from Time Series Data and its Application for Analysis of Physiological Data Streams

By

Catherine Inibhunu

A thesis submitted to the
School of Graduate and Postdoctoral Studies in partial
fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

Faculty of Science

University of Ontario Institute of Technology (Ontario Tech University)

Oshawa, Ontario, Canada

March 2020

THESIS EXAMINATION INFORMATION

Submitted by: **Catherine Inibhunu**

Doctor of Philosophy in **Computer Science**

Thesis title: A Method to Detect and Represent Temporal Patterns from Time Series Data and its Application for Analysis of Physiological Data Streams

An oral defense of this thesis took place on March 9th, 2020 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	DR. ALVARO QUEVEDO
Research Supervisor	DR. CAROLYN MCGREGOR AM
Examining Committee Member	DR. RICHARD PAZZI
Examining Committee Member	DR. AMIR RASTPOUR
University Examiner	DR. MARK GREEN
External Examiner	DR. KAAMRAN RAAHEMIFAR

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

ABSTRACT

In critical care, complex systems and sensors continuously monitor patients' physiological features such as heart rate, respiratory rate thus generating significant amounts of data every second. This results to more than 2 million records generated per patient in an hour. It's an immense challenge for anyone trying to utilize this data when making critical decisions about patient care. Temporal abstraction and data mining are two research fields that have tried to synthesize time oriented data to detect hidden relationships that may exist in the data.

Various researchers have looked at techniques for generating abstractions from clinical data. However, the variety and speed of data streams generated often overwhelms current systems which are not designed to handle such data. Other attempts have been to understand the complexity in time series data utilizing mining techniques, however, existing models are not designed to detect temporal relationships that might exist in time series data (Inibhunu & McGregor, 2016).

To address this challenge, this thesis has proposed a method that extends the existing knowledge discovery frameworks to include components for detecting and representing temporal relationships in time series data. The developed method is instantiated within the knowledge discovery component of Artemis, a cloud based platform for processing physiological data streams.

This is a unique approach that utilizes pattern recognition principles to facilitate functions for; (a) temporal representation of time series data with abstractions, (b) temporal pattern

generation and quantification (c) frequent patterns identification and (d) building a classification system.

This method is applied to a neonatal intensive care case study with a motivating problem that discovery of specific patterns from patient data could be crucial for making improved decisions within patient care. Another application is in chronic care to detect temporal relationships in ambulatory patient data before occurrence of an adverse event.

The research premise is that discovery of hidden relationships and patterns in data would be valuable in building a classification system that automatically characterize physiological data streams. Such characterization could aid in detection of new normal and abnormal behaviors in patients who may have life threatening conditions.

Keywords: Temporal Patterns; Temporal Abstractions; Knowledge Discovery; Pattern Recognition; Temporal Data Mining

AUTHOR'S DECLARATION

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University Of Ontario Institute Of Technology to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize the University Of Ontario Institute Of Technology to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

The research work that was performed in this thesis was in compliance with the regulations of Ontario Tech's Research Ethics Board/Animal Care Committee under the following REB Certificate numbers; (HiREB #3859-D) from McMaster Children's Hospital through the Artemis Project, (SRHC REB #0087-1516) from Southlake Regional Health Centre and (UOIT REB #14136) from Ontario Tech University.

CATHERINE INIBHUNU

STATEMENT OF CONTRIBUTIONS

I hereby certify that I am the sole author of this thesis. I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others. Furthermore, I hereby certify that I am the sole source of the creative works and/or inventive knowledge described in this thesis. As part of this dissertation, several publications have been submitted and presented in many international venues as follows;

Submissions under peer review

Journal Submisssion

Inibhunu, C., & McGregor,, C. (2020). A Temporal Pattern Discovery Method and its Application to Physiological Data: A Case Study. *Journal of Medical Information Systems. JMIR*. DOI: 10.2196/preprints.17488.

Conference Submisssions

Inibhunu, C., & McGregor, C. (2020). Identification of Temporal Changes on Patients at Risk of LONS with TPRMine: A Case Study in NICU, *33rd IEEE International Symposium on Computer-Based Medical Systems*. *IEEE CBMS*. Mayo Clinic, Rochester, Minnesota, USA.

Inibhunu, C., & McGregor, C. (2020). Application of TPRMine method for Identification of Temporal Changes on Patients with COPD: A Case Study in Telehealth, *33rd IEEE International Symposium on Computer-Based Medical Systems*. *IEEE CBMS*. Mayo Clinic, Rochester, Minnesota, USA.

Conference Publications

Inibhunu, C., & McGregor, C. (2018). Fusing Dimension Reduction and Classification for Mining Interesting Frequent Patterns in Patients Data. *Machine Learning and Data Mining in Pattern Recognition. MLDM* (pp. 1-15). New York: Lecture Notes in Computer Science, vol 10935. Springer, Cham. ***{Premier Machine Learning Conference, 33% Acceptance Rate}***

Inibhunu, C., Jalali, R., Doyle, I., Gates, A., Madill, J., & McGregor, C. (2019). Adaptive API for Real-Time Streaming Analytics as a Service. *41st EMB.* Berlin: IEEE.

Inibhunu, C., & McGregor, C. (2018). Fusing Dimension Reduction and Classification for Mining Interesting Frequent Patterns in Patients Data. *Machine Learning and Data Mining in Pattern Recognition. MLDM* (pp. 1-15). New York: Lecture Notes in Computer Science, vol 10935. Springer, Cham.

Inibhunu, C., & McGregor, C. (2018). State Based Hidden Markov Models for Temporal Pattern Discovery in Critical Care. *Life Sciences Conference (LSC)* (pp. 77-80). Montreal, Canada: IEEE.

Inibhunu, C., Schauer, A., Redwood, O., Clifford, P., & McGregor, C. (2017). Predicting Hospital Admissions and Emergency Room Visits using Remote Home Monitoring Data. *1st IEEE conference in Life Sciences.* Sidney: IEEE Press.

Inibhunu, C., Schauer, A., Redwood, O., Clifford, P., & McGregor, C. (2017). The impact of Gender, Medical History and Vital Status on Emergency Visits and Hospital Admissions: A Remote Patient Monitoring Case Study. *1st IEEE Conference in Life Sciences.* Sidney: IEEE Press. **{BEST PAPER AWARD}**

Inibhunu, C., & McGregor, C. (2016). Dimension Reduction and Similarity Measures for Temporal Pattern Recognition in Critical Care. *IEEE EMBS.* Orlando, FL.

Inibhunu, C., & McGregor, C. (2016). Machine learning model for temporal pattern recognition. *IEEE EMBS International Student Conference (ISC)*, (pp. 1-4). Ottawa.

Inibhunu, C., & McGregor, C. (2014). *An Overview of Temporal Abstraction in Critical Care*. *CSCBCE2014.*

Inibhunu, C., & McGregor, C. (2014). *Temporal Data Mining in Critical Care an Overview*. *CSCBCE2014.*

Conference Workshops

Inibhunu, C., & McGregor, C. (2017). Towards Temporal Discovery aided by Remote Patient Monitoring Services: A Case Study on ER Visits Factors. *WIML Workshop.* Long Beach, CA. ***Co-hosted with NeurIPS***.

Inibhunu, C., & McGregor, C. (2016). Dimension Reduction and Similarity Measures for Temporal Pattern Recognition in Critical Care. *International Conference on Health Informatics (ICHI). Chicago, IL**. {Doctorial Consortium}***

ACKNOWLEDGEMENT

I would like to give special thanks to Dr. Carolyn McGregor for the encouragement, support, and direction that she continuously provided during my PhD studies. Her insightful suggestions, enthusiastic commendations, and prudent advice has made the completion of this research possible. She provides an example to be emulated and has truly helped me expand the breadth of my research potential with many hours spent on reviewing this document.

I am extremely grateful to Dr. McGregor and the Ontario Tech University for the financial support that has enabled me carry out my passion in Knowledge Discovery and its potential application in Critical Care. I would also like to give special thank you to Dr. Richard Pazzi for the critical review of initial research proposal and providing important feedback which has enabled me incorporate crucial details in this thesis.

I am deeply grateful to my husband Genesis, our children Happy, Bryan and Genesis for their generous comments, continuous moral support, encouragement and acceptance of the long hours I stayed up working even on weekends and during holidays. Your resilience has allowed me to work unimpeded on my passion. I also want to thank my parents Fredrick and Jane for instilling a strong work ethic early in my childhood.

Finally, I have had the opportunity to meet many colleagues and collaborators while working many hours in Health Informatics Research lab, am truly grateful for your continuous encouragements and well wishes, thank you to everyone, it's been a well worth voyage.

I dedicate my dissertation work to my family and friends.

A special feeling of gratitude to my husband Genesis,

our lovely children Happy, Bryan and Genesis and my parents,

your support and encouragement have made me stronger, better and

more fulfilled than I could have ever imagined.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 5

CHAPTER 6

CHAPTER 7

GLOSSARY OF TERMS

Data Streams/ Streaming Data: Sequence of digital encoded data packets transmitted between computers.

Static Data: Data that is already stored in memory with size and structure already known in advance.

Time Series: series of data points that are indexed with time

Time Oriented Data: series of data points that have time as one of the key attributes

Temporal abstractions: High granular representation of time series data

Temporal patterns: Segment of data that occurs frequently in time series data

Knowledge discovery: The process of learning about hidden or unknown information from data

Knowledge Extraction: The process of retrieving hidden or unknown knowledge from data

Knowledge Management: This is a combination of knowledge discovery and knowledge extraction process

Temporal Pattern Recognition: The process of extracting temporal patterns from data

Temporal Data Mining: The process of discovering and retrieving time oriented knowledge from data

Classification System: A system for group objects based on predefined features

Administrative Databases: Data structures that house information about patients details while they receive healthcare services i.e. admission, transfer, or discharge.

Multivariate Time Series: Time series data that representing varying attributes about an object

Physiological Data: signal data that is captured from a human that represent the vital status of a human organ i.e. heart rate, blood pressure, respiratory rate , blood oxygen saturation, etc.

Bayesian Predictive Modelling: The use of Bayes probability theory to predict events using data

Correlation Analysis: The use of statistical measures to understand the similarity between data attributes

Distributed Computation: Running computer functions in different computers or different processes within one computer.

Multidimensional: Data which consists of more than one variable to represent an object, place or person

NICU: Neonatal intensive care unit

Neonatology: Field in medicine that looks at the care of infants and young children

Chronic Care: Field in medicine that looks at the care and services provide to patients with recurring medical conditions.

Frequent Patterns: Set of items or sequences that occur frequently in data

Frequent Pattern Mining: Process of discovering and retrieving frequently occurring patterns

Clinical Decision Support: Process geared to provide health care professionals and physicals information that can be utilized for making clinical decisions.

Chapter 1

1 Introduction

1.1 Overview

This thesis presents a method for knowledge discovery using time series data. For the last 20 years, there has been an explosion of data, from banking to healthcare, in US alone 2,657,700 gigabytes of data are generated every minute (Hale, 2017). In critical care setting billions of data are produced from various complex systems, monitors and sensors that record physiological features such as heart rate, arterial blood pressure averages, respiratory frequency and oxygen saturation from patients everyday (Inibhunu, et al., 2019). Anyone trying to make sense of all this time-series data, also known as temporal data, is faced with difficulties from acquisition, storage, analysis, and knowledge extraction to presenting the knowledge discovered in a way that is meaningful for healthcare providers as they make crucial decisions on the care of patients.

Researchers have tried to process the huge volumes of time-series data and often face many challenges; data collected could be in varying formats, captured in differing speeds, different levels of granularity thus making it very difficult for use by healthcare providers in decision-making. In addition, information gathered from patients data needs to be validated with actual medical knowledge as noted in (Silven, Dojat, & Garbay, 2005) (Combi, Keravnou-Papailiou, & Shahar, 2010) .

Temporal abstraction (TA) is one research area that has been explored with an aim of taking time series data to high levels of granularity that are easier for use in clinical settings (Shahar, 1997), (Hadad, Evin, Drozdowicz, & Chiotti, 2007), (Shabtai, Shahar, & Elovici, 2012). A lot of research has been completed to advance knowledge in TA. This includes the Resume framework in (Shahar, 1997) which derived inference from temporal data and was demonstrated in various clinical domains. This system was extended in (Shabtai, Shahar, & Elovici, 2012) to enhance computation power with distributed workflows and in (Spokoiny & Shahar, 2007) to accommodate interval based abstractions while Combi et al (2012) looked at visual definition and descriptions of TA. Researchers have identified the need to have systems that are able to transcend to continuous streams of time-series data generated every second from monitors and sensors in the critical care setting like Neonatal Intensive Units (NICU) as noted in (McGregor et al 2009, 2010, 2011). Having frameworks that are capable of handling the continuous flow of data streams thereby generating multidimensional abstractions is a key aspect of Artemis, a cloud based architecture for collecting and analyzing multidimensional patient data with application in the NICU (McGregor et al, 2009). Artemis is also the platform within which the case study research in this thesis is demonstrated.

The ability to create meaningful abstractions from time-series clinical data streams has several benefits: support of recommendations for diagnostic and therapeutic treatment; aid in therapy planning to create meaningful context of patients state of health; help in explaining recommendations derived from intelligent systems; represent clinical guidelines; and allow real-time review and aid in generating summaries that could explain trends in patient health status (McGregor et al, 2009, 2010, 2012, 2014).

Reasoning and interpretation of the abstractions creates knowledge. Systems aimed at valid interpretation of temporal data should have the following functionality: ability to determine the boundaries for when abstractions convey meaningful information; capability of understanding inconsistency and uncertainty as noted in (Combi, Keravnou-Papailiou, & Shahar, 2010) ; ability to capture new interpretations based on new abstractions; understand relationships between different abstractions; and hypothesise a truth based on the abstracted relationships. Additionally, reasoning systems should be able to accurately interpret the state of the world being examined at a particular abstraction (Silven, Dojat, & Garbay, 2005).

Temporal data mining (TDM), a more recent field of knowledge extraction has a main objective of reasoning and interpretation of hidden knowledge through mining large temporal data (Shabtai, Shahar, & Elovici, 2012). TDM is an intersection between several disciplines; pattern recognition, data mining, statistics, temporal databases, optimization, visualization, parallel and high performance computing (Combi, Keravnou-Papailiou, & Shahar, 2010).

Specifically, pattern recognition, within TDM is concerned with the algorithmic means of extraction of patterns and enumerated from temporal data (Lin, Williamon, Borne, & DeBarr, 2012). In clinical settings where raw data is at the lowest granularity such as the time of a breath or a valve movement of the heart, there are several challenges on how to identify association and patterns from this data over a given moment in time. Systems that are able to understand the temporal relationships in time-series data would be central to effective knowledge extraction.

In recognition that there are many complex processes that must be unified in the knowledge discovery process from data acquisition, standardization, abstraction, mining and interpretation of discovered knowledge, researchers have proposed varying frameworks for temporal abstraction and data mining such as the work described in (Catley, Smith, McGregor, & Tracy, 2009). The researchers proposed extensions to the data preparation process in CRISP-DM to include modules for temporal abstractions and knowledge management Catley et al (2009).

However, the ability to understand relationships and derive patterns from huge volumes of data, scale to real-time streaming, handle data anomalies, incorporate a domain knowledge base and semantics together with visualization of the extracted data remains a central challenge.

As noted in (Khazaei, McGregor, Eklund, El-Khatib, & Thommandram, 2014), a single patient's physiological features generate thousands of bytes of data every second from bedside monitors and sensors. For example, a Phillips monitor such as an Intellivue MP50 or MP70 can display multiple physiological data. Table 1 provides sample data from a single patient. In a single minute thousands of reading are generated and in an hour this can grow to millions of data.

Table 1. 1 : Sample Physiological Data Generated from a Phillips Intellivue Monitor
Metric: Physiological data that have a single sample every second. Waves: Readings that have multiple readings sampled per second.

Data Types	Physiological Data Elements	Reading per Second	Reading per Minute	Reading per Hour	Reading per 24 Hrs
	Heart Rate	1	60	3600	86400
	Pulse Rate	1	60	3600	86400
	Pulse Rate from Plethsmogram	1	60	3600	86400
	SP02 - Blood oxygen saturation	1	60	3600	86400
	Respiratory Rate	1	60	3600	86400
	non-invasive blood pressure - Systolic	1	60	3600	86400
	non-invasive blood pressure - Diastolic	1	60	3600	86400
	non-invasive blood pressure - Mean	1	60	3600	86400
	Arterial Blood Pressure - Systolic	1	60	3600	86400
	Arterial Blood Pressure - Diastolic	1	60	3600	86400
Metrics	Arterial Blood Pressure - Mean	1	60	3600	86400
	Unspecific ECG wave	128	7680	460800	11059200
	ECG Lead I	128	7680	460800	11059200
	ECG Lead II	128	7680	460800	11059200
	ECG Lead III	128	7680	460800	11059200
	IRW: Impedance Resp Wave	16	960	57600	1382400
Waves	PLETH wave	32	1920	115200	2764800
Total Readings for a single patient		**571**	**34260**	**2055600**	**49334400**

It's a significant challenge to understand hidden relationships or patterns that may exist in such large volumes of time oriented data streams. For example, in a clinical setting, within a given hour, a patient can be given some medication. This medication could lead to changes in the patient's body either positively or negatively both in its pharmacokinetic behaviours, the body's distribution of the medication, and its pharmacodynamic behaviours, the medication's impact on the body. The ability to quantify these changes through identification of sequence of patterns in data collected in the consecutive timeframes before and after the medication is administered has great potential to healthcare providers who could utilize this information as they make critical decisions on care of patients.

The mechanisms to detect, represent and classify the various patterns that may exist in data streams captured continuously from monitors and sensors in critical care is a central research problem being tackled in the health informatics community (Chaovalitwongse w., 2007), (Sacci, Larrizza, Combi, & Bellazzi, 2007), (Batal, Sacchi, Bellazzi, & Hauskrecht, 2009), (Chen, Hong, & Tseng, 2012), (McGregor C. , James, Eklund, & et al., 2013).

Classification is one approach where modelling patient's data can be extremely useful in monitoring and detection of various conditions. Current classification is based on averages, minimum, maximum, standard deviation or down sampling of the patient's data leading to results that could be skewed or inaccurate. Therefore, robust classification methods centred on temporal patterns derived from patient data are needed.

1.2 Research Question

The following research questions are the focus in this thesis;

"Does development of a method to detect and represent changes in patterns that may be exhibited in time series physiological data streams lead to discovery of any hidden relationships that may exist in the underlying data?"

"How can the developed method be instantiated in a big data analytics platform for clinical case studies?

The premise of this thesis is that development of such a method can lead to discovery of patterns and previously unknown patterns relationships that may exist among the many sequences of time series data streams. Those discovered patterns could build a classification system with which to classify any new data streams.

1.3 Research Challenges

This thesis identifies several limitations on existing research that seek to discover patterns from time series data captured from patients; specifically the lack of,

(a) Algorithms for generating temporal abstractions that maintain the temporal nature of the original data. (Stacey and McGregor, 2007; Combi et al, 2010).

(b) Processes to detect, represent, understand relationships and classify temporal patterns from real-time data streams with capabilities to filter and prune irreverent patterns. (Laxman and Sastry, 2006).

(c) Algorithms that combine temporal abstractions and data mining thereby filtering and pruning irreverent abstracts. (Chen 2012, 2013; Alvarez et al, 2013).

(d) Effective methods that are able to scale and handle data from various processes be it static or real-time, distributed or not (McGregor et al, 2009, 2010, 2011, 2013).

(e) Algorithms with clinical input which reduces the adoption and trust of results from mining algorithms, this can be accomplished by incorporating medical ontologies or case base reasoning (Combi, et al, 2011).

(f) Storage capabilities that are able to handle large volumes of data (Robison, 2012),

(g) Visualization techniques that make it easy for one to comprehend the results of knowledge systems and inclusion of user interaction whereby hidden details might not be obvious (Lammarsch et al, 2013; Kamaleswaran, 2016).

1.4 Research Objective

This thesis aims to address the first four challenges discussed in section 1.3 by extending existing knowledge discovery frameworks thereby developing a method to detect and represent relationships that may exist in temporal abstractions (TA) and temporal patterns (TP) derived from time oriented data. This process involves incorporating techniques for

downsampling a data stream and in so doing introducing information about the behaviour of the stream. In clinical care, downsampled streams captured from patients can represent higher abstraction of behaviours in the physiology of patients. Providing such abstractions as input to mining algorithms can lead to discovery of unknown relationships among physiological time oriented data. Consequently, the discovered relationships could lead to detection of onset of conditions and aid in classifying abnormal or normal behaviours in patients. This process could be vital to clinicians and health care providers as they make critical decisions on best care for patients.

Currently, bedside monitors generates thousands of records per second as shown in table 1.1, i.e. a single electrocardiogram wave contains about 7680 readings a second, in a 60 minute period, this can result to about 460,800 readings. Trying to analyse this data to understand any temporal patterns that may exist is not a trivial process and the current methods to process such data become like a black box to end users. Instead, manual annotations from monitors are recorded every 30-60 minutes (McGregor et al., 2012). This process is ineffective for the detection of life threatening conditions such as sepsis where early discovery can lead to prompt treatment. Therefore there is a need to develop a method that can detect and represent any temporal relationships and patterns from time oriented data while maintaining the temporal nature of the underlying data as well as ability to process multiple time series data that can be captured from diverse clinical data domains.

This thesis achieves this goal by developing a method that extends the existing knowledge discovery frameworks to include components for detecting and representing temporal

relationships in time series data. Instantiation of the developed method is completed within a big data analytics platform and applied to the analysis of streaming patient data. This process leads to research contributions in three research domains as follows;

(a) Computer Science: Extension of frameworks for Knowledge representation and discovery using probabilistic principles in machine learning for data driven pattern recognition, dimension reduction and data mining proposing new approaches in the integrated use of frequent pattern mining and classification;

(b) Health Informatics: Advancement in frameworks for real-time knowledge discovery using physiological data streams and classification of events and episodes in a clinical context; and

(c) Medicine:

    I. Application in neonatology: a case study application for detection of conditions such as neonatal Sepsis using temporal patterns generated using physiological data streams from patients in NICU.

    II. Application in chronic care of the elderly: a case study application for understanding hidden relationships on key factors attributing to lengthy hospitalization and multiple emergency room visits on patients participating in a remote patient monitoring (RPM) program.

1.5 Research Contribution

This thesis proposes a method for knowledge discovery that provides research contributions to computer science, health informatics and medicine as follows:

Contribution to Computer Science: The thesis focus is on extending the Cross Industry Standard in Data Mining (CRISP-DM) framework (Shearer, 2009). Historically CRISP-DM model consists of 6 components; 1) business understanding, 2) data understanding, 3) Data Preparation 4) Modelling, 5) Evaluation and 6) Deployment. However this framework does not integrate components for data reduction, generation of temporal abstractions or discovery of temporal relations and patterns in data preparation phase before passing the data to the modelling phase. In addition, the modelling phase does not provide clear guidelines on how mining frequent patterns can be performed when dealing with real time data streams. This thesis makes this contribution by;

a) Generating temporal abstractions using principles in (Catley et al, 2009) and discovery of relationships among abstractions using dimension reduction and similarity principles.
b) Deriving temporal patterns from generated abstractions by exploring principles in pattern recognition techniques.
c) Identification of Frequent Patterns: Selection of frequent patterns utilizing the principles described in association rule mining shown in (Aggrawal, 1996) and event sequence mining in (Li et al, 2001). Integration of mathematical modelling will be explored to ensure the temporal nature of the underlying patterns is maintained

(Povinelli, 1999), (Wang, Liu, Shw, Nahavandi, & Kouzani , 2013). In addition, the thesis explores incorporation of clinical ontologies thereby generating temporal patterns valid in a clinical setting.

d) Classification model: The discovered patterns will be used to build a rule based classification system to generate a new hypothesis on a given dataset. The thesis investigates a classification technique that is able to scale to changes in real-time data streams and incorporation of temporal data storage.

Health Informatics Contribution: This thesis proposes a method utilizing the contribution in computer science to build components that can support construction of a classification model that can handle time series physiological data streams. This involves use of pattern recognition methods for classifying normal and abnormal patterns from real-time-physiological data streams captured from bedside monitors within the case study context of the NICU or data captured from patients participate in the RPM with respect to chronic care context. The discovered patterns could describe the state of a patient at any given moment of time. As a result, the patterns can be used for detection of conditions not obvious to clinicians about a critically ill patient.

Contribution to Medicine: This thesis utilizes two case studies as follows;

First Case Study: Application to derive patterns and understand relationships exhibited in physiological data collected from patients in NICU. The thesis premise is that certain patterns may exist among patients with similar conditions such as Late Onset Neonatal Sepsis (LONS) and therefore if a classification system can correctly identify those unique patterns, then it's

possible to characterize such a condition on data that exhibit similar patterns in the future or in detection of onset of such conditions.

Second Case Study: Application in chronic care of the elderly for understanding risk factors contributing to lengthy hospitalization (HA) or multiple ER Visits and understanding temporal relationships and patterns that maybe exhibited in data before an adverse event like HA or ER Visit. The thesis premise is that understanding temporal relationships in data could be vital when making decisions on care of such patients leading to reduced hospitalization, better management of the exponential costs associated with hospitalization and better outcomes for patients.

With this contribution, this thesis addresses three of the limitations identified in section 1.3, namely;

- Efficient methods for generating temporal abstractions that maintain the temporal nature of the original data. (Stacey and McGregor, 2009; Combi et al, 2010).
- Processes to detect, represent, understand relationships and classify temporal patterns from real-time data streams with thresholds to filter and prune irreverent patterns. (Laxman and Sastry, 2006; Chen 2012, 2013; Alvarez et al, 2013).
- Algorithms have clinical context therefore increasing the adoption and trustworthiness of results from mining algorithms, by incorporating medical ontologies or case base reasoning (Combi, et al, 2011).

A detailed discussion on the proposed research contributions are presented in Chapter 5.

1.6 Research Methods

To facilitate the proposed research contributions, the Constructive Research method in (Kasanen, Lukka, & Siitonen, 1993) has been adopted in this thesis using similar principles of computing and information technology described in (McGregor, 2018). In this thesis, constructive research is adopted in order to have a systemic approach to generate knowledge. This is accomplished first by identifying a practical problem and then understanding the current research and challenges pertaining to the problem. This process then allows construction of innovative methods, modules or frameworks that seeks to address the identified problem followed by a demonstration and evaluation of the methods developed to address the problem. Constructive research method has been broadly utilized in computer science, operation analysis, mathematics and clinical medicine (Kasanen, Lukka, & Siitonen, 1993) . To accomplish this process, this thesis follows a phased approach as follows.

First, this research identified the need to find patterns in physiological data captured from diverse cohort of patients as a practical research problem in knowledge discovery using data. To this respect, it is necessary to have a clear understanding of the current research and challenges faced by researchers in knowledge discovery using temporal abstraction and data mining. As such, a comprehensive literature review was completed and presented in chapter 2. This was the backbone guiding the research forward as gaps in existing systems were identified leading to clear problem space for future research on knowledge discovery utilizing time oriented data in clinical care.

After identification of the practical research problem it was clear that this research aims to understand patterns exhibited in time series datasets generated from patients physiological features. Therefore, a comprehensive review was completed to understand the current techniques and algorithms utilized for pattern recognition and presented in chapter 3.

In continuing with the work carried out by (McGregor et al 2009) on clinical decision support in NICU, this research aims to have a clear impact in the Artemis knowledge discovery component, therefore it is important to understand neonatology.

Within this respect, a literature review in neonatology has been completed to understand the data sets collected from patients, the types of research carried out in this field utilizing this data and the existing challenges faced by researchers especially on detection of life threatening conditions. This review is present in chapter 4.

After the comprehensive literature reviews, this research was able to fully identify and define the underlying problem leading to formulation of key research contributions, the approach and plan for fulfilling the research. As a result, a comprehensive methodology is represented to address this research need and is presented in chapter 5.

Demonstration of the developed methodology and its application to clinical care is presented in chapter 6 while the implementation and evaluation of the designed components to support the developed methodology is presented in chapter 7. A detailed tabulation of the phases completed in the adopted research method is presented in table 1.2.

Phase	Constructive Research	Temporal Pattern Recognition Constructive Research
Phase 1	Find a practically relevant problem which has research potential	Currently there are multiple devices that capture a patient physiology and generate millions of records every hour, does development of a method to detect and represent patterns that may exist in such data streams lead to early detection of life threating conditions about patients in critical care?
Phase 2	Obtain a general and comprehensive understanding of the topic	Understanding current state of research in systems that enable knowledge discovery using physiological data generated from patients and their application in clinical domains. Understand the methods for pattern recognition from time series data. Understand the current state of research in clinical care in particular in neonatology and the type of data utilized for clinical oriented research.
Phase 3	Innovate (i.e., construct a solution idea)	**Contribution to computer science:** Extends the CRISP-TDMn model by introducing components with functions for understanding relationships among abstractions, generating temporal patterns, mining frequent occurring patterns and building a classification system. **Contribution to Health Informatics**: Extends the patented multi-dimensional temporal data mining framework presented by (McGregor, 2011) and integrating these enhancements to Artemis System (McGregor et al, 2013) for understanding temporal relationships in abstractions and generating temporal patterns. **Contribution to Medicine:** (a) Understand temporal patterns on patients scored through the LONs detection algorithms in McGregor, Catley, & James (2012). (b) Understand the temporal patterns among a select set of variables before, during and after a hospital admission or ER visit.
Phase 4	Demonstrate the solutions feasibility	This research will utilize data collected from two clinical settings; (a) data captured from McMaster Children's Hospital through the Artemis Project under ethics approval (HiREB 3859-D), and (b) data captured from patients participating in a remote patient monitoring program through a collaboration with Alaya Care, We Care, Southlake and UOIT under OCE Grant (MIS#23823) and ethics approval from both Southlake Regional Health Centre (SRHC REB #0087-1516) and the University Of Ontario Institute Of Technology (UOIT REB #14136).
Phase 5	Show theoretical connections and research contribution of the solution concept	**First Case Study:** Using physiological data streamed from monitors at the bedside in NICU; how effective is the method at enabling the generation of temporal patterns, identification of frequent patterns and building a classification. Additionally, how effective is the method in identifying normal and abnormal behavior in data from differing neonatal patients who may be diagnosed with a condition such as neonatal sepsis. **Second case study:** Using data captured from elderly patients, how effective is the method at identifying temporal patterns which could be associated with an adverse event.
Phase 6	Examine scope of applicability of the solution	This system is capable of discovering temporal patterns from multiple domains in clinical settings as well as outside the healthcare domain.

Table 1. 2 : Constructive Research Method for Detection and Representation of Temporal Patterns

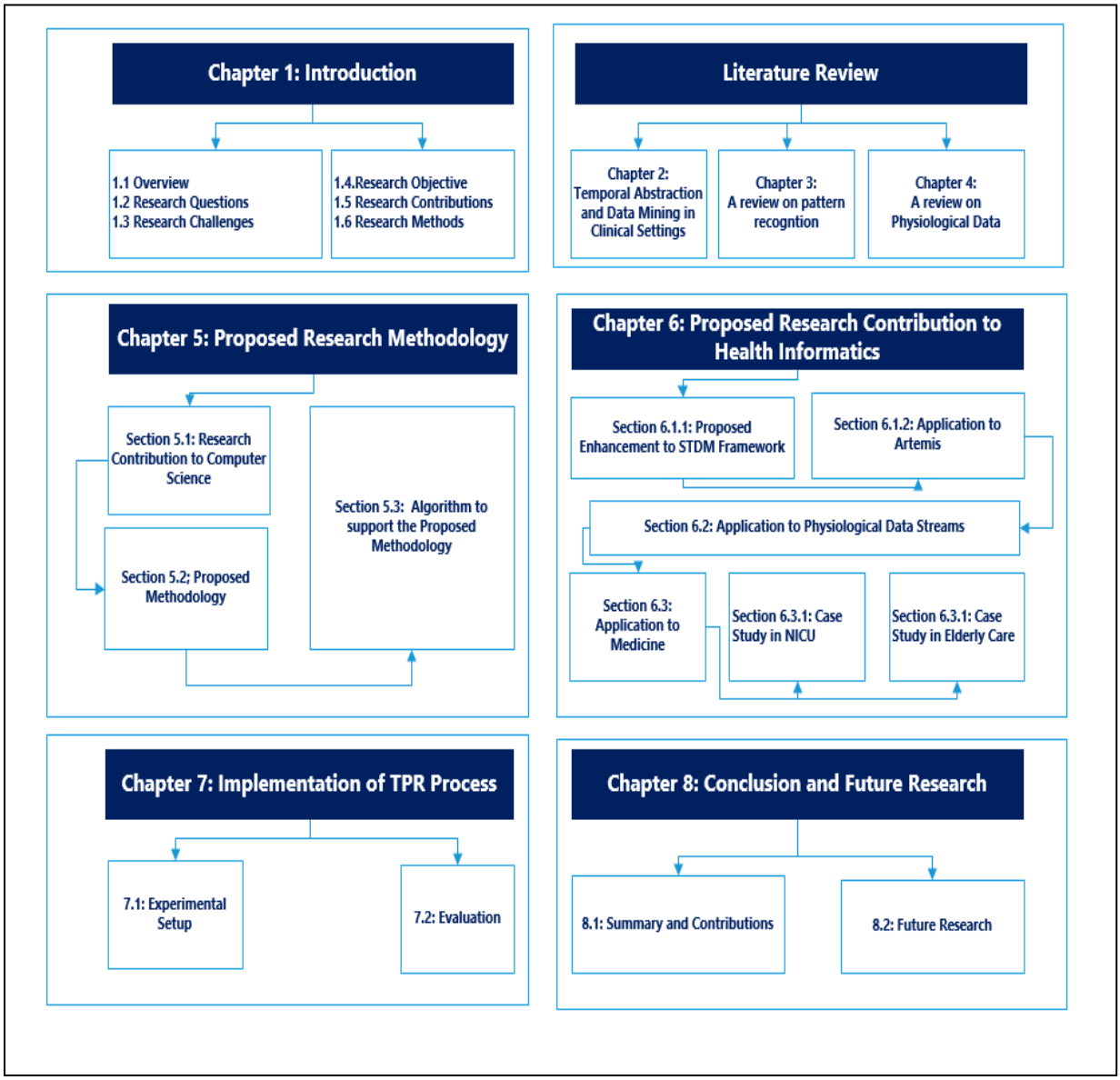
1.7 Thesis Outline

In this chapter, a brief overview on knowledge discovery from massive data streams generated every second from complex medical systems including the various challenges encountered by researchers working in this field was discussed in section 1. Section 1.2 provides the problem statement. Section 1.3 provides the limitations of existing systems, section 1.4 details the research objective, section 1.5 outlines the thesis contributions in computer science, health informatics and field of medicine to tackle some of the challenges outlined in section 1.4 and section 1.6 details the research method adopted for successful completion of the proposed research contributions.

In conclusion, this research aims to provide an innovative method to detect and understand temporal relationships that may exist in time series data generated from patient physiological features such as heart rate, respiratory rate, blood pressure and blood oxygen saturation. Understanding temporal relationships from patient data and then building a classification system can be applied to many clinical domains. This thesis investigates the application of the method developed in analysis of patients in NICU in order to understand patterns which may exist before, during and after a life threatening condition such as neonatal infections. Another application is in provision of care to an aging population using remote patient monitoring programs to understand factors contributing to multiple emergency room visits and lengthy hospitalizations on patients with chronic obstructive pulmonary disease (COPD) and heart failure (HF).

The rest of the document is structured as follows: chapter 2 discusses a literature review on temporal abstraction and mining systems in clinical settings. Chapter 3 provides a review of pattern recognition methods. Chapter 4 provides an overview of neonatology, the data types collected from patients and conditions affecting preterm babies. Chapter 5 introduces the proposed method and details of the research contribution in computer science. Chapter 6 details application of the proposed method in contribution to health informatics and demonstrated using two clinical case studies in contribution to medicine. Chapter 7 details the experimental setup and evaluation of the proposed framework against existing frameworks. The conclusion and future works are presented in chapter 8. A pictorial flow of the entire thesis is presented in Figure 1.1.

Figure 1. 1 : Pictorial Representation of Key Thesis Chapters



Chapter 2

2 Temporal Abstraction and Data Mining in Clinical Settings

This thesis aims to address the research questions in chapter 1, in order to accomplish this process, it is necessary to understand the current research in knowledge discovery utilizing patient data. Temporal abstraction and data mining are two research fields where there has been a focus on synthesising time-oriented data. This chapter presents a literature review on temporal abstraction and temporal data mining and applications in clinical settings. The focus is on review of various methodologies applied for knowledge retrieval from raw patient data, the data types processed, techniques used in temporal abstractions and the algorithms used for temporal data mining. The review also looks at varying medical environments where temporal abstraction has been applied. This process provided a methodical way to expose open research areas in temporal abstraction and data mining research fields and their application in clinical settings.

To facilitate this review, papers were retrieved through the use of a number of keywords such as “patient monitoring”, “real-time mining”, “critical care monitoring”, “knowledge-retrieval in critical care”, “mining time-series data”, “temporal abstraction”, “temporal data mining” and combination of these terms. Artificial Intelligence in Medicine, ACM, IEEE, PubMed databases were examined. Emphasis was given to papers published since the last comprehensive review in temporal abstraction completed in (Stacey & McGregor, 2007).

More than 110 papers where reviewed and the focus was in the following areas: temporal abstraction, a comparable review on the types and techniques in the abstraction process, clinical application for temporal abstraction and knowledge retrieval, research direction on temporal abstraction and data mining, identifying shortcomings of temporal abstraction and temporal data mining. The chapter concludes with a summary on key findings that led to the formation of the research questions of this thesis.

2.1 Temporal Data Abstraction

In varying clinical settings, extensive research work has been published in the area of temporal abstraction. Starting from RESUME, a system introduced as capable of deriving inference from temporal data and evaluated in various clinical domains (Shahar, 1997). A framework for capturing temporal abstractions on data captured from hepatitis patients was proposed where researchers developed algorithms that were able to estimate the stage of liver fibrosis (Ho, Kawasaki, Le, & et al., 2003). Abstraction of multi-level temporal data was attempted in (Silven, Dojat, & Garbay, 2005) with an aim of extracting sequence patterns from monitored medical data. They incorporated the use of ontology for interpretation of the data.

A system for assessing the quality of hemodialysis series given to patients was introduced in (Bellazi , Larizza, & Magni, 2005) where temporal data mining and analysis techniques were applied to gain more insights and potentially discover unsatisfactory clinical results. The researchers analysed 5800 dialysis sessions from 43 patients who were monitored in a period of 19 months, the data was captured at each dialysis session. The researchers claim that the

proposed method proved to be a good tool for discovery of knowledge in time series data. K-Medoid clustering algorithm was introduced in (Altiparmak, Ferhatosmanoglu, Erdal, & Trost, 2006) to capture the high dimensions of clinical trials by detecting blood substances that had a strong correlation and clinicians verified the resulting clusters. They stated that they were able to identify a small set of analyses that was effective for modeling the state of normal health. However, as the frequency of the data utilized is only at a specific point, there is a lot of data not captured which if utilized could provide a complete picture of a hemodialysis patient.

A framework for generating minimal datasets for prediction of the risk of developing heparin-induced thrombocytopenia is described in (Batal, Valizadegan, & Cooper, 2012). The researchers relied on Resume in (Shahar, 1997) and incorporated temporal logic for defining temporal interactions in multiple time series data. They adopted the frequent pattern-mining algorithm and introduced the Mining Predictive Temporal Patterns (MPTP) framework for effectively filtering of false temporal patterns.

A framework for visual specification and defining queries of clinical databases was proposed in (Combi, Pozzi, & Rossato, 2012) and tested using temporal data from hemodialysis patients. The researchers aimed at providing a framework where clinicians could do consistent temporal abstractions of different granularity. Based on a specific clinical domain, one could tailor the abstractions to different levels of detail. In trying to understand the trends of temporal data, another system was proposed that was aimed at management of temporal patterns and granularities derived from the temporal nature of the data (Combi

& Oliboni, 2012). The researcher's claim that current systems are only able to deal with one patient's temporal data at a time, their work aimed at being able to query the entire static temporal database comprised of different levels of granularity.

Scalability is another problem in temporal abstraction and systems are faced with having to process large amounts of raw data. In an attempt to deal with increasing volumes of time series data, the work in (Shabtai, Shahar, & Elovici, 2012) proposed the distributed knowledge based temporal abstraction (DKBTA) architecture that they claim could efficiently distribute abstractions processed from parallel computation nodes. The system assign subtasks to nodes which all perform abstractions of data in parallel thereby increasing the amount of data processed in a single unit of time.

After abstractions are generated, several techniques are proposed for identifying temporal patterns; yet, significant challenges exist in trying to understand relevant or irrelevant patterns. MEMURY was proposed in (Lammarsch, Aigner, Bertone, & et al., 2014), a procedure for finding frequent patterns using temporal relationships and effectively filtering out the unimportant patterns. Frequent pattern mining is also adopted by (Alvarez, Felix, & Carinena, 2013) who introduced ASTPMiner an algorithm that incorporated a seeding technique for reduction of the search space and the number of frequent patterns generated. Two models were developed in (Minne, Ludikhuize, Jonge, & al, 2011) by first collecting temporal data and deriving patterns from ICU patients and then incorporate logistic regression for predicting the probability of survival rates of patients with major health complications.

In the neonatal setting, a system for generating summaries based on data recorded by nurses in various shifts was developed (Hunter, Freer, & Gatt, 2012). In recognition that mechanisms for processing data continuously generated from monitors are needed, (Thommandram, Pugh, Eklund, McGregor, & James, 2013) introduced algorithms that process physiological data streams collected live from infants in NICU for classifying neonatal spells using real-time temporal analysis. These algorithms even though effective at creating temporal rules in a specific neonatal condition, they are not able to explain the temporal relationships among physiological variables. To address this problem, (Inibhunu & McGregor, 2016) introduced models for understanding relationships in time oriented data and it's potential application in critical care.

2.2 Comparative review of Temporal Abstraction Systems

In this section, a careful review of what is incorporated in systems that process temporal abstractions and the techniques used are evaluated.

Techniques in Temporal Abstraction

Various techniques are proposed for the management of temporal abstractions; RESUME system proposed in (Shahar, 1997) incorporated domain ontology, an advancement to the system was presented in (Boaz, 2005) where a distributed system acted as a mediator in a distributed framework. Multi-level temporal abstractions were proposed where first temporal abstractions were presented by enhancing the work in (Hoppner & Klawonn, 2002) which looked at dependencies in multivariate time series and then finding temporal

relationships (Bellazi, Larizza, & Magni, 2005). Adaptive abstractions using ontologies was proposed by (Silven, Dojat, & Garbay, 2005). The complexities of dealing with the temporal nature have been identified by various researchers who have proposed techniques for capturing a subset of data; dimensional vectors were proposed in (Chaovalitwongse w., 2007) Sliding windows were attempted in (Sacci, Larrizza, Combi, & Bellazzi, 2007), (Denai, King, Ross, & Mahfouf, 2007), (Batal, Sacchi, Bellazzi, & Hauskrecht, 2009), a (H., 2014) (Chen, Hong, & Tseng, 2012) (Alvarez, Felix, & Carinena, 2013).

Maintaining validity of earlier generated temporal abstractions in (Spokoiny & Shahar, 2008) attempts an incremental approach. Others have used mathematical models, such as the polynomial shape space representation in (Fuchs, Gruber, Pree, & Sick, 2010), varying granularity in (Combi & Oliboni, 2012) and process mining of temporal pathways is performed in (Huang, Lu, & Duan, 2012). Natural language process methods are also explored in (Hunter et al, 2012) where shift summaries are introduced in their BT-Nurse framework. Others have used Fuzzy logic as described in (Campos, Juarez, Palma, Marin, & Palacios, 2011).

The use of electronic health records for mining disease trajectories is attempted in (Beck, et al., 2016) where researchers utilized clinical data for uncovering temporal patterns and predicting mortality in patients with sepsis. Using similar patient datasets, a recent study in (Campbell, Bass, & Masino, 2020) describes a method for identifying temporal patterns around an initial diagnosis of pediatric asthma. The method was applied to datasets collected from patients who had received care in a children's hospital in Philadelphia. Other works have attempted detection of patterns within a certain temporal range utilizing timestamps

datasets (Titarenko, Titarenko, Aivaliotis, & Palczewski, 2019). Although the methods demonstrated are effective at processing select static data, it is not clear how they can scale to identifying temporal patterns in high frequency real time data streams.

Additionally, the complexity in a clinical setting makes it quite difficult to handle real time physiological data streams which is continuously generated from monitors, sensors and various other complex systems. Within the context of the review for this research, out of 110 papers, 10 are highlighted in Table 2.1 that have attempted to process real-time clinical data, handling multi-dimensional datasets captured from distributed systems and generating temporal abstractions.

Table 2. 1 : An overview of Various Techniques used in Temporal Abstraction

Author/Year	Summary	Temporal Abstraction Technique	Abstraction Type
Spokoiny, Shahar, 2008	Aim at handling the redundancy of earlier generated abstracts.	Domain independent technique of incremental temporal abstractions and maintaining the validity of earlier generated abstractions	Complex
Garbay et al 2011	In this paper they try to solve the problem identified in their earlier work on dealing with ambiguity and the complexity of interpretations	Hidden Markov Model (Bayesian Network)	Complex
Shahar et al, 2012	Research proposes improvement of the KBTA method to enable scalability in handling large volumes of static multivariate time oriented data	Parallel computational nodes that handle data in a distributed architecture	Complex
Combi and Oliboni, 2012	A system that allows for visually defining and composing temporal abstractions	A visual query language for visual description of metaphors that are used for temporal abstraction	Complex
Álvarez et al, 2013	Mining frequent patterns where a predefined seed id given by the user to guide the search space	Sliding window with seed for directed search space	Complex
Gozt et al, 2013	The researchers advanced the ICDA system by adding functionality that allow for ad hoc temporal constrained based queries	Visual analytics that combine data mining with visualization technique	Simple
Lammarsch et al, 2014	Utilize th SAPPERLOT procedure for visualization of patterns explored	Aim at reduction of unimportant patterns and keeping the important ones, by omitting repetitive patterns	Complex
(Beck et al, 2016)	Using electronic health data for mining temporal trajectories	Model based on stratified trajectory	Complex
(Titarenko et al, 2019)	Temporal event mining	Temporal reasoning using probabilities and vectorisation	Complex
Campbell et al, 2020	Identifying temporal patterns around an initial diagnosis of pediatric asthma.	Apriori-based sequential pattern mining using equivalence classes	Complex

Types of Temporal Abstractions

There exists different types of abstractions, some are complex and others are simple. In (Silven, Dojat, & Garbay, 2005), (Shahar, 1997) different types of abstractions can be found. The authors distinguish three types of abstractions; time based (horizontal type), parameter based (vertical type) and linking parameter and time (trend/oblique type). Based on the type of abstraction, different mechanisms for abstractions are considered. A more systematic categorization of abstraction is presented in (Combi , Keravnou-Papailiou, & Shahar, 2010) where two main abstraction types are categorized; Simple abstractions which attribute the time aspect of the data in current moment (the now state) and are referred to as *atemporal* abstractions. Complex abstractions are the other type, these involve varied user-defined approaches on how time is captured and the knowledge derived from the abstracts created, these are referred to as *temporal* abstractions. Table 2.2 provides a detailed overview of these different types of abstractions. We have utilized these 2 types of abstractions as an evaluation criteria of the abstraction features in Table 2.3 and data mining systems highlighted in Table 2.4.

Table 2. 2 : Types of Temporal Abstractions (Combi et al, 2010)

(Simple) Atemporal Types	
Qualitative Abstraction	Converting numerical expressions to qualitative expressions
Generalization Abstraction	Mapping instances into classes
Definitional Abstraction	Mapping across different conceptual categories
(Complex) Temporal Types	
Merge Abstraction	Deriving maximal intervals where properties for concatenation on groups of time-stamped data with those properties
Persistence Abstraction	Applying default persistence rules to protect maximal intervals for some property, both backwards and forwards in time and possibly on the basis of a single datum
Trend Abstraction	Deriving significant changes and rates of change in the progression of some parameter
Periodic Abstraction	Deriving repetitive occurrences, with some regularity in the pattern of repetition

Temporal abstractions of medical data involve reasoning about the underlying medical context. This requires handling data of multiple granularities which may exhibit relationships of temporal nature and concepts of semantics. The work in (Combi , Keravnou-Papailiou, & Shahar, 2010) indicates that abstractions are very different from statistical analysis as the derivation of temporal abstraction is depended on contextual knowledge of the underlying data.

Mining Temporal Abstractions

Temporal abstraction systems generate high level granularity of data with an aim of providing meaningful information about the underlying raw data. However, the creation of these abstractions is a difficult task given one has to handle data of varying formats captured in complex systems some of which are static, real-time, stand alone or in distributed frameworks. Another challenge is in the process of generating temporal abstractions where huge volumes of abstractions are generated. In an attempt to select those abstracts deemed as relevant for a certain domain being examined, association rule mining has been applied in (Spokoiny & Shahar, 2008), (Batal, Sacchi, Bellazzi, & Hauskrecht, 2009), (Sacci, Larrizza, Combi, & Bellazzi, 2007), (Alvarez, Felix, & Carinena, 2013)), (Lammarsch, Aigner, Bertone, & et al., 2014). However, the current approaches do not provide a mechanism to understand temporal relations in time oriented data streams. Additionally most of these techniques are applied to static data sets therefore not equipped to handle the dynamic nature of real-time data streams.

Frequent pattern mining has been applied in (Batal, Sacchi, Bellazzi, & Hauskrecht, 2009) and (Ali, Elhelw, Atallah, & et al., 2008), however, these approaches do not provide clear strategies on management of exponential patterns generated or how these can be applied to real-time data streams.

Machine learning approaches have been used in (Kamsu-Foguem, Tchuente-Foguem, Allart, & et al., 2012), (Yet, Perkins, Fenton, & et al., 2014), (Borchania, Bielzaa, Torob, & et al., 2013). Other approaches have used various types of clustering (Denai, King, Ross, &

Mahfouf, 2007) where fuzzy clusters were used, the K-Medoid in (Altiparmak, Ferhatosmanoglu, Erdal, & Trost, 2006) and outlier detections in (Lara, Moreno, Perez , & et al., 2008). Bayesian classifiers approach is used in (Borchania, Bielzaa, Torob, & et al., 2013). Ensemble models which combine different algorithms such as fuzzy logic and clusters are proposed in (Haghighi, Gilick, Krishnaswamy, & et al., 2010), Rules generation in (Portela , Santos, Silva, & et al., 2013), (Minne, Ludikhuize, Jonge, & al, 2011), Neuro networks in (Huang, Juarez, Duan, & Li, 2014).

Probabilistic approaches have also been used in various proposed frameworks, (Silven, Dojat, & Garbay, 2005), (Fuchs, Gruber, Pree, & Sick, 2010), (Amate, Forbes, Fontecave-Jallon, & et al., 2011), (Batal, Valizadegan, & Cooper, 2012) and (Heidjen & Lucas, 2013). Biomedical time-series processing is proposed in (Wang, Liu, Shw, Nahavandi, & Kouzani , 2013), where a process that groups biomedical time series data collections are grouped utilising bag of words emulating a natural language processing technique that uses probabilistic latent semantic analysis for checking similarities in a time series.

Frameworks to support the methods for mining temporal abstractions follow the knowledge discovery process detailed in Cross Industry Standard in Data Mining (CRISP-DM) model in (Shearer, 2000). The CRISP-DM model is a neutral framework for data mining and has been adopted by researchers and analysts in vast domains in an attempt to discover hidden knowledge in data (Catley et al, 2009). CRISP-DM consists of 6 sequential phases that start from (1) business understanding, (2) data understanding, (3) data preparation (4) modelling, (5) evaluation and finally (6) deployment.

A case study in NICU investigated the potential for modelling clinical data by performing temporal data mining was detailed in Catley et al (2009). The researchers noted that CRISP-DM does not account for the temporal abstractions that needs to be performed when handling high frequency clinical data streams before the modelling phase. To support the ability to mine temporal abstractions, Catley et al. (2009) introduced the CRISP-TDM model and demonstrated the potential for such a system in neonatal intensive care units (NICUs). However, with respect to understanding temporal relationships that may be exhibited in temporal abstractions generated in time oriented data, the CRISP-TDM model does not provide modules to support this.

An evaluation of the data mining approaches utilized is provided in Table 2.3 which highlights the data type been processed whether real time or static, the systems used to generate the data whether they are distributed or not and the temporal abstraction type. Out of the 110 papers reviewed, 23 papers tried to generate complex abstractions and use that as input for mining algorithms. There is a need for efficient algorithms that are capable of processing varying data types; static, multi-dimensional which can scale to structured and non-structured. Another aspect identified is the need to effectively select the most relevant data from huge volumes of abstracted data thereby leading to efficient mining processes.

Type of Data (real time, off-line)

Health care organizations have very many systems collecting patient data, monitors, sensors, administrative databases, patient charts, testing results from labs to pathology reports. In order to handle all these varying data types and sources, researchers have

proposed varying frameworks in an attempt to deal with data of two distinct types, static or real time data. Static data collected from medical sensors and stored in static databases for various analysis have been utilized in, (Silven, Dojat, & Garbay, 2005), (Altiparmak, Ferhatosmanoglu, Erdal, & Trost, 2006), (Verduijn, Sacchi, Peek, & et al., 2007), (Chaovalitwongse w., 2007), (Spokoiny & Shahar, 2008), (Combi & Oliboni, 2012), (Bouarfa & Daneklman, 2012) , (Huang, Lu, & Duan, 2012), (Wang, Liu, Shw, Nahavandi, & Kouzani , 2013), (Lammarsch, Aigner, Bertone, & et al., 2014). Recently use of electronic medical records has been utilized in (Beck et al 2016), (Titarenko et al, 2019) and (Campbell, Bass, & Masino, 2020).

Other systems have looked at real time abstraction of temporal data as it's generated from sensors and monitors as noted in (Bellazi , Larizza, & Magni, 2005), (Denai, King, Ross, & Mahfouf, 2007) , (Spokoiny & Shahar, 2007), (Minne, Ludikhuize, Jonge, & al, 2011), (Haghighi, Gilick, Krishnaswamy, & et al., 2010), (Portela , Santos, Silva, & et al., 2013), (Thommandram, Pugh, Eklund, McGregor, & James, 2013).

Processing real-time streams is a key factor in Artemis where data is captured from monitors and sensors in NICU as noted in (McGregor & Kneale, 2007) (McGregor & Stacey, 2007) (McGregor, Kneale, & Tracy, 2007) (McGregor C. , James, Eklund, & et al., 2013) and recently in (Inibhunu, et al., 2019).

Artemis as described in (McGregor et al 2009) is capable of capturing multiple individual streams of physiological data from bedside monitors and process data such as ECG, Heart Rate, Respiratory Rate and Blood Oxygen saturation SpO_2. The researchers recognized that

in clinical settings, a temporal abstraction system must be able to first capture the high frequency data generated every second from bedside monitors, McGregor et al., (2011; 2012; 2013).

Temporal Databases

Storage of time-oriented data requires temporal databases that are capable of holding events and transaction times. Various systems have been proposed that utilize relational models and related algebra and calculus to build temporal dimensions. Storage of valid and transaction time is explored in (Combi , Keravnou-Papailiou, & Shahar, 2010) where the use of relational databases is proposed. Others like (Ke, Gong, Li, & et al., 2014) have explored the use of NOSQL (non-relational) databases for storing semantically indexed temporal data. In the temporal database research community two temporal dimensions are recognized (Combi, Keravnou-Papailiou, & Shahar, 2010). One is the transaction time when moments are recorded or deleted from the database; the second is the valid time when the event is modelled in reality. User defined time is another aspect identified for modeling temporal dimensions. In this approach the system would be able to support semantics of time. Extending relational databases where time dimensions are added to the attribute level could to be explored in effective storage of temporal data, (Combi , Keravnou-Papailiou, & Shahar, 2010).

Temporal Reasoning

Making decisions using time oriented patient data involves reasoning and interpreting the information captured. To effectively utilize abstractions, systems aimed at valid

interpretation should be able to determine the boundaries for when abstractions convey meaningful information, understand inconsistency and uncertainly, capturing new interpretations based on new abstractions, understand relationships between different abstractions and being able to hypothesis a truth based on the abstracted relationships (Campos, Juarez, Palma, Marin, & Palacios, 2011). In aid for reasoning, several techniques have been proposed mostly from the computer science theories on artificial intelligence, Knowledge representation and reasoning, natural language and machine learning. Some of the popular machine learning techniques involves the use of data mining techniques for temporal pattern mining and knowledge representation. Association rule mining technique in (Agrawal & Srikant, 1995) is one of the major works sited by researchers such as (Chen, Hong, & Tseng, 2012), (Alvarez, Felix, & Carinena, 2013) (Lammarsch, Aigner, Bertone, & et al., 2014) where improved Apriori is performed to enable reductions of the huge amount of abstractions that can be generated from raw patient data. Others have utilized Bayesian networks such as the works defined in (Borchania, Bielzaa, Torob, & et al., 2013) (Yet, Perkins, Fenton, & et al., 2014) or combinational approach such as (Portela , Santos, Silva, & et al., 2013) who have proposed ensemble models that integrate rules from multiple algorithms.

Type of Systems

Clinical data is generated from various systems static or dynamic, in order to make sense of this data, a fusion of the data from the various systems is needed, and researchers have chosen different approaches to the acquisition of the temporal data. An ideal approach would be having a framework for easy integration of systems in a distributed flow, however a lot of

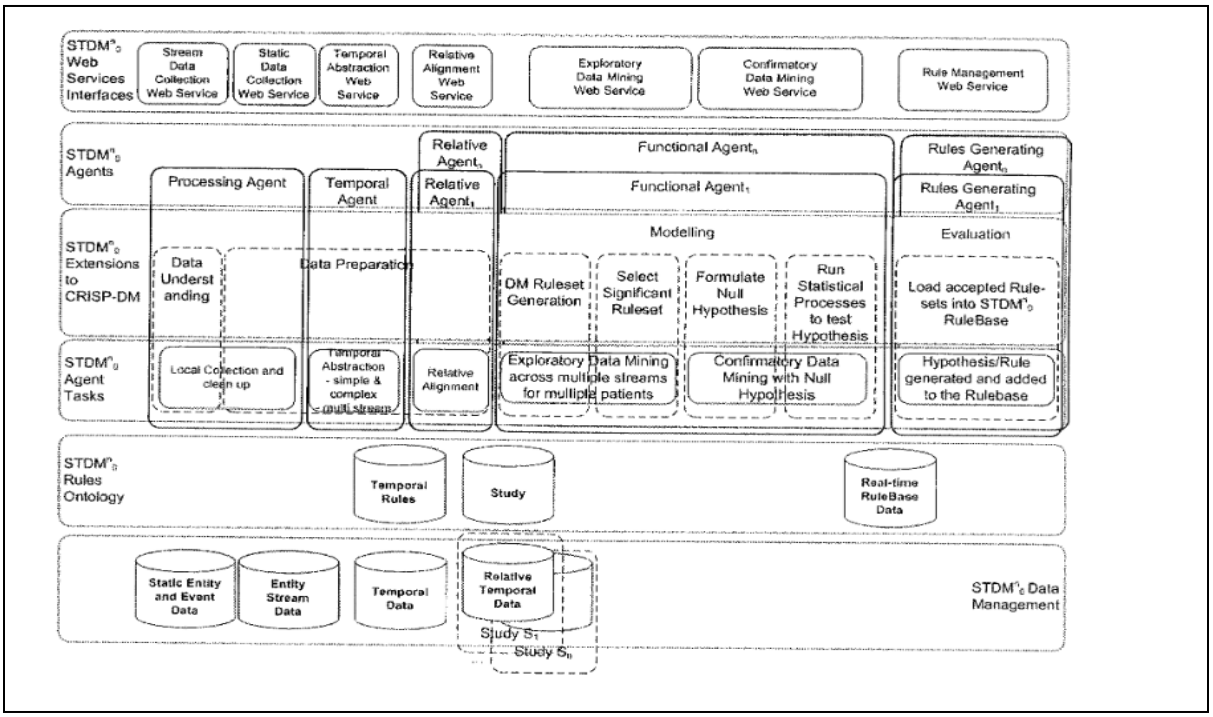
challenges exist on how such a framework would work in a clinical setting where equipment's and data are from very different systems and very different formats which would require a lot of configuration and formatting to get systems to integrate and communicate seamlessly with each other.

This is a central challenge and several groups of researchers are opting to use non-distributed approach such as (Bellazi , Larizza, & Magni, 2005), (Silven, Dojat, & Garbay, 2005), (Altiparmak, Ferhatosmanoglu, Erdal, & Trost, 2006), (Verduijn, Sacchi, Peek, & et al., 2007), (Campos, Juarez, Palma, Marin, & Palacios, 2011), (Amate, Forbes, Fontecave-Jallon, & et al. , 2011) , (Bouarfa & Daneklman, 2012), (Alvarez, Felix, & Carinena, 2013), (Heidjen & Lucas, 2013). Recognizing the need for a framework to enable synchronization of various data sources is one of the key aspects in the works that have incorporated distributed systems. From their resume system in (Shahar, 1997) to their other work dealing with Distributed KBTA in (Boaz. A., 2005) Incorporates distributed computing. The use of computational nodes is proposed in (Shabtai, Shahar, & Elovici, 2012) where distributed abstractions are processed in parallel to understand radical colon cancer surgery as a critical clinical pathway for colon cancer.

Another approach for synchronizing vast data sources captured from multiple distributed systems at the same time providing a platform for mining knowledge from data is detailed in (McGregor, 2011). The researcher introduced the service-based multi-dimensional temporal data mining framework ($STDM_0^n$) . This framework bridges the gap between management of data and data mining research by providing an environment for continuous data fusion and

multiple agents processing this data (McGregor, 2011). $STDM_0^n$ as shown in Figure 2.1 can be described as consisting of hierarchical layers with multiple components to support varying functions.

Figure 2. 1: Service-Based Multi-Dimensional Temporal Datamining Framework



A Brief overview of $STDM_0^n$.

Starting from bottom layer in Figure 2.1 is Data Management, this comprises components for management of static data entities, events data, data streams, and relative temporal data. Next is, the Rules Ontology that has components to support ontological rules such as temporal rules, rules for each cohort study and real-time rule base data. To support the various stages of the extended CRISP-TDM, there are components comprising of Agents and

Tasks. The processing agent has functions to support data understanding, temporal and relative agents for generating temporal abstractions and temporal rules in support of data preparation.

In the modelling component, there is a functional agent to support data mining rule generation, selection of significant ruleset, support formulation of null hypothesis and run statistical processes to test the null hypothesis. Finally there is an agent for rule generation to support the evaluation component. Web Services Interface is the top layer in $STDM_0^n$ and contains components for web services and user interfaces to support data stream collection, static data collection, temporal abstraction. To facilitate modelling and evaluation there are web service components for exploratory and confirmatory data mining, and rules management.

As $STDM_0^n$ incorporates CRISP-DM principles in the functions for temporal data mining, it inherits the shortcomings of CRISP-DM model which provides no components to support temporal abstraction, understanding temporal relationships that may be exhibited in temporal abstractions or formulation of temporal patterns that could be information utilized as input for further knowledge discovery functions.

Table 2. 3 : An Overview of Data Mining Systems

Author/Year	Data Mining Approach	Processing (Real Time, Static)	Distributed Data	Temporal Abstraction Types
Bellazzi et al, 2005	Apriori algorithm	Real time	No	Simple and complex
Silven et al, 2005	Probabilistic	Static	No	Simple and complex
Batal et al, 2009	Frequent mining	Static	Yes	Complex
Altiparmak, 2006	K-Medoid clustering algorithm	Static	No	Complex
Chaovalitwongse, 2007	KNN classification	Static	Yes	Complex
Denai, Mahouf 2007	Fuzzy clustering	Real time	Yes	Complex
Spokoiny, 2007	Temporal-abstraction	Real time	Yes	Complex
Ali et al, 2008	Frequent pattern mining	Real time	Yes	Complex
Spokoiny 2008	Associative property	Real time	Yes	Complex
Campos et al, 2011	Fuzzy interval representation	Static	Static	Complex
Garbay et al 2011	Probabilistic models that are adaptive to the domain	Static	No	Complex
McGregor, 2011	Framework for vast mining approaches	Static or real time	yes	Simple and complex
Haghighi et al, 2012	Fuzzy Logic and clusters	Real time	Yes	Complex
Combi et al, 2012	Relational Calculus	Static	Yes	Complex
Huang, 2012	Process mining	Static	Yes	Complex
Batal et al, 2013	Apriori based mechanisms	Static	Yes	Complex
Wang et al 2013	Bag of Words clustering	Static	No	Complex
Portela et al, 2013	Ensemble rules models	Real time	Yes	Complex
Álvarez et al 2013	Apriori strategy	Static	No	Complex
Heijden, lucas 2013	Probabilistic	Static	No	Complex
Lammarsch et al, 2014	Improved Apriori	Static	No	Simple
Gozt et al, 2014	Pattern mining and statistical pattern analyzer	Static	No	Complex

2.3 Clinical Application for Temporal Abstraction and Knowledge Retrieval

It's quite evident that there is a significant focus in knowledge retrieval in clinical settings; hospitals are full of medical equipment; monitors, sensors, PDAs, complex diagnostic and information systems that all hold a lot of very valuable data about patients. The ability to effectively utilize all this information in aid of decision making in various pathways of a patient journey from screening, diagnosis to choosing the right treatment for a patient is at the heart of a lot of research in medical systems (Combi , Keravnou-Papailiou, & Shahar, 2010).

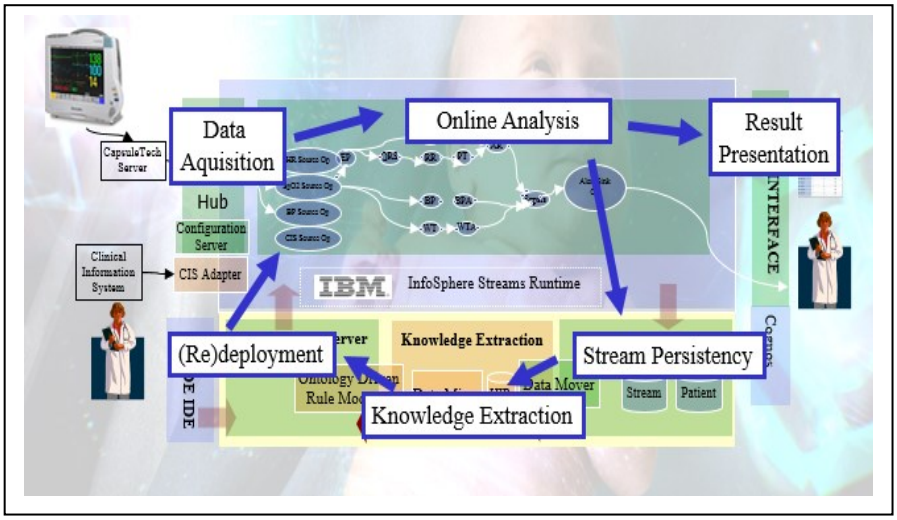
Analysis of Hemodialysis data is attempted in (Bellazi , Larizza, & Magni, 2005), (Sacci, Larrizza, Combi, & Bellazzi, 2007), (Combi, Pozzi, & Rossato, 2012), ICU patients in (Silven, Dojat, & Garbay, 2005), prediction of heparin induced thrombocytopenia is performed in (Batal, Sacchi, Bellazzi, & Hauskrecht, 2009), evaluation of phase III clinical trials data to understand the toxicology measures is presented in (Altiparmak, Ferhatosmanoglu, Erdal, & Trost, 2006). Understanding of cardiac surgery patients is performed by (Verduijn, Sacchi, Peek, & et al., 2007) (Denai, King, Ross, & Mahfouf, 2007). Understanding bone marrow transplant data is presented in (Shahar, 1997), (Spokoiny & Shahar, 2008) and (Klimov, Shahar, & Taieb-Maimon, 2010). In neurology medicine, (Lara, Moreno, Perez , & et al., 2008) looked at stabilometric and posturographic systems.

In neonatal intensive care units (NICUs) a lot of research has been completed using the ARTEMIS platform where researchers acquire data real-time from preterm babies in ICU. They classify conditions in real-time based on physiological response relating to infections, drugs, neonatal spells (pauses in breathing, decrease in blood oxygen saturation, or decreased heart rate), pain and detection of late onset of neonatal sepsis (LONS) (McGregor

& Kneale, 2007) (McGregor & Stacey, 2007) (McGregor, Kneale, & Tracy, 2007) (McGregor C. , James, Eklund, & et al., 2013) and (Thommandram, Pugh, Eklund, McGregor, & James, 2013). Others who have looked at NICU include (Hunter, Freer, & Gatt, 2012), (Minne, Ludikhuize, Jonge, & al, 2011) in the detection of organ failure on patients who had 2 to 7 days of ICU, (Portela , Santos, Silva, & et al., 2013) also predicted organ failure in ICU. Detection of sleep apnea-hypopnea syndrome using polysomnography tests was looked at by (Alvarez, Felix, & Carinena, 2013). Accurate prediction of ICU stay is presented in (Huang, Lu, & Duan, 2012). Avian Influenza (H5N1) is studied in (Campos, Juarez, Palma, Marin, & Palacios, 2011) while monitoring of human health.

Fourteen systems are highlighted in Table 2.4 where researchers have attempted to generate temporal abstractions from real-time in varying clinical domains. To effectively assist in decision making in a specific domain, accommodation of medical knowledge ontologies are needed as highlighted in (Silven, Dojat, & Garbay, 2005) . However to avoid the development of segregated clinical systems, research on mechanisms that can scale to varying clinical domains if explored could steer towards more comprehensive solutions which can serve multiple clinical domains. Such an approach is attempted in Artemis, an architecture for capturing multiple real-time high frequency data streams from NICU (McGregor et al, 2013). The Artemis framework as shown in Figure 2.2 is comprised of several components as detailed next.

Figure 2. 2: Artemis Framework



The Data Acquisition component enables the real-time collection of data from monitors and electronic health data from clinical information systems. The Online Analysis component processes the data forwarded by the Data Acquisition component in real-time. IBM InfoSphere Streams is utilized as the technological solution to support real-time processing of incoming data streams. This approach allows deployment of clinical rules representing correlation between behaviours of interest in physiological streams for any condition that is been monitored. The Knowledge Extraction component emulates the $STDM_0^n$ model in (McGregor, 2013).

Data Acquisition and Online Analysis occur at the same location where medical devices acquire data. In addition, Artemis allows analysis of stored physiological data for new knowledge discovery of unknown conditions onset physiological behaviours, through an instantiation of $STDM_0^n$ McGregor (2011). The result presentation component provides the results of the online analysis to end users. The Data Persistence component allows storage of

analysis generated from Online Analysis component. The Deployment Server contains modules that allow clinical rule generation and modification which are then deployed within InfoSphere Streams Runtime.

To support the components in the Artemis system, various clinical research has been carried out including; models for clinical knowledge discovery as defined in CRISP TDM (Catley et al, 2008), an assessment of trends and opportunities for real time clinical decision support in neonatology (Bressan, James, & C., 2012), using morphine pharmacokinetics and pharmacodynamics (PKPD) parameters to estimate morphine plasma concentrations and their correlation to heart rate variability in neonatal population (Bressan, James and McGregor, 2012), analysis on incidence of neonatal vagal spells at different gestation age in (Naik, Bressan, James, & McGregor, 2013) potential for using HRV analysis for detection of late onset neonatal sepsis (LONS) was demonstrated in (McGregor, Catley & James 2012), classification of neonatal spells using real time streaming was completed in (Pugh et al, 2013), correlation of neonatal conditions in detection of life threatening conditions such as sepsis in (Thommandram et al, 2013), assessment of physiological streams of data from Artemis for translation into clinical rules for pain profile in infant (Choi, Bressan, James, Pugh, & McGregor, 2013) detection of apnea by analyzing one minute epochs of respiratory impedance using feature selection and kNN classification described in (Thommandram, Eklund, McGregor, 2013), correlation of retinopathy of prematurity with blood oxygen saturation in (Fernando, McGregor, & James, 2016) and visualization of real-time data streams in (Kamaleswaran, Collins., James, & McGregor, 2016). Although Artemis has provided a foundation for several clinical research, there is opportunity for enhancing the

existing architecture to include capabilities for detecting and representing temporal relationships that may exist in data captured from vast clinical domains.

Table 2. 4 : An Overview of systems on Temporal Abstractions in Clinical Domains

Author/Year	**Clinical Domain/Purpose**
(Bellazi , Larizza, & Magni, 2005)	Quality assessment of hemodialysis patients
(Denai, King, Ross, & Mahfouf, 2007)	Cardiac intensive care monitoring
(Spokoiny & Shahar, 2008)	Incremental monitoring of bone-marrow transplant patients
(Martins, Shahar, Goren-bar, & et al., 2008)	Intelligent querying of real-time clinical data
(Ali, Elhelw, Atallah, & et al., 2008)	Ear-worn activity recognition (E-AR) sensor
(Spokoiny & Shahar, 2007)	Evaluated in many clinical domains, i.e. Monitoring children growth, insulin-dependent diabetes and guided-based care in oncology
(Haghighi, Gilick, Krishnaswamy, & et al., 2010)	Mobile Health monitoring of ECG signals
(Minne, Ludikhuize, Jonge, & al, 2011)	2 to 7 days of ICU days for the detection of organ failure
(Vettier, Amate, & Garbay, 2012)	Human health monitoring, using ambulatory physiological sensors.
(McGregor, Catley & James 2012)	Detection of late onset neonatal sepsis
(Portela , Santos, Silva, & et al., 2013)	Predict Organ Failure in ICU
(Pugh et al, 2013)	Classification of Neonatal Spells
(Thommandram, Pugh, Eklund, McGregor, & James, 2013; Thommandram et al, 2014)	Classifying Neonatal Spells
(Fernando, McGregor, & James, 2016)	Correlation of retinopathy of prematurity with blood oxygen saturation

2.4 Research Direction on Temporal Abstraction and Data Mining

It's quite evident that for the ability to effectively utilize clinical data in aid of decision making in healthcare, mechanisms such as temporal abstraction and data mining for processing and interpreting the information generated are needed. Since the literature review presented in (Stacey & McGregor, 2007), there is a trend on the evolution of temporal abstraction and data mining on the clinical side. Researchers have recognised the need to have systems that are able to transcend on data streams generated every second from monitors and sensors in intensive care units; having frameworks that are able to capture the continuous flow of data is a key aspects of the ARTEMIS project (McGregor & Smith, 2009), (Minne, Ludikhuize, Jonge, & al, 2011), (McGregor, James, Eklund, & et al., 2013).

Various temporal abstractions techniques have been suggested to work on data of varying formats static or real-time (Shabtai, Shahar, & Elovici, 2012), looking at dependencies in multivariate time series and then finding temporal relationships (Bellazi , Larizza, & Magni, 2005) and using ontologies for medical validation in (Silven, Dojat, & Garbay, 2005).

An evaluation of temporal abstraction techniques is provided in Table 2.1 while Table 2.2 highlights the 2 main types of abstractions identified in the papers reviewed and used as one of the attributes in evaluating systems which have attempted mining temporal abstractions as presented in Table 2.3.

Data mining as a technique for knowledge retrieval from huge volumes of data has been applied in different areas of research, and in particular in healthcare. The adoption of the association rule mining and the frequent pattern mining is seen in most of the papers reviewed (Ali, Elhelw, Atallah, & et al., 2008), (Batal, Sacchi, Bellazzi, & Hauskrecht, 2009) and

recently in (Titarenko, et al. 2019). Other researchers have used Bayesian networks; a good review on use of data mining using neuro-networks is presented in (Orphanoua, Athena, & Keravnough, 2013). Probabilistic models are still a dominant part of data mining and have been used in (Fuchs, Gruber, Pree, & Sick, 2010). Other researchers have investigated clustering, k means, k-Mediod (Chaovalitwongse w., 2007), (Denai, King, Ross, & Mahfouf, 2007), (Lara, Moreno, Pere, & et al., 2008), (Haghighi, Gilick, Krishnaswamy, & et al., 2010).

A comprehensive evaluation of data mining systems reviewed is provided in Table 2.3. All these systems aim at retrieving potential knowledge that is hidden in the temporal abstractions extracted from the data. However, most of the systems are tailored to a specific clinical problem and none of the techniques utilized looks at the potential relationships or temporal patterns that could be exhibited in the time oriented data. Discovery of search patterns could help in early discovery of onset of conditions in various clinical domains.

The extraction of knowledge from temporal abstractions generated is performed using data mining techniques and most of the techniques demonstrate mining using static data collected from various domain in healthcare as described in (Altiparmak, Ferhatosmanoglu, Erdal, & Trost, 2006), (Campos, Juarez, Palma, Marin, & Palacios, 2011), (Amate, Forbes, Fontecave-Jallon, & et al., 2011) (Combi , Keravnou-Papailiou, & Shahar, 2010) (Combi, Pozzi, & Rossato (2012), (Batal I. , Valizadegan, Cooper, & et al., 2013), (Alvarez, Felix, & Carinena, 2013), (Heidjen & Lucas, 2013), (Lammarsch, Aigner, Bertone, & et al., 2014). Others have also taken account of the temporal nature in data and attempted real time mining (Denai, King, Ross, & Mahfouf, 2007), (Spokoiny & Shahar, 2007), (Ali, Elhelw, Atallah, & et al., 2008) , (Vettier, Amate, & Garbay, 2012), (Minne, Ludikhuize, Jonge, & al, 2011), (Portela, Santos,

Silva, & et al., 2013). In other works, (Sacci, Larrizza, Combi, & Bellazzi, 2007) (Lammarsch, Aigner, Bertone, & et al., 2014) have developed algorithms that run on real-time stream of data been captured in NICU generating rules with an aim of detection of clinically significant changes in blood oxygen saturation and heart rate.

A diverse number of clinical applications as shown in Table 2.4 from quality assessment of hemodialysis patients in (Bellazi , Larizza, & Magni, 2005) to classifying neonatal spells (Thommandram, Pugh, Eklund, McGregor, & James, 2013) where researchers processed real-time clinical data.

Pruning of temporal abstraction is another item tackled in the research community. Temporal abstractions generate a huge amount of data that is quite complex to synthesise. There is a need to identify outliers and thresholds for when to create an abstraction. Several methods have been attempted to use Allen logic (Batal, Valizadegan, Cooper, & et al., 2013) for filtering non-predictive abstracts.

As noted in this review, various techniques and algorithms have been introduced with a main aim of retrieving knowledge from complex volumes of clinical data. The need to understand the retrieved knowledge is a critical part of a clinical setting where such knowledge could be quite helpful in aid of decision in health care and possibly improve the outcome of a patient. However, the current algorithms generate information that is quite difficult to comprehend and clinicians might feel like drowning with the information. To address this challenge, visualization techniques have been incorporated in various frameworks (Combi, Keravnou-Papailiou, & Shahar, 2010; Kamaleswaran et al, 2016)

Two kinds of visualization are identified in the literature as noted in (Kamaleswaran & McGregor, 2016), the first one deals with visualization of patient data that is already mined and the second one deals with the interactive visual exploration where the user provides input for knowledge extraction (Combi & Oliboni, 2012). Interactive mining is also proposed in (Alvarez, Felix, & Carinena, 2013) where use of seeds allows users to drive the mining process. Interactive visualization of mined results and parameterization is also in (Lammarsch, Aigner, Bertone, & et al., 2014). Informative visualizations in NICUS is attempted in (Kamaleswaran, Collins., James, & McGregor, 2016). These systems although effective at providing canned visualizations are limited to a specific data set and do not have mechanisms for adapting to a users demand or accommodating changes in data. A highlight on the different types of frameworks that have integrated some aspects of visualization is provided in Table 2.5.

Table 2. 5 : An Overview of Systems with Visualization Capabilities

Author/Year	Summary	Visualization Integrated
(Kamsu-Foguem, Tchuente-Foguem, Allart, & et al., 2012)	Machine learning knowledge discovery with visual methods	Interactive graphical user interface
(Haghighi, Gilick, Krishnaswamy, & et al., 2010)	Situation aware adaptive processing (SAAP) on data streams	Adaptive visualization in mobile phones
(Gotz, Wang, & Perer, 2014)	The researchers advanced the ICDA system by adding functionality that allow for ad hoc temporal constrained based queries	Web-based Interactive pattern diagrams with querying interfaces
(Lammarsch, Aigner, Bertone, & et al., 2014)	Utilize the SAPPERLOT procedure for visualization of patterns explored	Patterns represented using river and arc views
(Kamaleswaran, Collins, James and McGregor, 2016)	CoRAD, Visualization system in Clinical Care	Task analysis and alignment of physiological data to clinical events, visualization tested in NICU

2.5 Key Findings

There is a need for having systems that can take all the information generated in medical organizations from monitors, sensors, PDAs, administrative databases, lab results, images that are generated on patients every second. However, several challenges have been identified in the literature on how one can take all that data and present it to clinicians in an intuitive manner so they are able to use it when making decisions on what is the best treatment therapy for a specific patient.

Temporal abstraction has attempted to bridge the gap between raw data and generate high level granularities that are easier to understand and better represent pathophysiology pathways for various conditions and diseases (Combi et al, 2010) (Combi and Oliboni, 2012).

The existing frameworks that aim to provide a process for effective temporal abstraction such at the CRISP-DM and CRISP-TDM^n models do not integrate components for data reduction, generation of temporal abstractions or discovery of temporal relations and patterns in data preparation phase before passing the data to the modelling phase. Similarly the adoption of CRISP-DM into the multiple dimensional framework in $STDM_0^n$ does not incorporate components for temporal abstraction, understanding relationships in the temporal data and subsequent patterns that may result from processing multiple data streams. As a result, unnecessary data is passed to subsequent components within those frameworks which could be computationally expensive especially when handling huge volumes of real-time data streams. In addition, both models do not provide clear guidelines on the modelling phase on how mining frequent patterns can be handled when dealing with real time data streams.

In this literature review, several aspects with respect to the process of taking raw patient data and generating meaningful information have been determined as open research areas, as outlined below;

1. There is a need to build algorithms that are capable of generating meaningful abstracts from various data types and maintain the temporal nature of the data coming from static or multi-dimensions systems, include the ability to scale to changing data types structured or non-structured.

2. Utilizing data mining algorithms for filtering irrelevant data or picking the right intervals that maintains the temporal nature of the data that can process real-time volumes of data streams.
3. Temporal abstractions from medical data require clinical validation to ensure the knowledge retrieved is relevant to critical care. There is a need to utilize ontologies in the mining process to ensure that the knowledge extracted from the underlying data is relevant to patient care and outcomes.
4. Storage of real time data generated from multi-dimensional systems is also something that needs to be explored, the traditional relational databases needs to be improved to be able to handle temporal data including ability to incorporate user defined time and time semantics.
5. Temporal data maintenance needs to be incorporated in systems built in clinical systems. There is a need to explore the integrity of constraints based on temporal reasoning for data validation during acquisitions.
6. Reliability and trustworthiness of the results generated is a huge challenge identified in (Banaee et al, 2013) where the experts in the decision making tasks are reluctant to embrace any automated data mining process especially if there is minimal or no interactions with the systems.
7. Presenting the information generated in a manner that is intuitive and easy to comprehend for healthcare providers is another aspect that needs further research. Visualization techniques are incorporated in most works that are now emerging in the field of temporal abstraction and data mining; this will be an area that will continue

to grow especially now that there is a huge appetite to utilize data in mobile networks and smart phone.

2.6 Identified Shortcomings of Temporal Abstraction and Temporal Data Mining

This literature review provides a clear overview on temporal abstraction and temporal data mining in clinical settings. In particular, the varying methodologies applied on knowledge retrieval from raw patient data, data types processed, techniques used in temporal abstractions, the algorithms used for temporal data mining and the various medical environments where temporal abstraction has been applied. The review also details the challenges that come with handling huge volumes of data while trying to utilize it for aid in critical decision making in clinical settings.

Open research areas to ensure the usability of information generated from data mining algorithms in clinical setting has been determined and as noted below;

1) Robust processes for generating temporal abstractions that maintain the temporal nature of the original data.

2) Efficient algorithms that combine temporal abstractions and data mining thereby filtering and pruning irreverent abstracts and can scale to continuous real-time data streams.

3) The need for algorithms that have clinical input therefore increasing the adoption and trustworthiness of results from mining algorithms, this can be achieved by incorporating medical ontologies or case base reasoning.

4) Systems that are able to store and process large volume of data from various processes be it static or real-time, distributed or not, in addition to integrate with multiple data sources.

5) Visualization techniques that make it easy for one to comprehend the results of knowledge systems and inclusion of user interaction whereby hidden details might not be obvious and can use recommendation style views thus reducing the click based approached in current visualization systems.

6) The ability to have information presented in real-time and incorporation of mobile technology is another area that needs exploration especially since we are in the age of instance information collection. Systems that can take data in real-time and provide analytics in similar fashion and having the reliability needed in clinical settings needs to be explored.

This thesis research investigates the problems detailed 1 to 4 in the computer science contribution. In application to health informatics and medicine this thesis investigates use of techniques developed in computer science contribution for processing of physiological data streams. The reliability and visualisation open areas as detailed in 5-6 are outside the scope of this thesis.

2.7 Conclusion and Research Impact

The literature review discussed in this chapter has provided a clear background of temporal abstraction and data mining in clinical settings. The motivation for the review was to form a backbone for research and development into the area of temporal knowledge acquisition, retrieval, reasoning and interpretation. The review has assessed and synthesised the following areas: temporal abstraction, comprehensive and comparable review on types and techniques in the abstraction process, data mining algorithms used in abstraction process, data types, storage, clinical applications and visualization.

Key areas where research is still needed to foster the usability of time series data mined from clinical systems has been identified, this includes; 1). Generation of temporal abstractions and understanding relationships among abstractions, 2) Efficient algorithms that combine temporal abstraction and data mining complimented by use of ontologies, 3) Algorithms that maintains temporal nature of mined data, 4) Visualization techniques with interactive capabilities, and 5) Systems that can process data, provide analytics in real-time and having the reliability needed in clinical settings.

Researchers have attempted to provide varying techniques for utilizing clinical data, however, there are several shortcomings identified in the literature as discussed in chapter 2. In particular, current systems lack the ability to handle the variety, volume and velocity of continuous data streams especially when trying to derive any hidden relationships and patterns that may exist in the data (Combi et al 2010). As noted in (Shahar, 1999) and (McGregor, 2009), physicians who make critical decisions on diagnostic and therapeutic modalities are overwhelmed by the amount of data generated continuously at the bedside. In addition, the existing systems lack the ability to effectively present useful information to end users when most needed which results in cognitive overload (Kamaleswaran, 2016).

In an attempt to address these challenges, several systems have been proposed for deriving hidden patterns from time-series data Shahar et al (1999) , Belazzi et al (2005) , Chaovalitwonge (2007) Hunter et al (2012), Lui et al (2014). However, the current approaches are computationally expensive as extensive data loads need to be processed at each iteration of the knowledge discovery process. In particular, in the NICU, as shown in Chapter 1(Table 1.1), the amount of data generated in an hour from a single patient is millions of records and

this can grow to billions of data within 24hrs. As deriving hidden patterns requires multiple iterations on data, processing such data volumes using existing techniques can grow exponentially, even for a single iteration.

With respect to a unified system for temporal data mining, the literature review shows that there has been an attempt to mine temporal aspects of time series data through frameworks such as CRISP-TDM for generating temporal abstractions (Catley et al., 2009), the multiple system framework $STDM_0^n$ in (McGregor, 2011) and the instantiation of these frameworks in ARTEMIS cloud architecture in (McGregor et al, 2011; 2012; 2013). However, research still remains open on how to generate temporal abstractions, understand relationships in abstractions and then creating patterns which if quantified could explain a phenomena about a subject or a particular data domain.

To provide meaningful contribution in clinical settings such as neonatology, the ability to understand any relationship that may exist in time series data generated from patients physiological features such as heart rate, respiratory rate, blood pressure and transcutaneous blood oxygen saturation can be translated to a computational problem known as a pattern recognition and classification. Humans are able to identify objects that are similar using various measures; type, size, shape, color among others and given the sequence of occurrence of the measures, one can deduce that as a pattern (Doughtry, 2013). This is easy on a simple set of objects or data sets, however the ability to deduce and classify any patterns from millions of physiological time series data generated every hour by monitoring devices and sensors at the bedside is extremely difficult.

Pattern recognition is a complex research problem and as such, a review has been carried out and presented in Chapter 3 to understand the techniques and methodologies applied for temporal pattern recognition.

Chapter 3

3 A Review on Pattern Recognition

This chapter presents a review on pattern recognition methods. As noted in chapter 2, data types captured about patients are mostly time based resulting in continuous time series data points. This thesis recognizes that discovery of any hidden knowledge that may exist in such data types falls under pattern recognition problem space extensively studied in data mining and machine learning research as noted in (Dougherty, 2013) and (Rajaraman, Ullman, & Leskovee, 2010). The focus of this review is to understand the techniques and methodologies applied for temporal pattern recognition and representation of data.

A good analogy on patterns is provided in (Doughtry, 2013). The researcher highlights that, a combination of structures about an object can form a pattern, for example, a human body has specific features that form the patterns of the human body. Therefore if there is a difference in the way joints are interconnected, it is possible to deduce a pattern between different groups of people. With this notion, it would be possible to then create a specific class of people with similar characteristics (patterns) and then can use this class to assign a new person to the group if they fit the characteristics of the identified class, this is called classification. Dougherty (2013) notes that classification of objects to specific groups doesn't mean they are identical but do have some properties that distinguishes from each other.

Other researchers like (Murty & Devi, 2011) describes pattern recognition as a method for classifying data based on knowledge gained either from extracted features which forms

patterns or statistical information. A typical pattern recognition system involves taking raw data and converting it to a form that is machine readable, an example of such a system is signal processing from medical equipment.

Pattern recognition and classification are at the core of various research fields such as machine learning, statistical analysis, data mining, knowledge discovery and their application to a broad range of spectrum; fraud detection, speech recognition, image processing, commerce and health informatics (Dougherty, 2013).

There are two categories identified in the literature on the approaches utilized for pattern recognition, Statistical and Syntactic, Murty et al (2011). Statistical pattern recognition is more popular given the potential for handling noisy data with statistics and probability while syntactic pattern recognition is utilized to determine structural relationships in variables and in natural language processing to infer grammatical relationships in strings (Murty, 2011). Syntactic pattern recognition when applied to time series data is a research problem that seek to preserve the structural nature of data and is described as temporal pattern recognition in (Combi, Keravnou-Papailiou, & Shahar, 2010). In this literature review the focus is to understand methods used in statistical pattern recognition and temporal pattern recognition.

The process of discovering patterns from data is complex as it can involve a combination of multiple methods, techniques and steps. In this review we take a scaled approach to understanding this complex process and have several sections as follows. First a summary of the methods used for representing patterns are discussed in section 3.1. The use of

dimension reduction strategies for handling large volumes of data are discussed in section 3.2. Statistical pattern recognition methods are discussed in section 3.3. Methods for temporal pattern recognition are discussed in section 3.4 while section 3.5 discusses application of pattern recognition research in critical care. A summary of temporal pattern recognition is provided in section 3.6 and key findings are presented in section 3.7

3.1 Types of Pattern Representation

One of the steps in pattern recognition process involves an ability to describe the aspects of a pattern and this is referred to as pattern representation. In this section a summary of the methods presented in the literature for pattern representation is discussed. This includes techniques that use data structures and proximity measures (Murty, 2011) as well as probabilistic methods described in (Rabiner, 1989).

Pattern Representation with Data Structures

(a) Patterns as Vectors: A data matrix is comprised of columns and rows. The columns represents features or attributes while rows represents the vectors and each element in a vector can represent a pattern feature (Murty & Devi, 2011).

(b) Pattern as Logical description: Using logic to represent features, for example, (color = red ∪ blue) and (Make = leather) and (shape = sphere) to represent a cricket ball

(c) Patterns as Strings: A sentence in language mostly used in DNA sequencing

(d) Patterns as Fuzzy sets: Membership to a particular set is based on incomplete and imprecise data, normally associated with some approximation between 0 and 1.

(e) Patterns as Trees and Graphs: Various types of trees (spanning, frequent pattern) where nodes represent classes and edges leading to a particular node represent some relationships, trees can represent one or more patterns.

Patterns Representation using Proximity Measures

(a) Distance Measures: Weighted similarity measures look to understand relationships between different objects, such is the Hamming distance that takes two vectors and calculates the sum of the number of features that differ in value between the two. A user provides a threshold to measure the level of relevancy, however this method is seen as ineffective when handling highly correlated features.

(b) Non Metric functions: When dealing with non numerical data, attributes are treated as features instead of real numbers. Methods for grouping and discriminating attributes include; information entropy, K-median, Edit Distance, Mutual Neighbourhood Distance (MND), Conceptual Cohesiveness or Kernel Functions. A comprehensive review on these methods is provided in (Dougherty, 2013).

Patterns Representation using Probabilities

A set of time oriented observations can be modelled as a finite state machine representation where for a given period of time the associated time series data segments transitions randomly to one or many states based on some associated weights. In Markov chains this weight is normally based on some probability distribution.

Given $S = (s_1, s_2, \dots, s_i, \dots, s_j, \dots, s_n)$, as a set of n states and each of the states s_i are unique in that, for all s_i, $s_j \in S$ then $s_i \cap s_j = \emptyset$ where $1 < i < j < n \in \mathbb{R}$

$X = (x_{11}, x_{12}, \dots, x_{n1}, \dots, x_{nn})$ is a set of transition matrix and each x_{ij} represent the probability of transitioning from state s_i to s_j, such that $\sum_{j=1}^{n} x_{ij} = 1, \ \forall i$

Let P be a process such that (Pr) is the probability of P being in a particular state and follows the Markov property where by if P is a stochastic process then conditional probability distribution of future states in P is dependent only upon the present state, not on the sequence of events that preceded it. Therefore, if there exists some state $s_i \in S$, $s_1 \ and \ s_n$ are start and end states respectively, then the conditional probability $Pr(s_i)$ of process P been in s_i at some time t is defined as,

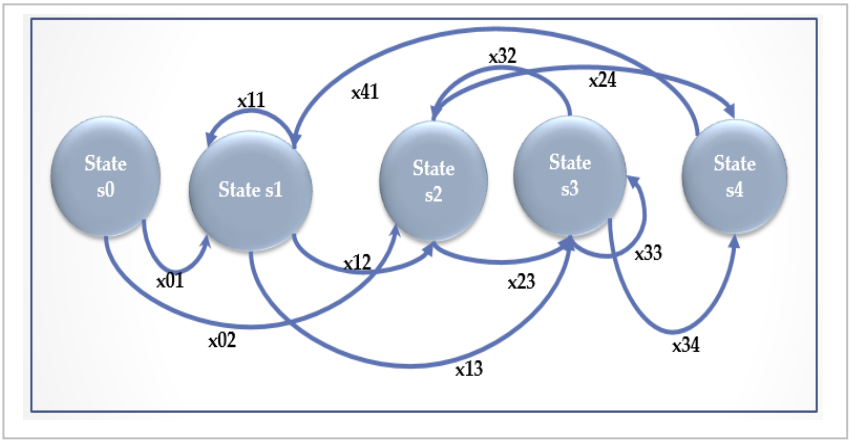
$$Pr(s_i(t) | s_1(1) \dots . s_{i-1}(t-1)) = Pr(s_i(t) | s_{i-1}(t-1)) \quad (1)$$

However, it's not possible to know ahead of time what's the most probable sequence of states an observation in a data stream may transition to in a given time interval. In particular, given a time series data stream D with observations captured at each second, how can one identify the sequence of probable states $(s_1, \dots, s_i, \dots, s_j, \dots, s_n) \in S$ that elements in D may transition to in a given time interval, where $1 < i < j < n$. To facilitate this process, two additional definitions are added to the Markov chains to satisfy the hidden markov principle (Rabiner, 1989). To this respect, given a set $O = \{o_1, o_2, \dots, o_K\}$ that contains sequence of *K* observations each drawn from some abstraction $A = \{a_1, a_2, \dots, a_n\}$ and a sequence of observation likelihood each representing the probability of an observation o_i being at some

state s_i leads to the output probability $Pr(o_i|s_1, \ldots, s_i, \ldots s_K, o_1, \ldots, o_i, \ldots, o_K) = Pr(o_i|s_i)$ such that o_i depends only on the state s_i.

Suppose there exists a process M and there are 5 known states *i.e. (state s0, state s1, state s2, state s3, state s4)* that M can transition to as shown in Figure 3.1. Application of HMM to M calculates the probabilities of transitioning from the start state to each of the other 4 states. M has an initial state *(state s0)* and end state *(State s4).* The arcs in blue indicates the probability of transitioning from one state to another. For example *x01* indicates the probability of transitioning from state *s0* to state *s1*. There are several sequence of transitions from a specific state and these sequences represent patterns. From Figure 3.1, starting from *state s0*, there are several sequences possible i.e. *{s0, s2, s3, s4}, {s0, s1, s4}, {s0, s1, s1, s2, s3, s3, s2, s4}* and the probability *(Pr)* of the sequence *{s0, s2, s3, s4}* = *Pr (s4|s3) Pr(s3|s2) Pr(s2|s1)Pr(s0).*

Figure 3. 1: A sample process with states and transition probabilities.



There are three computation problems associated with formulation of transitions with probabilities as noted in (Rabiner, 1989) as follows;

•Evaluation problem: Given some observation sequence O, a set of random variables X, Y such that $X = (x_1, \dots, x_n)$, $Y = (y_1, \dots, y_n)$ and a model λ = (X, Y) where X and Y are the model parameters, what is the likely hood of the occurrence of an observation sequence Pr (O| λ).

•Decoding problem: What is the optimal hidden state sequence to produce an observation sequence.

•Learning Problem: Given an observation O and the set of states in the model (λ), learn the HMM parameters X and Y.

3.2 Dimension Reduction Methods

Pattern representation of a small set of data is a simple task for humans but becomes a computationally expensive process when faced with massive datasets such as the high frequency data generated every second from bedside monitors resulting to millions of records every hour (Inibhunu, et al., 2019). In order for pattern representation methods to be effective when handling complex datasets, mechanisms for reducing the amount of data processed are studied in dimension reduction research. There is a need to represent features in simpler ways as noted in (Murty & Devi, 2011). Two-dimension reduction approaches are utilized in an attempt to reduce the computational complexity of handling huge dimensions of data, feature extraction and feature selection. A brief summary on these techniques are highlighted below, a comprehensive review is detailed in (Liu, Wu., & Zhang, 2014).

Feature Extraction

(a) Fisher's linear discriminant: A method closely related to analysis of variance and provides a combination of features that separates or characterizes two or more events, classes or objects. The main aspect is the attempt to express one feature using a linear combination of other features i.e. regression analysis.

(b) Principal component analysis: The goal in this approach is to represent the data in a lower dimensional space by preserving the original space with as little variance as possible. Each principle component is a linear combination of the original attributes. There are various techniques to creating the principal components such as multidimensional scaling and eigen value transformation describes in (Jensen & Shen, 2008).

Feature Selection

This involves selection of attributes that might be seen as fitting some user criteria or interest and then using these as basis for classification by ensuring that a limited number of features are used to represent patterns. A measure of good is determined using deterministic or probabilistic techniques. Some methods for selecting features includes; (a) Exhaustive Search: Involves searching through all the features and identifying best features, (b) Branch and Bound Search: This approach avoids the exhaustive search through divide and conquer mechanisms, (c) Sequential search: This looks as starting with an empty list using varying techniques from forward, backward or floating forward searching to get best attributes.

3.3 Statistical Pattern Recognition Approaches

Given a set of patterns of known classes, statistical pattern recognition involves designing a classifier that can assign a new pattern to a class. This approach is known as supervised learning in the machine learning research and involves integration of pattern representation techniques in order to build classifiers as part of a pattern recognition process. Several methods have been developed that utilize some of the pattern representation techniques discussed in section 3.1 as part of a pattern recognition process. A highlight on some of these approaches is discussed next.

(a) Nearest Neighbourhood classifiers (NN): This method assigns a pattern to a predefined category based on characteristics of its nearest neighbours. There are several variations of NN, k-nearest neighbors (kNN), modified kNN (MkNN), fuzzy kNN, r nearest neighbours, (rNN) (Murty & Devi, 2011).

(b) Bayesian Classifiers: This approach uses Bayes theorem of probabilities where by prior probabilities and distributions of a class are known. A classifier then uses the known priori probability to derive posterior probabilities using likelihood values to choose the most probable hypothesis.

(c) Hidden Markov Models: These are used for classifying patterns where each pattern is a sequence of sub-patterns. Markov models starts with a state and then goes through sequence of states based on the probabilities associated with each sequence.

(d) Decision Trees: This approach uses tree-like graphs where tree leaves represent a decision state and branches represent the conditions satisfying the decision. Decision trees can be used to classify patterns.

(e) Support Vector Machines: Developed by (Vapnik, 1995), this approach uses data points as support vectors which are then utilized for identifying patterns in differing groups of data using structures such as hyperplanes. Various approaches for minimizing the errors in picking the optimal vectors for classification is a continued research area in optimization.

(f) Ensemble classifiers: A combination of classifiers is used in this approach where imperfection of a classifier is enhanced by assembling of multiple classifiers. Methods for combining different classifiers includes weighting measures and probabilistic estimates for deriving the ensembles (Murty & Devi, 2011).

(g) Clustering: This involves grouping of patterns with similar characteristics into similar groups known as clusters. Clustering is a part of Gaussian Mixture Models (GMM) which assumes that there are certain number of Gaussian Distributions (or the Normal Distribution) and each of these distributions represent a cluster. Given k Gaussian Distributions indicates there are k number of clusters each with a mean and covariance values. Suppose there are four Gaussian Distributions i.e. GausD1, GausD2, GausD3 and GausD4. These distributions each have a mean (mean1, mean2, mean3, mean4) and a variance (var1, var2, var3, var4) respectively. For each given data points, clustering identifies the probability of each data point belonging to each of these distributions. There are several clustering algorithms;

k-means, k-mediod, survey on clustering algorithms is detailed in (Murty & Devi, 2011).

3.4 Temporal Pattern Recognition

Pattern recognition has been applied to various datatypes and structures (structured and non-structured). Structured data types include time series data while non structured data involves text or natural language. Each of these data types have unique challenges. Time series data contains high dimensions and massive amount of features with a huge percentage of irrelevant or noisy data but the ability to dissever what is noise or not is a huge problem and most data mining algorithms do not scale well to time-series data (Lin, Williamon, Borne, & DeBarr, 2012). Another challenge is the need to preserve the temporal structure of the original time series sequence, a problem studied in temporal pattern recognition research (Combi, Keravnou-Papailiou, & Shahar, 2010). In this section we provide an overview on works that have worked in this field.

Methods in Temporal Pattern Recognition

This section provides a discussion on prior works in the domain of temporal pattern recognition specifically those that have utilized time series data. Selection of papers on google scholar and other internet sources utilized a combination of key words i.e. “temporal patterns”, “time series mining”, “temporal data mining”, “mining temporal patterns”, “time series patterns”, to maximize the potential of finding articles relevant to pattern recognition where temporal representation was a main factor. A tabulation of the different methodologies is presented in Table 3.1.

Sung and Priebe (1988) indicate that effective information processing involves the consideration of its relative time-order. The researchers describe a system that takes sequences categorized to produce a Gaussian classification that represents temporal data using moving average and moving covariance (Sung & Priebe, 1988). They claim the system is able to detect out of order sequences. Critical intervals to define relationships between pairs of time events called T-Patterns were introduced in (Magnusson, 2000). Two years later, (Borrie, Jonsson, & Magnusson, 2002) used the T-patterns as the basis for the analysis of temporal patterns and their application in analyzing time-based events and real-time behaviours in sports.

A revisit of the T_patterns is detailed in (Sarah, Pauwels, Tavenard, & Gevers, 2010) with a focus to improve the time complexity of the algorithm by reducing the number of steps to iterate on pattern search.

Others have used neural networks. A survey on neuro network systems is detailed in (Chen, 1990) The researchers noted that statistical based pattern recognition methods rely on probability densities and are not able to handle nonlinear decision boundaries. However as noted in Chen (1990) the fact that neuro networks exhibit some black box computations that are hard to explain make it hard to transit effectively in the critical setting.

Temporal complexity measures are detailed in (Shmulevich & Povel, 2000) where the idea that a temporal pattern can be described based on elaboration of small patterns that occur simultaneously was defined early on by (Tanguiane, 1994). Although this approach proved

to be effective in application to music chords, it's not clear how this can be applicable huge volumes of time series data streams.

The use of dimension reduction techniques in time series has been explored in (Dash & Liu, 1997) using feature selection. However, without taking into account the element of time, variable connections among features is lost.

Quantitative intervals are described in (Guyet & Quiniou, 2008) who utilizes the classical Apriori algorithm for identification of temporal sequences represented as hyper-cubes and the temporal information is processed using density estimation by distribution of events across intervals.

The notion of discovery of spatial temporal patterns is described in (Mohan, Shekhar, Shine, & Rogers, 2012) where the researcher finds subsets of event-types which are located together and with serial occurrence. The challenge of handling computational complexity and statistical interpretation are noted as difficult to balance in pattern recognition (Mohan, et al 2012).

Other researchers have used connectionist models enhanced with back propagation techniques to reduce the amount of information for recreating the original structures, (Mozar, 1989).

The work in (Laxman & Sastry, 2006) details a comprehensive survey on temporal data mining where various techniques in data mining have been utilized in mining patterns from

large volumes of data from clustering, search and retrieval to pattern discovery of sequential patterns or episodes and incorporation of time constraints.

Use of clustering on temporal data mining has been utilized in various works as noted in the survey by (liao, Chu, & Hsiao, 2012) where most of the works cited used distance based techniques to measure similarities for classification. Liao et al (2012) notes the lack of correlation among time samples. Distance based outlier detection in combination with auto regression modelling are utilized in (Gul & Carbas, 2009)

Several applications in the real world on mining temporal patterns (Kruse, Steinbrecher, & Moewes, 2010), (Kruse, 2010). Supervised feature selection for dimension reduction is proposed in (Liu, Wu., & Zhang, 2014) using information entropy theory. (Marchione & Johnson, 2013) introduced a spatial temporal pattern discovery in maritime piracy while (Hui, Chen, Weng, & Lee, 2012) proposed incremental approach for pattern discovery using frequent sequential pattern mining.

A common theme among researchers in this domain is the need to identify temporal patterns from volumes of time-oriented data, however, the challenges remain on effective representation on patterns and the ability to scale systems to handle huge volumes of continuous data streams. The computational complexity of effective pattern generation is a significant problem. Techniques in data windowing could be utilized in discovery of patterns on user-defined widows rather than trying to identify all possible patterns on large volumes of data. This approach would allow discovery of patterns among data windows and then applied to real-time data streams.

Table 3. 1: Varying Techniques in Pattern Recognition

Researchers	Methods of Temporal Pattern Recognition	Problem Solved
Sung and Priebe (1988)	Gaussian classification to represent temporal data	Effective information processing
Magnusson (2000)	Critical Intervals	Relationships among event pairs
Borrie et al (2002)	T-Patterns	Temporal pattern analysis
Salah et al, 2010	Improved T-Patterns	Improve time complexity of algorithm in Borrie et al, 2002
Chen (1990)	Neuro Networks	Statistical pattern recognition
Schumulevic and Povel (2000)	Temporal complexity measures, enhanced the algorithm in Tanguane, 1993	Small pattern elaboration
Mohan (2012)	Discovery of event types closely located with serial occurrence	Spatial temporal patterns
Mozer (1989)	Connection models	Reduction of information to recreate original structures
Laxman and Sastry (2006)	Review on temporal data mining models	Systems applied to mining patterns from large volumes of data
Liao (2013)	Clustering on temporal data mining	Distance based techniques to measure similarities
Liu et al (2014)	Information entropy	Dimension reduction
Marchione and Johnson (2013)	Spatial temporal discovery	Application in maritime piracy

3.5 Pattern Recognition Research in Critical Care

The potential for understanding patterns from clinical data has been studied extensively and in this section an overview of works that have presented research in this domain is provided.

First an overview of works that have utilized statistical pattern recognition methods is presented.

Detection of abnormal drop in arterial blood pressure using HMM is described in (Singh, Tamminedi, Yosiphon, Ganguli, & Yadegar, 2010). A Probabilistic model for early prediction of abnormal clinical events using vital sign correlations in home-based monitoring (Abdur & Ibrahim, 2016).

Nearest neighbor technique for generating small prototypes that have similar properties are used for classification, (Pekalska & Duin, 2001). Another approach utilizing k-Nearest neighbour is described in (Mai, Tim, & Rob, 2012) for diagnosing heart disease.

Using data captured from elderly patients, Bayesian classification was presented in Inibhunu and McGregor (2017) to determine features contributing to adverse events such as hospital admission or emergency room visit. A follow-up on their research identified decision trees as the best feature identifier resulting to better accuracy of predicting adverse events compared to bayes classification (Inibhunu & McGregor, 2018).

Decision trees are also utilized to identify factors that are considered by healthcare providers in ICU when making clinical decisions in care for patients (Berney, et al., 2017).

A combination of multiple pattern recognition methods which included decision trees, support vector machines and KNN are described in (Reza, Tanvi, & William, 2019) where researches proposed a method for predicting mortality using heart signals captured from patients in ICU.

Identifying subgroups of ICU patients with similar clinical trajectories and clinical needs is attempted with clustering as described in (Vranas, et al., 2017). Automatic detection of

patterns on ECG data using neuro networks was explored in (Sternickel, 2002). The system includes two phases; feature extraction using dynamic time warping and classification by selecting 10 to 20 patterns on 10 min sequences. The researchers do not look at varying lengths on patterns but rather used a static window (Sternickel, 2002).

The use of support vector machines is attempted in (Lin, Xie, Hu, & Kong, 2018) to predict acute injury in ICU. Another use of SVM attempts to build a classifier for prediction of mortality in ICU by utilizing patient vital signals (Amer, et al., 2019).

The methods discussed above have attempted pattern recognition using adult population vital status data such as heart rate and except the work in (Inibhunu & McGregor, 2018) all the other systems utilize vital status data from MIMIC public database in (Saeed, Lieu, Raber, & Mark, 2002).

In the term and preterm infant population, algorithms for recognizing and characterizing neonatal spells have been developed for the Artemis platform using the real time processing capabilities of IBM InfoSphere Streams (Thommandram et al, 2013). The algorithms rely on the ability to detect relative changes in heart rate, blood oxygen saturation, and pauses in respiration in real time. Before clinical testing, the functional behavior of each algorithm was assessed against 24 hours of clinically annotated physiological data to assess the accuracy of the algorithms (Pugh, Thommandram, Ng, & McGregor, 2013) (Thommandram, Eklund, McGregor, Pugh, & James, 2014).

3.6 Summary on Pattern Recognition

The ability to discover patterns from data is a complex process and various methods have been proposed from how to detect, represent and classifying patterns to modelling recurring patterns on new datasets. Techniques for representing patterns includes data structures and proximity measures. Pattern recognition has been applied to varying data types, structured and non-structured. On structured data, time series is the dominant data in temporal pattern recognition research given its reliant on the time aspect which results to huge volumes of data generated and a vast majority of it, is noisy (Dougherty, 2013). Dimension reduction principles such as feature selection and feature extraction have been utilized in trying to reduce the amount of data processed during pattern recognition.

Feature extraction identifies some key attributes on data that can be used to represent the data in a higher granularity while feature selection aims to look at transforming a subset of the raw data. There are potential limitations exhibited in each technique as noted in (Dougherty, 2011).

3.7 Key Findings

Statistical and Tactical pattern recognition methods derived from mathematical theory and that have been utilized extensively in handling time series data are presented in this review. Statistical methods evaluated includes the use of bayesian probabilities, hidden markov models, clustering, neural networks and support vector machines.

In Syntactic pattern recognition special attention was given to temporal pattern recognition and its application in critical care. This review notes that some statistical techniques have been applied for temporal pattern recognition. Some of the techniques used includes; sequence of intervals from medical records (Hui, 2015), Nearest neighbour (Pekalsak and Diun, 2001), Automative ECG detection in (Sternickel, 2002), distance based outlier detection (Gu & Catbas, 2009), quantitative intervals in Guyet and Quiniou (2008), spatial temporal patterns in (Mohan 2012), connectionist models (Mozer, 1989) and the use of spatial temporal in piracy discovery in (Hui, Chen and Weng) and classifying of neonatal spells in (Pugh et al, 2013).

The review on the various techniques applied for temporal pattern recognition demonstrate promising results on generating patterns when methods are applied to small datasets, however, they lack techniques for quantifying temporal relationships in data and do not address the challenges when faced with large volumes of data generated continuously from complex systems, especially in critical care such as the big data framework described in (Inibhunu, et al., 2019). Processing such data makes it extremely difficult to detect, represent, classify temporal patterns and build classification systems. As such, research is still needed in

temporal pattern recognition to ensure that potential knowledge hidden in time series clinical data is discovered and presented to health care providers and clinicians as they make important decision about care of patients when they need it most.

In conclusion, this review has provided the background needed to; (a) understand the techniques and methodologies utilized in pattern recognition, (b) identify some of the challenges in handling massive volumes of data in pattern mining, and (c) highlights some of the techniques in dimension reduction and similarity measures that can be adopted for filtering and pruning out irreverent or repetitive data.

Chapter 4

4 A Review on Physiological Data

This thesis presents a method for knowledge discovery utilizing data collected from monitors and sensors that continuously or discontinuously monitor critically ill patients in clinical domains. To accomplish this, it's necessary to have a clear understanding of the nature of the data collected about patients and the types of research undertaken utilizing physiological data. This chapter presents a brief overview of two clinical domains where patient data has been utilized to demonstrate the contribution of this thesis. The first is neonatology and the other is in telehealth where elderly patients participate in a remote patient monitoring program.

4.1 An Overview of Neonatology

Neonatology is a field in medicine that is involved in the study, prevention, teaching and the provision of medical care to premature and ill term infants in NICU (American Academy of Pediatrics, 2019). As noted in (McGregor et al, 2009), a preterm or ill term infant in a NICU could have many conditions which may not be obvious to physicians until it's too late to save a fragile life. Such a problem is neonatal infections, a common cause of death for premature infants. Most infants acquire bacteria infections around delivery time while others are exposed to bacteria while receiving care in a NICU. Accurate diagnosis of the presence of bacteria remains a challenging task for clinicians due to the frequent presence of multiple comorbidities. The ability to utilize patient physiological data for early detection of neonatal

infections could greatly improve the health outcome of a critically ill infant (McGregor et al, 2009).

4.2 Physiological Features for Diagnosis of Diseases

The need to utilize a combination of physiological data from a patient's data such as heart rate, blood pressure, RR, oxygen saturation in the attempt to detect conditions on preterm infants has been studied extensively (White-Taut, 2003; Sadleir, 2009; Krueger, 2009; Tang, 2010; Weissman, 2012; Thommandram et al 2013). This section provides a brief overview of each of these features and highlight the potential for utilizing these to detect unknown conditions such as neonatal sepsis. A tabulation of the physiological features utilized in the varying preterm infants research is presented in Table 4.1.

Heart Rate

The measure of the number of contractions of a heart per minute is referred to as heart rate which varies in response to the body's cells needs for oxygen absorption and nutrients under varying conditions. In order to respond rapidly to the changing requirements of the body's tissues, the heart rate and contractility are regulated by the nervous system, hormones, and other factors (Gordan, Gwathmey, & Xie, 2015).

A resting adult has a heart rate which ranges between 60 – 100 beats per minute (bpm). A fast heart rate is called Tachycardia where heart rate is above 100 bpm and a slow heart rate is called bradycardia where heart rate is below 60 bpm. In deep sleep, slow heartbeats are around 40-50 bpm and this is considered normal (American Heart Association Guidelines, 2005). Normal range of a preterm infant is 100-170 with tachycardia identified when bpm are

beyond 180 (Singh, Garekar, Epstein, & L'Ecuyer, 2005) while bradycardia when bpm drops below 80 (Chandrasekharan et al, 2017). When the heart's beating patterns increases, this is considered an abnormality and could indicate present of an infection (McGregor et al, 2009).

Electrodes placed on the skin are used to measure the electrical activity of the heart over a period of time. This process is called electrocardiography (ECG). The electrodes detect any electrical changes from the heart muscles electrophysiological pattern during each heartbeat.

The degree of variation on the time between heart beats is called heart rate variability (HRV) measured by the beat-to-beat interval variation. Reduced HRV has been shown in research to be a predictor of acute myocardial (heart attack) and could indicate neonatal sepsis (Frencken et al, 2018).

As noted in (Hanna, et al., 2000) patterns of HRV are closely correlated to clinical outcomes in many pathological situations. Their study aimed at characterizing the relationship between HRV, length of stay and diagnosis of cerebral palsy. The work in (Tory, et al., 2003) aimed at investigating the cardiovascular autonomic function in juvenile uremia by analysing of HRV. The researchers looked at heart rate variability for cardiac mortality in hemodialysis patients. The results showed a clear difference on HR between patients who had dialysis compared to those who had a transplant (Tory, et al., 2003).

Fetal heart rate patterns are analysed in (Goncalves, Bernardes, Rocha, & Ayres-de-Campos, 2007) to understand association with fetal behaviour states in the antepartum (prenatal) period such as body and eye movements. The researchers indicate that

understanding and interpretation of fetal heart rate (FHR) could be helpful in diagnosing pathological fetal conditions.

The investigation of complex heart rates using non-linear based methods is details in Weissmann et al (2012). Their work aims at examining the dynamic changes on heart rate fluctuations and variability during painful stimulus in newborns. HRV indicated the interplay of sympathetic and parasympathetic nervous system and a significant change in these can be attributed to many pathological conditions such as impending intraventricular hemorrhage and sudden infant death syndrome, Weissman et al (2012).

An evaluation of preterm infants' heart rate characteristics (HRC) is studied in (Fairchild, Sinkin, Davalian, Blackman, & et, 2014) for detection of neonatal sepsis.

Respiratory Rate

Localization of disorders can be determined by Identifying abnormal breathing patterns and help refine diagnosis (Yuan, Drost, & Mclvor, 2013). Normal breathing includes phases of inspiration and expiration which occurs in chronous thorax and abdomen movement. An infant less than one year of age has 30 to 60 breaths per minute. A respiratory rate consistently greater than 60 beats per minute in an infant is abnormal.

As noted in (Yuan, Drost, & Mclvor, 2013) there are several abnormal breathing patterns:

(a) thoracoabnominal paradox; respiratory muscle dysfunction with increased work in breathing. The thorax and abdomen move in completely different directions at the same time.

(b) Kussmaul's breathing; increased frequency on tidal wave and often seen as gasping, this is normally associated with severe metabolic acidity in the blood and other body tissues.

(c) Apneustic breathing; this is breathing where each inspiration is followed by a prolonged inspiration pause and each expiration is followed by a prolonged pause.

(d) cheyne-strokes respiration: cyclical crescendo-decrescendo pattern of breathing followed by periods of central apnea. This type of breathing is seen in patients with conditions such as brain injury, carbon monoxide poisoning as noted in (Yuan, Drost, & Mclvor, 2013).

(e) Ataxic and Biot breathing; breathing with irregular frequency with tidal volume followed by unpredictable pauses in breathing or periods of apnea.

(f) Biot breathing: high frequency and regular tidal volume combined by apnea periods.

(g) Agonal breathing: irregular pattern and sporadic breathing with gasping normally seen in dying patients before the terminal apnea.

(h) Central sleep apnea: brain temporarily stops to send signal to the muscles that control breathing.

As noted in (Taylor, 2015), respiratory rate is vital but mostly overlooked nonetheless is seen as a critical measure that could help identify early deterioration of patients.

Blood Pressure

The pressure of blood circulating on the walls of blood vessels is referred to as blood pressure. This pressure is normally expressed in two measures, maximum during one heart

beat (systolic) and minimum in between two heart beats (diastolic) and measured in millimeters of mercury (mmHg). A normal blood pressure is recording with systolic less than 120 mmHg and Diastolic less than 80 normally abbreviated at 120/80 mmHg. Blood pressure that is higher than this range is categorized as High Blood Pressure hypertension (HBP) which is a leading cause of conditions such as stroke, vision loss, heart failure, heart disease or kidney disease failure (Baton, Li, Newman, Das, & al., 2014).

Low blood pressure is associated with hypotension, this is normally associated with systolic less than 90mmHg or diastolic less than 60 mmHg. Severe low blood pressure can impact vital organs including the brain which needs oxygen and nutrients resulting to life threatening condition known as shock (Fanaroff & Fanaroff, 2006).

In neonatal research, blood pressure has been studied extensively as researchers seek to understand the impact of blood pressure to a patient's conditions including identifying which is viable approach to collection of the measures from a patient. Fanaroff J. and Fanaroff A. (2006) indicate there is a potential link between hypotension and neonatal diseases such as IVH, and argue that there is no clear definition of hypotension and hypertension in the neonatal period.

On collection of blood pressure, Fanaroff J. and Fanaroff A. (2006) indicate that simple non-invasive and reliable means if not available, direct invasion methods that include; arterial or peripheral artery catheter can be used and are considered more optimal. In this case, mean pressure rather than systolic pressure is used as it is believed to be free of artifacts i.e air bubble, thrombi or resonance. However, the downside of this method is the risk of thrombus

formation, hemorrhage or infection. Non-invasive methods include oscillometric and Doppler techniques that are automated. Oscillometric measurements seems to be accurate within normal range but at lower levels it is seen to overestimate blood pressure.

Fanaroff (2016) noted that, in the term neonate, abnormal blood pressure such as hypertension is often defined as systolic and /or diastolic blood pressure more than 2 standard deviation (2SD) above the mean values. The researcher indicated that gender is also a factor where males weighing <1000g known as extremely low birth weight (ELBT) have lower blood pressure than females. However, for preterm infants, this is hard to define. About 20-50% of ELBW infants are diagnosed with hypotension while Hypertension is seen in about 3% in NICU admissions (Fanaroff, 2006).

Batton et al. 2014, described the dynamics of evolving blood pressure in extremely preterm infants by examining Arterial Blood Pressure (ABP) on ELBW preterm infants. Systolic, diastolic and mean ABP were collected.

Blood Oxygen Saturation

The percentage of hemoglobin molecules in the arterial blood that are saturated with oxygen is recorded as blood oxygen saturation. This is referred to as SaO2, however when this is determined by a pulse oximetry, its referred to SpO_2. The readings vary between 0 to 100% where a normal healthy adult reading range from 94%-100%. Target ranges are lower for preterm infants but not consistent internationally. In newborn application of pulse oximetry sensor is recommended to be placed in the right hand (pre-ductal) where SpO_2 from right hand are more representative of brain oxygenation. SpO_2 between (right hand =

pre-ductal and Left hand=post-ductal) and vary by 25%. Levels below 75% in new born may indicate abnormalities (Askie, 2003).

Neonatal research on the impact of low blood oxygen saturation is studied extensively such as the work in (Tin, Milligan, Pennefather, & Hey, 2001) where researchers looked at the relationship between preterm babies that develop retinopathy of prematurity and the survival rate with or without cerebral palsy at age one. Askie et al's (2003) work on target and outcome of oxygen saturation on preterm infants noted that an improved survival of extremely preterm babies has been associated with increase in chronic lung disease of infancy such asbronchopulmonary dysplasia. The researchers note that infants with such conditions have higher rates of oxygen consumption and lower oxygen saturation than infants without the conditions.

There is a potential to combine the varying data types i.e. blood pressure, heart rate, blood oxygen saturation, respiratory rate which could be used to understand any hidden patterns that might exists and could be associated with life threatening conditions.

4.3 A Brief Overview on Neonatal Sepsis

There are a range of conditions that a preterm or ill term baby could have which may not be obvious to physicians until it's too late to save a life (McGregor et al., 2009). Such a condition is neonatal sepsis, which is a bacterial infection in neonates known as neonatal sepsis. Neonatal sepsis is a leading cause of morbidity and death among infants admitted to NICUs and for those babies born premature, the incidence of infections is 3 to 10 times more than term babies (Hoogen, 2009). Neonatal sepsis affects the bloodstreams, organs and

meninges. Diagnosis of neonatal infection is difficult to establish and is often confirmed when clinical signs of the infection are present and positive blood cultures.

There are two categories of neonatal sepsis based on the time of condition onset; early-onset and late-onset. Early-onset is sepsis that occurs in the 72 hours of life, generally between 48 to 72 hrs after birth and late-onset sepsis are the infections that usually occur

Table 4. 1: Varying Research Utilizing Physiological Features from Preterm Infants

Researchers	**Physiological Features**	**Research Highlights on Preterm Infants**
White-Taut (2003)	Heart Rate Variability	Highlight clinical indicators or respiratory and cardiac abnormalities are two common form of premature IVH
Krueger et al, (2009)	Heart Rate Variability	Relationship between Heart rate and IVH
Weissman (2012)	Heart Rate Variability	Dynamic changes on heart rate during pain simulation in newborn
Thommandram (2013)	Heart Rate Variability	Detection of Sepsis
Hanna (2000)	Heart Rate Variability	Detection of IVH
Goncalves et al (2007)	Heart Rate Variability	Association between heart rate and fetal behaviour state; body and eye movements
Fairchild (2014)	Heart Rate Variability	Detection of Sepsis
Yuan (2013)	Respiratory Rate	Identifying abnormal breathing patterns
Taylor (2015)	Respiratory Rate	Identify patients early deterioration
Fanaroff (2006)	Blood Pressure	Highlight potential link between hypertensions and IVH
Milligan (2000)	Blood Oxygen Saturation	Relationship between preterm babies and retinopathy of maturity
Askie et al (2013)	Blood Oxygen Saturation	Association between blood oxygen saturation and chronic lung diseases

after 72 hrs of life. Center for disease controls terms any infection that occurs during admission to the hospital as nosocomial infections, that is hospital acquired infection, and this occurs in 20 to 25 % of preterm infants who weigh <1500g and the rate of infections increases as the gestation age and birth weight decreases (Hoogen, 2009).

Clinical symptoms of neonatal sepsis are non-specific, subtle and most of the time interpretations of clinical symptoms may be difficult as symptoms could be associated as pathophysiologies of other clinical conditions. There are risk factors associated with invasive methods needed to confirm sepsis incidences as clinical diagnosis can involve evaluation of blood cells, spinal fluid or other specimen to confirm sepsis cases. This is a process that involves complex interventions which are identified as independent risk factors for infections (Hoogen, 2009). Antibiotics are the most frequent used drugs in NICUs for treatment of neonatal sepsis and treatment needs to be administered immediately when sepsis is suspected. As such, early detection of sepsis can have a life long impact on the health of an ill infant.

Non invasive methods for early detection of neonatal conditions remains a continuous research area that evaluates physiological data such as blood pressure, HR, RR, blood oxygen saturation to understand diseases or conditions that may be present in patients. As noted in section 4.2, different studies have tried to use either a specific physiological feature or a combination of one more features to detect conditions in neonates. This includes works such as evaluation of HRV patterns described in Hanna (2000), Tory et al (2003) and Weissmann et al (2012), studying breathing patterns to understand abnormal behaviours in (Yuan, Drost

and Mclvor, 2013), blood pressure circulation in Fanaroff (2006) and blood oxygen saturation in Tin, Millligan and Hey (2000). A combination of three features is detailed in (Thommandram et al, 2014) where the evaluation of heart rate, respiratory rate and blood oxygen saturation is explored in an attempt to classify neonate spells. McGregor et al propose non-invasive methods for early detection of neonatal sepsis utilizing multiple patient physiological data streams (McGregor, Catley, & James, 2012). The researchers demonstrated the potential for identifying patients at risk of neonatal sepsis by analysing the variability of the patient's heart and respiratory rates.

As there are multiple physiological data streams captured from neonatal patients, there is potential for utilizing such data to discover patterns that may be exhibited in patients who are at risk of neonatal sepsis. Further more, patient's data could highlight time-oriented patterns before during and after a potential neonatal infection. This is information if captured early, can be utilized in making decision on care of patients in NICU.

4.4 A Brief Overview on Elderly Care with Remote Patient Monitoring

Another population that could benefit from analysis of physiological data is the elderly patients especially those with chronic conditions such as heart failure (HF) and chronic obstructive pulmonary disease (COPD). In particular, remote collection and analysis of such patients data could aid in effective decision making as a proactive and earlier warning of the need for intervention. Remote Patient Monitoring (RPM) is such an approach that is considered to have potential to improve the quality of life on patients diagnosed with cardiac

conditions as well as lead to reduction on healthcare costs through a reduction in emergency room visits and hospital admissions (Gorst, Armitage, Brownsell, & Hawley, 2014).

Conditions such as HF and COPD contribute significantly to the burden placed on patient's quality of life as well as substantial costs to healthcare systems through repeated emergency room visits and lengthy hospitalization. The direct cost associated with caring for patients diagnosed with HF and COPD in 2012 was estimated at over $50 billon in USA (Gorst, Armitage, Brownsell, & Hawley, 2014). In addition, Gorst et al. indicated that escalation of COPD and HF has a direct link with an aging population. In Canada, it is estimated that by 2036 1 in 4 Canadians will be a senior aged 65 and over and 85% of that population will have some type of chronic condition (Ward, Schiller, & Goodman, 2014). To this respect, there is a need for effective management of chronic conditions like COPD or HF facilitated through remote patient monitoring services in telehealth (Gorst , Armitage, Brownsell, & Hawley, 2014).

As noted in a publication resulting from this research, earlier research on predictive analytics utilizing RPM data in (Inibhunu, Schauer, Redwood, Clifford, & McGregor, 2017), potential for use of RPM in providing care for patients was recognized by several healthcare organizations. In particular, a partnership was formed between AlayaCare, a software company for community care providers and We Care, a division of CBI Health Group Canada's largest provider of rehabilitation and home health services to produce innovative approach utilizing telehealth to reducing adverse events leading to leading to emergency room (ER) visits and hospital admissions (We Care & Alayacare, 2013).

In 2015, these two partners then started collaborative research with Southlake Regional Health Centre and Health Informatics Research at Ontario Tech University in order to perform a remote patient monitoring research study whose focus was on three areas; (a) risk score prediction using predictive analytics utilizing data collected from patients participating in RPM study, (b) validate that remote patient monitoring and telehealth coaching of chronically ill patients reduces the frequency of adverse events such as re-hospitalizations, ER visits, and falls an (c) validate the service delivery model whereby an acute care facility, in partnership with the local HealthLink, engages a private sector service delivery partner like CBI/WeCare to deliver the RPM/Telehealth services.

To facilitate this process (Inibhunu, McGregor, Schauer, Redwood, & Clifford, 2017) utilized predictive modeling to determine key factors that are significant determinant to hospitalization and multiple ER Visits. The results showed that gender, past medical history and vital status are key factors to hospital admissions and ER Visits. Men were more likely to have hospital admissions compared to women and that the probability of presence of past medical history was statistically significant for seniors age 65 and over. Additional models that included a flag to indicated period before during or after an adverse event showed SpO_2 and pulse rate as significant predictors of an adverse event. However, it was not clear how the time flag impacted the predictive model and further analysis as proposed in this thesis is required to understand the patterns before, during and after an adverse event.

4.5 Key Findings

High frequency data collected from monitors and sensors about patients' vital status in intensive care units together with home-based monitoring has potential to provide valuable insights which can be crucial for making decisions about care for preterm infants. Similarly, monitoring the vital status of an aging population especially those with chronic diseases can potentially reduce the multiple emergency room visits and hospitalizations if patients are provided with information that might help them make informed decisions on appropriate cause of actions. However, this analysis is not trivial with huge volumes of data captured every second and the ability to collect and understand any hidden relationships in the data that might be crucial for decision making is a central challenge.

Converting research that utilizes physiological data streams to a pattern recognition research problem, there are several questions one can try to ask, such as; is it possible to categorize the changes in heart rate in combination with other physiological features collected at the same time from a patient such respiratory rate and blood oxygen saturation to understand any temporal patterns exhibited in the data? Is it possible to distill differences in patterns before, during and after a neonatal diagnosis? Can any patterns discovered be used to build a classification system for potential diagnosis and characterization of unknown conditions before they happen?

The ability to understand the relationship among time series data points generated from a patient's physiological features such as heart rate, respiratory rate, blood pressure and transcutaneous blood oxygen saturation can be translated to a computational problem

known as a pattern recognition and classification problem. Humans are able to identify objects that are similar using various measures; type, size, shape, color among others and given the sequence of occurrence of at the measures, one can deduce that as a pattern. This is easy on a simple set of objects or data sets, however on complex data sets such as those captured continuously in bedside monitors, the ability to deduce any patterns from such data is not a simple exercise given the amount of physiological data generated in a single second.

Integration of monitoring systems and increased accessibility of high frequency data allows the potential for advanced analysis of patient data which can be utilized in characterizing patterns of disease. Results of such analysis could offer opportunities to identify patients who are at high risk of conditions and detection of the onset of disease prior to clinical signs manifestation (Huvanandana, et al., 2017).

With respect to providing telehealth services to elderly patients with chronic conditions such as HF or COPD, several models are presented through contributions related to this thesis in Inibhunu et al, (2017), where initial models assessed if gender, patient clients medical history and vital status are significant determines to hospitalization. However, further understanding of pulse and SpO_2 behaviour in the time preceding an ER visit or a hospital admissions could uncover new pathophysiologies. This process can be facilitated using temporal abstraction and pattern recognition models. McGregor has patented a temporal data mining method known as the Service-Based Multi-Dimensional Temporal Data Mining ($STDM_0^n$) (Canada Patent No. WO2011009211 A1, 2011).

There is potential in utilizing temporal abstraction and deriving temporal patterns to understand the underlying temporal relationships on vital status in data collected from patients participating in RPM program in combination with other data sets in order to get a complete patient flow. Such discovery can highlight when an elderly patient is at risk of an adverse event, this is information that can be utilized for provision of appropriate care for the patient.

Chapter 5 discusses adoption of pattern recognition techniques for developing an effective process that seek to understand relationships in high frequency time series data streams and their application in health informatics for analysis of physiological data streams.

Chapter 5

5 Proposed Research

This chapter presents details the thesis contribution to propose a method to detect and represent relationships that may exist in temporal abstractions (TA) and temporal patterns (TP) derived from time oriented data. As research continues to advance the technologies and processes in health care, so is the exponential growth of data generated from various complex systems in clinical settings. In particular, in critical care, monitors and sensors recording a patient's physiology generates large volumes of time series data every second. As concluded in chapter 2, there is currently no ability to understand any relationships that may exist between multiple time oriented data streams that are acquired in parallel. Additionally, it is a challenge to discern any hidden knowledge from such high frequency data which if discovered promptly could be utilized for care of the patients as concluded in chapter 4.

As noted in chapter 1, this thesis is motivated by the question: Does the detection of changes in patterns that may be exhibited in time series physiological data streams lead to discovery of any hidden relationships that may exist in the underlying data? The premise of this thesis is that the discovery of such patterns may quantify previously unknown hidden relationships that may exist among the many sequences of time series data streams. In addition, those discovered patterns could be utilized to build a classification system with which to classify any new data streams. In clinical care, discovered relationships on

physiological data could be utilized in detection of onset of conditions, aid in classifying abnormal or normal behaviours or derive patterns of a patient's future state.

Researchers have attempted to provide varying techniques for utilizing clinical data, however, there are several shortcomings identified in the literature as discussed in chapter 2. In particular, current systems lack the ability to handle the variety, volume and velocity of continuous data streams especially when trying to derive any hidden relationships and patterns that may exist in the data (Combi et al 2010). As noted in (Shahar, 1999) and (McGregor, 2009), physicians who make critical decisions on diagnostic and therapeutic modalities are overwhelmed by the amount of data generated continuously at the bedside. In an attempt to address these challenges, several systems have been proposed for deriving hidden patterns from time-series data (Shahar et al (1999) , Belazzi et al (2005) , Chaovalitwonge (2007) Hunter et al (2012), Lui et al (2014)).

However, the current approaches are computationally expensive as extensive data loads need to be processed at each iteration of the knowledge discovery process. In particular, as noted in Chapter 1 (Table 1.1), the amount of data generated in an hour from a single patient in the NICU is in millions of records and this can grow to billions of data within 24hrs.

As discussed in chapter 2, trying to analyse such data to understand any temporal patterns that may exist is not a trivial process and the current algorithms to process such data become like a black box to end users. Instead manual annotations from monitors are recorded every 30-60 minutes (McGregor et al., 2012). As discussed in chapter 4, this process

is ineffective for the detection of life threatening conditions such as sepsis where early discovery can lead to prompt treatment and potentially save a life.

To address the current limitations, this thesis proposes to develop a method to detect and represent relationships that may exist in TA and TP from time-oriented data. The development of this method involves extensions to the temporal data mining model, CRISP-TDM detailed in (Catley, 2009) and the multidimensional temporal abstraction and data mining framework, $STDM_0^n$ detailed in (McGregor, 2011) for real-time knowledge discovery, and this extension contributes to the following research domains:

(a) Computer Science: Advancement in knowledge representation and discovery by proposing an innovative method that utilizes a phased fusion of machine learning principles in pattern recognition, dimension reduction and data mining for frequent pattern mining and classification.

(b) Health Informatics: Advancement in frameworks for real-time knowledge discovery using physiological data streams and classification of events and episodes in a clinical context.

Details on contribution to computer science are discussed next while contributions to health informatics and application to medicine are discussed in chapter 6.

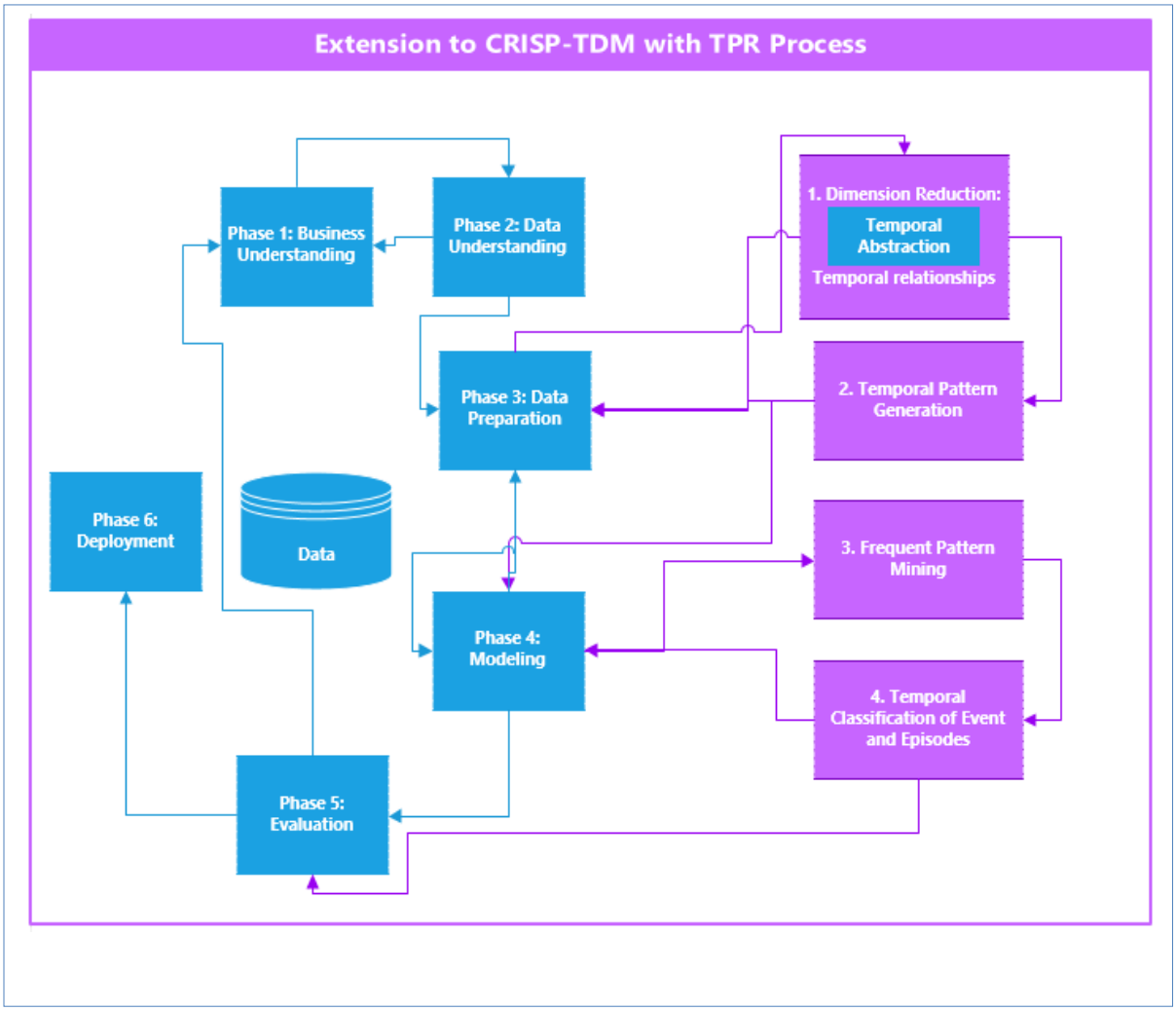
5.1 Research Contributions to Computer Science

This research extends the CRISP-TDM^n model in (Catley, 2009) by introducing four components with functions for discovery of relationships among abstractions, generating temporal patterns, mining frequent occurring patterns and building a classification system.

This is a novel knowledge discovery model that uses a fusion of some of the principles described in chapter 3 such as; dimension reduction, similarity measures, temporal abstraction, pattern recognition and frequent pattern mining for efficient generation of temporal patterns in real-time data streams.

The proposed model is presented in Figure 5.1 where blue shows the existing components and purple indicates the proposed extensions. Components 1 and 2 extends the data preparation phase while component 3 and 4 extends the modelling phase. The proposed components include techniques for dimension reduction with temporal abstraction, deriving temporal relations, discovery of temporal patterns, mining frequent patterns and using the resulting outputs for developing a classification system. Details on the proposed components are provided next.

Figure 5. 1 : Proposed Extension of CRISP-TDM^n model.
New model incorporates additional components for dimension reduction, derivation of temporal relations, temporal patterns, frequent pattern mining and classification. Blue color shows the existing model while purple shows the proposed enhancements.



Component 1: Dimension Reduction: This component includes modules for generating temporal abstractions and deriving temporal relationships. This process builds on the principles for converting raw data into temporal abstractions as described in (Catley et al, 2009). Generation of temporal abstraction in this thesis is accomplished by formulating a process as a finite state machine using hidden markov models augmented with data driven

clusters as described in section 5.2.1. After generation of temporal abstractions is the derivation of temporal transactions and temporal relationships as described in 5.2.2 and 5.2.3.

Component 2 - Temporal Pattern Generation: This component adds functions for deriving temporal patterns by aggregating all the temporal relationships generated in component 1. This process is accomplished by formulating temporal patterns as temporal paths as described in section 5.2.4.

Component 3 - Identification of Frequent Patterns: This component enables identification of frequent patterns by utilizing the principles described in association rule mining in (Agrawal & Srikant, 1995) and event sequence mining in (Liu, Wu., & Zhang, 2014).

Several works have been proposed in the literature on frequent pattern mining and have seen some success such as in market basket analysis (Laxman & Sastry, 2006). However, the primary constrains of frequent pattern algorithms remain, such as time and space complexities and ability to analyze significant volumes of data. With respect to processing real-time data streams, these challenges become even more complex. To address these challenges, this thesis uses thresholds as discussed in section 5.2.5.

Component 4 - Classification Model: The discovered frequent patterns are used to generate association rules which forms a dynamic classification system that can be used to generate new hypothesis on a given data stream. This process is described in sections 5.2.6 and 5.2.7. Additionally, all the frequent patterns and the association rules generated are stored in a temporal database emulating the temporal data storage mechanism in (Combi et al, 2012).

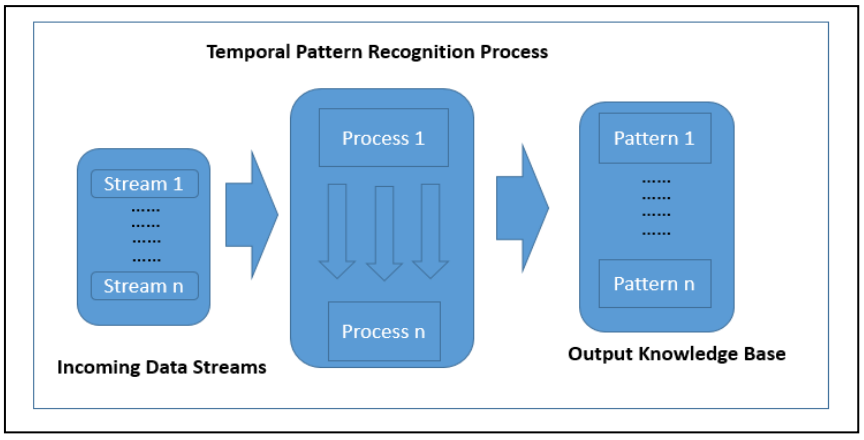
This approach allows scaling to real-time data streams and adapts to changing data landscape.

5.2 Methodology

Facilitating the enhancements to CRISP-TDM as described in section 5.1 is completed in this thesis within the computer science domain by introducing a Temporal Pattern Recognition Process (TPR Process) and an associated Temporal Pattern Recognition and Mining algorithm (TPRMine) which adopts a stepwise approach to temporal pattern discovery.

To facilitate such a knowledge discovery process, this thesis formulates the research questions in chapter 1 as a pattern recognition problem space thus introducing the TPR Process as shown in Figure 5.2 where incoming data streams are input to the multiple processes within the TPR Process and the outputs are multiple patterns that form a knowledge base.

Figure 5. 2 : Temporal Pattern Recognition Process



The execution of the TPR Process is accomplished by adopting a stepwise approach to temporal pattern discovery where first a scaled mathematical formulation of the incoming multiple data streams is completed as described next in section 5.2.1 and then passing the resulting output to subsequent steps which are described in sections 5.2.2 to 5.2.7.

5.2.1 Temporal Representation with Markov Models

Mathematical formulation of incoming multiple data streams is achieved in this thesis by utilizing the state model developed as part of this research and detailed in (Inibhunu & McGregor, 2018). To this respect, a problem space is modelled as a finite state machine representation where by for a given timeframe, a time series data segment transitions from one state to another based on some associated weights. Representation of time series data with weights is a form of dimension reduction strategy and in Markov chains this weight is normally based on some probability distribution as noted in (Inibhunu & McGregor, 2018). Adoption of Markov chains in this research is described as follows.

As noted in Chapter 3, a process can be modelled as a finite state machine using some weighted measures. As such, for a given process P, Pr is the probability of P being in some state $s_i \, \epsilon \, S$ and follows the Markov property where by, if P is a stochastic process then conditional probability distribution of future states of P is dependent only upon the present state, not on the sequence of events that preceded it. Therefore,

$$\forall i, 1 \leq i \leq j, Pr(s_i | s_1 \ldots . s_{i-1}) = Pr(s_i | s_{i-1}) \quad (2)$$

and each s_{ij} expresses the probability $Pr(s_j|\ s_i)$, and as per the laws of probability, all arcs leaving a particular state forms a stochastic vector with probabilities that add up to 1. A sample Markov chain is shown in Figure 5.3, where for a given time interval, a time series dataset comprised of multiple data streams transitions from one state to any state at random.

The arcs shows the potential of a transition and the text on the arcs represents the probability of the transition. For example the label x01 indicates that a process can transition from state s0 to s1 with the probability x01.

Figure 5. 3 : Probabilistic State Transitions with Hidden Markov Chains

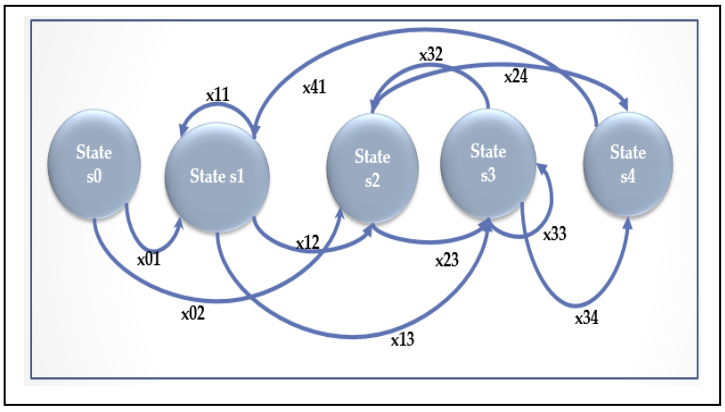


Figure 5.3 can represent the possible state a process can be in at a given time interval *T* and the length of *T*, $len(T)$ can be dynamic or fixed windows. The choice of $len(T)$ could have an impact on the overall performance of the HMM process. *T* can be a combination of multiple time segments(TS), such that $T = \{TS_1, TS_2, \dots, TS_n\}$ and each TS_i is comprised of a set of observations in *O*. For all *(i, j* $\in \mathbb{R}$), then $(TS_i, TS_j) \in T$ are two different time segments where,

$$TS_i = \{o_{i1}, o_{i2}, \dots, o_{im}\}, TS_j = \{o_{j1}, o_{j2}, \dots, o_{jk}\} \quad (3)$$

such that $(i \neq j)$, $(1 < m \leq k < n)$, $TS_i \neq TS_j$ for all observations $o_i \in O$.

With this, each segment $TS_i \epsilon T$ can be termed as a temporal process and elements within the segment can be characterized as a chain of linked events which randomly change (Gagniuc, 2017). Application of HMM to such a process allows quantification of the characteristics within that time segment with probabilities. To this respect, each time segment $TS_i \epsilon T$ is a temporal pattern discovery problem that falls into the HMM learning problem space in that, given a time series observation set O = $(o_1, o_2, .., o_i, .., o_j, .., o_n)$ and a set of possible states S = $(s_0, s_1, .., s_i, .., s_j, .., s_z)$, discover all possible hidden state sequences in *O* for a given time segment *TS* in time interval *T* (Inibhunu & McGregor, 2018). To facilitate this, the adoption of the Baum-Welch algorithm is applied by utilizing the Expectation Maximization (EM) algorithm to find the maximum likelihood estimate of parameters in a hidden markov model (Baum, 1972). As noted in chapter 3, the maximum likelihood estimate of the probability a_{ij} for some transition from state i to j is calculated by counting the total number of times the transition was observed.

For a given time segment the adoption of HMM generates the sets of probable states that a time series dataset can transition to and the inferred paths of such transitions. States themselves are unique and can be inferred from data or predefined per data domain.

This leads to the next questions, what are the most frequent transitions and paths leading to this transition in any period of time and, with multiple time series data sets, are there any hidden relationships among these state transitions?

5.2.2 Temporal Transactions

To answer these questions, this thesis introduces temporal transactions as a technique to characterise temporal aspects in time series data by formulating the process as a frequent pattern mining problem and adopt association rule principles. In particular, a temporal transaction is defined in this thesis as a finite set containing several elements i.e., {time interval, states and inferred probabilities} such that for a given time period T, there are a set of finite states that a given signal (data stream) can transition to.

Assuming a snapshot is taken at some timeframe in $T = (t_1, \dots, t_n)$, of a time series data D (see Figure 5.3). Let S = $(s_0, s_1, s_2.., s_i, .., s_j, .., s_{n-1})$ be a set of *n-1* states, such that, $s_i \cap s_j = \emptyset$, s_o is the initial state, s_{n-1} is the last state and $(0 < i < j < n)$. At time interval $(t_i, t_j) \epsilon\, T$, the time series data can transition to multiple states and these transitions are represented by a set of arcs $A = (x_{0i}, .., x_{i0}, .., x_{ij}.., x_k)$ $1<k\in\mathbb{R}$. Each arc in A is comprised of the hidden probabilities such that an arc $x_{ij} \in \beta_t(s_i)$ is the probability of transitioning to a state s_j from s_i. Within the time interval (t_i, t_j), all states $(s_i, s_j) \in S$ where $x_{ij} > 0$ are characterised as set of temporal transitions of the form $\{(s_o, s_i), (s_0, s_j), \dots, (s_i, s_j), \dots (s_n)\} \in S$. The length of a temporal transaction is determined by the number of states and arcs leading to a specific state. For example, a temporal transition of the form (s_0, s_2, s_1) indicates that a process can transition from the initial state s_0 to s_2 and then finally to state s_1.

5.2.3 Temporal Relationships

As there are many temporal transitions that a time series signal can take in any given time interval, then the temporal relationship can be derived as the behavior of state to state transition. This approach allows two relationships to be modelled in this thesis as follows.

Transitive Temporal Relationship (TTR): TTR is a relationship between states. If there is a path from s_i to s_j passing through some state s_l. This relationship is denoted as $\{s_i, s_l, s_j\}$ over some time interval $(t_i, t_j) \epsilon\, T$. Note that it's possible for inverse path from s_j to s_i, where 1 or more arcs leads to state j from i. $\{s_j, s_l, \dots, s_m, \dots, s_i\}$.

Equivalence Temporal Relationship (ETR): ETR is a transition from a state s_k to s_j s.t. $\{s_k, s_j\}$ and there is a reciprocal path $\{s_j, s_k\}$, then this can be inferred as an equivalence relationship. As these two relationships are augmented with time, we refer these as temporal relationships.

Generating such relationships over several time oriented data can result in multiple relationships with differing paths which introduces another aspect; how to define the start or end of a temporal relationship. This is a known problem in path finding research and some of the approaches suggested in the literature apply distance measures to group similar elements together or using predefined thresholds. In this research a time dependent windowing strategy is adopted which can be adjusted as needed. The idea is to quantify the patterns observed in a select window of time and as demonstrated in chapter 7 (Figure 7.4),

the size of the selected window can have a significant impact on the space and time complexity in computation of the proposed process.

5.2.4 Temporal Pattern

Let X be a set of temporal transactions comprised of temporal relationships $\{R\}$ in a given time interval $\left(t_{ij}\right)$ ∈ T where $1 \leq\ i \leq j \leq n$ then

$$\mathrm{T} = \left(t_1, \ldots, t_i, t_{i+1}, .., t_{j,}, t_{j+1} .., t_n\right) \tag{4}$$

A Temporal pattern TP is the sequence of the all state transitions over interval t_{ij} where,

$$\{S\} = \left\{\left(s_i, ., s_k, .., s_j | R\right)\right\} \tag{5}$$

and a sub-pattern

$$SP\epsilon\ \{TP\} = (s_i, ., s_m), \}\ \text{over}\ (t_{im})\ \text{and}\ (1 \leq i \leq m \leq\ j \leq n)\ \epsilon\ \mathbb{R} \tag{6}$$

The resulting output are sets of temporal pattern vectors (TPV), which are comprised of temporal patterns, their associated probabilities and the time interval where the patterns where observed, such that,

$$\left[TPV = \{\, TP_i, \ldots, TP_{i+1}, .., TP_j, .., TP_n\}\right]. \tag{7}$$

5.2.5 Generating Frequent Temporal Patterns

Next is to apply frequent pattern mining (FPM) principles to discover the most frequent items in a given set of temporal pattern vectors. Let $TR = \{TR_1, TR_2, \dots, TR_n\}$ be a transaction database where each transaction TR_i in TR is a set of items such that $TR_i = \{a_1, a_2, \dots, a_n\}$. A set $SS \subseteq TR_i$ is an itemset and the size of SS is determined by the number of items it contains. The proportion of the transactions containing SS in TR is the referred to as the support of SS and is calculated by counting the number of times transactions in TR contains SS over all the transactions TR. SS is termed as frequent if it's support is greater or equal to a user defined threshold. Figure 5.4 shows a sample transaction database with 5 transactions $\{TR1, TR2, \dots, TR5\}$, the second column indicates the number of items in each transaction and each item in a transaction is represented as *var1, var2,..., var8.*

Suppose a user defined threshold support is given as 3, then frequent items are presented in third column where $TR1 = \{var1, var2, var3, var6\}$ contains items that appear 3 or more times in the sample transactions and therefore are categorized as frequent.

Figure 5. 4 : Sample Frequent Patterns: Transactions comprised of multiple items frequent or infrequent

Transactions	Items	Frequent Items	InFrequent Items
TR1	var1, var2, var3, var4, var6, var7	var1,var2, var3, var6	var4, var7
TR2	var1, var6, var7	var1, var6	var7
TR3	var2, var5, var3, var1	var2, var5, var3, var1	
TR4	var3, var5, var6, var8	var3, var5, var6	var8
TR5	var3, var2, var5, var4, var6, var 8	var3, var2, var5, var6	var4, var8

Unlike existing FP approaches where the output is only the set of frequent items based on an initial threshold, In this thesis, an additional iteration is introduced which generates a second set of output comprising of all the sets of infrequent patterns and this is facilitated by another call to the mining step with a second threshold. The research premise is that some infrequent patterns could lead to discovery of rare behaviors that could be missed if focus is only on frequently occurring patterns. In Figure 5.4 column 4 shows a sample of infrequent patterns, *TR1* contains *var4, var7* as items which appears with a support less than 3. Note that items which are not frequent are not in the 3rd column, if we do not have such a secondary threshold, these 2 infrequent items would not be discovered.

5.2.6 Generation of Association Rules

After identifying the frequent itemsets in a given transaction set, next is to generate association rules adopting the work in (Han & Kamber, 2001). For example $(J \rightarrow L)$ is a rule *R*, and *J* and *L* are items or itemsets. *J* is the antecedent while L is the consequence. The confidence of rule *R* is equal to the ratio of the union between J and U denoted as $(J \cup L)$ to that of the support of J. This is the conditional probability $P(L|J)$ such that;

$$\text{Confidence } ((J \rightarrow L) = \text{probability } Pr(L|J) = \frac{support\ count\ (J \cup L)}{Support\ Count\ (J)} \quad (8)$$

The support of J with respect to dataset a *D* is defined as the proportion of transactions TR_j in *D* that contains the itemsets in *J*. Another aspect about association rules is to determine the correlation of the antecedent and the consequence in a generated rule. This

correlation is referred to as a lift (lf) calculated using probabilities such that for a rule $(J \rightarrow L)$,

$$lf_{J \rightarrow L} = \left(\frac{P(J \cup L)}{p(J)P(L)}\right) \quad (9)$$

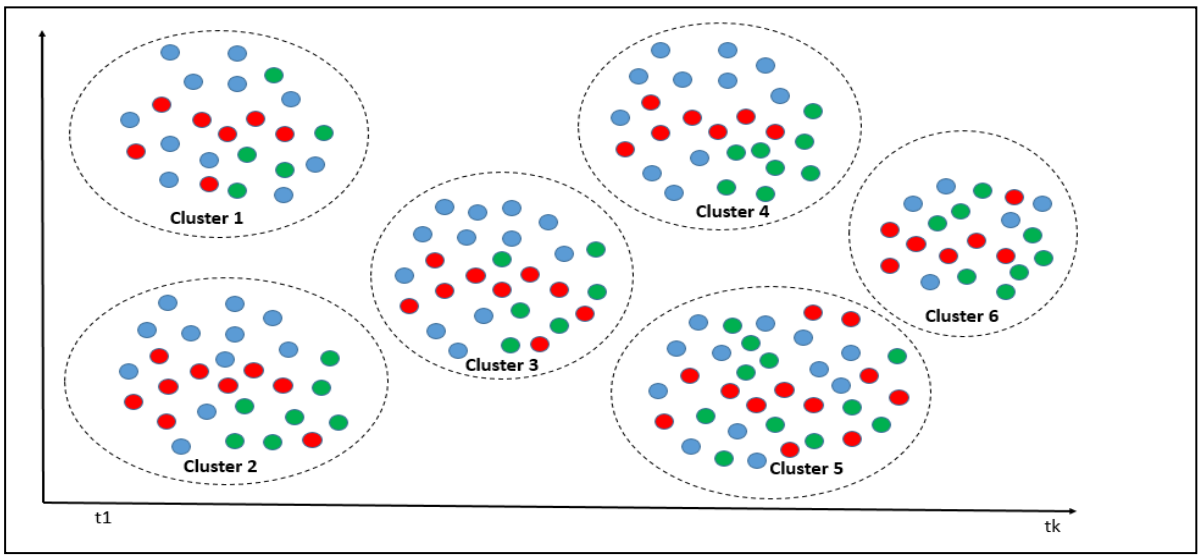
If $lf_{J \rightarrow L} < 1$ then occurrence of *J* is negatively correlated with occurrence of *L*, if $lf_{J \rightarrow L} > 1$ then *J* and *L* are positively correlated and if $lf_{J \rightarrow L} = 1$ indicates *J* and *L* occurs independent of each other (Han & Kamber, 2001).

Handling Multiple Patterns

Even though the process described above can identify both frequent and infrequent patterns, it is still faced with one of the main challenges in frequent pattern mining (FPM) algorithms as noted in the literature review in chapter 3. FPM algorithms generate too many patterns making it difficult to choose which are most important. To tackle this problem, this research utilizes dimension reduction principles by generating a more scaled representation of input data streams. Supplying a scaled data as input to the FPM algorithm allows processing less data streams per iteration when counting itemsets in the data and determining which are frequent. Data driven clustering approach that models data density is adopted in this thesis such that for a given time period, data streams are represented by the derived clusters and their cluster means thereby generating cluster based transactions. This process is similar to locating center points in a sliding widow and then merging those data points in the window which has some similarity to a common centroid. See Figure 5.5 for an illustration of 5 clusters about data captured within some time *t1* to *tk*, each cluster

has 3 different data streams depicted by the colors; blue, green and red. Further details on the clustering approach adopted is detailed in section 5.3.1.1 step1.

Figure 5. 5 : Sample Dataset Classified with 6 Clusters



Clustering aims at partitioning a given datasets D into M clusters, s.t.

$$\left(c_i, c_{i+1}, .., c_j, .., c_m | M\right), (c_i \cap c_{i+1}) = \emptyset, \text{ and } \left(c_i \cup c_{i+1} \cup .., c_j \cup .., c_m\right) = D. \quad (10)$$

For each cluster c_i, the center of a cluster termed as a cluster centroid $(cd_i \in c_i)$ or the cluster mean is generated and

$$(cd_x \cup cd_{x+1} \cup, .., cd_m) < |M| \quad (11)$$

This approach allows elimination of duplicates within a cluster c_i given its content can be represented in the centroid cd_i, s.t. $cd_i < c_i < D$, and efficient identification of both frequent and infrequent patterns. Additionally, representing the many data elements within

a cluster such as the ones depicted in Figure 5.5 with their respective cluster centroids would allow a reduction of the original dataset.

5.2.7 Classification with Frequent Temporal Clusters

Finally the temporal clusters and their associated cluster centroids augmented with the generated association rules forms a classification system $CS = \left(TC_i, .., TC_j, .., TC_n | \, C_n\right)$ such that a new temporal pattern (TP) is categorized as $(TP \, \epsilon \, TC_i)$ if TP falls under the similarity cluster centroid of TC_i.

5.2.8 Summary

The contribution described in section 5.1 provides the techniques used in development of the advancement described in Figure 5.1. In particular, component 1 is accomplished in temporal representation, transactions and relationships detailed in sections 5.2.1, 5.2.2 and 5.2.3 respectively. Component 2 is accomplished in methods described in section 5.2.4 and component 4 is accomplished in methods described in section 5.2.5, 5.2.6 and 5.27. Tasks to facilitate the methodology described in section 5.2 are detailed next in section 5.3.

5.3 The Temporal Pattern Recognition and Mining Algorithm

There are many steps that are needed to facilitate the techniques detailed in section 5.2 especially when applied to multiple data streams simultaneously. In this respect, a step wise approach is adopted in this thesis by development of an algorithm termed as Temporal Pattern Recognition and Mining (TPRMine). TPRMine is the main algorithm and each step

contains its own sub processes and modules. To give a clear picture of the proposed algorithm, this section includes some contents from the literature in order to tie together the fusion of the methods adopted in development of TPRMine algorithm.

Main Algorithm: Temporal Pattern Recognition and Mining (TPRMine)

A time series observation $O = (o_1, o_2, .., o_i, ., o_j, ., o_n)$, where each $o_i \epsilon O$ is a data stream that contains values each captured at some unit time $t \epsilon T$. If t=1, there exists time segments ts_i , $ts_j \in T$ and $ts_i = \{t_1, .., t_k\}$, $ts_{i+1} = \{t_{k+1}, .., t_m\}$ such that o_i over ts_i, ts_j forms sets $ts_i(o_i) = \{o_{i,t_1}, o_{i,t_2}, \dots, o_{i,t_k}\}$ and $ts_{j(o_i)} = \{o_{i,t_{k+1}}, o_{i,t_{K+2}}, \dots, o_{i,t_m}\}$ where $1 < i < j < k < m \leq n$ and k and m are predefined or randomly determined.

SO is a subset of O over some $ts_i \epsilon T$ and forms a set of data streams such that,

$$ts_i(SO) = \left\{\{o_{i,t_1}, o_{i,t_2}, \dots, o_{i,t_k}\}, \{o_{j,t_1}, o_{j,t_2}, \dots, o_{j,t_k}\}, \dots, \{o_{m,t_1}, o_{mt_2}, \dots, o_{m,t_k}\}\right\} \quad (12)$$

Therefore,

Step 1: For a given $SO \subseteq O$ captured at some $ts_j \in T$,

Identify all possible states $(S_1, S_2, \dots, S_m)$ the observations in SO may transition to.

Step 2: Formulate the mathematical distribution of the elements in SO with respect to states in S. At time segment ts_j;

Identify the probability of *SO* being at a particular state S_j in some time $k \in ts_j$

Identify the probability of transitioning from a state $s_i \; to \; s_j$ over *SO*

Identify the likely hood of staying in the same state in the next time period

Quantification of states

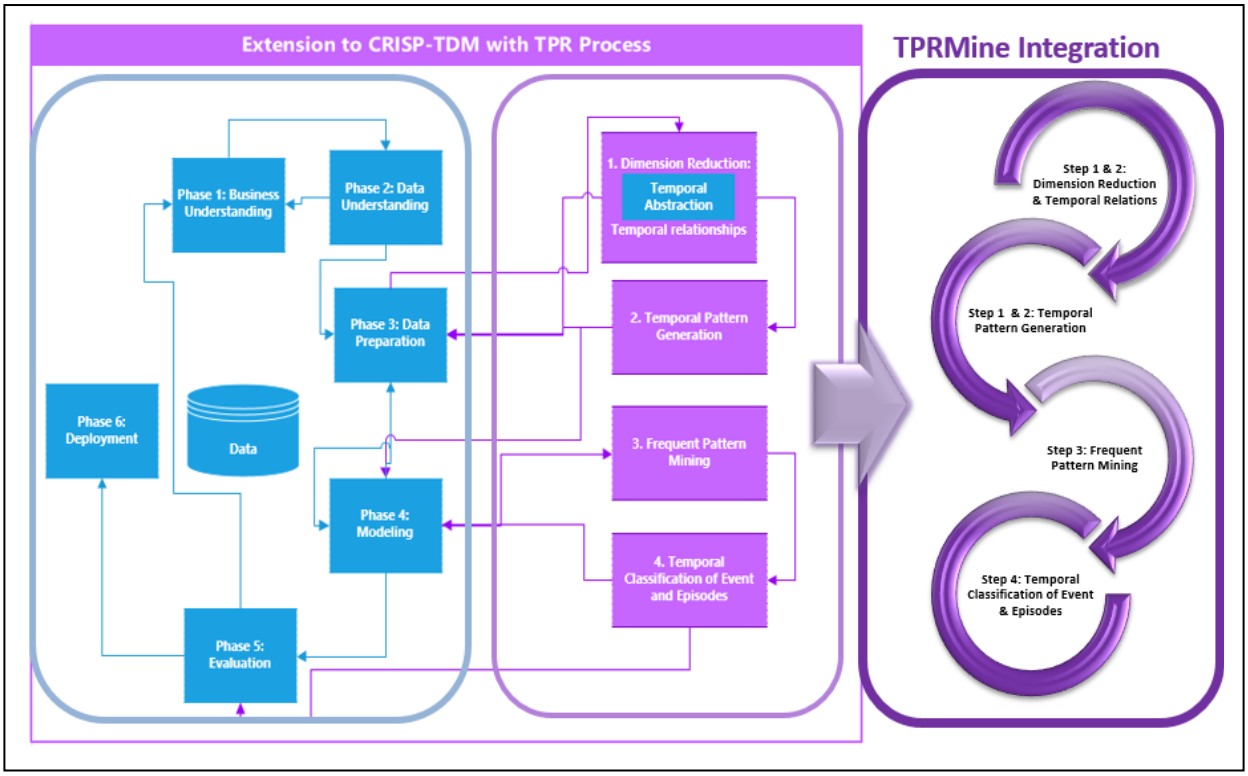
Step 3: Identify the frequent states in time segment ts_j

Step 4: Classify the frequent states based on some domain context associated with the derived states.

Integration of TPRMine Algorithm to Proposed CRISP-TDM Enhancements

Integration of the steps in TPRMine are presented in Figure 5.6 where, functions to support components 1 and 2 are accomplished in steps 1 and 2 while component 3 is accomplished in step 3 and component 4 is accomplished in step 4. The next section provides a walk through of the steps in TPRMine when applied to a sample dataset.

Figure 5. 6: TPRMine Steps integrated within the Proposed CRISP-TDM Extension
Functions in steps 1, 2 support components 1 and 2, components 3, 4 by steps 3 and 4 respectively.



5.3.1 Methods in TPRMine Algorithm

Using a time series sample data collected over some time interval, this section provides a walkthrough on the developed algorithm when applied to a sample data set comprised of four variables similar to the data represented in Figure 5.7(a).

5.3.1.1 Step1: Data Driven Clustering

Suppose a subset of data streams captured from patients in a NICU contains four unique variables such as the sample data in Figure 5.7 (a) and each patient has their own independent data. The data contains random values and a unit of time in seconds when each value was generated. The data is assumed to be random as independent random external events can lead to changes in physiological data for example changes that are resulting from the nervous system response to external stimulus. This thesis assumes that by combining all the patient's data captured at a select time segment, one can derive hidden facts about the data and in so doing, about the patient state. This can be characterized using Gaussian Distributions and modelled with Gaussian mixture models in order to identify grouping identities that are not obvious. As noted in chapter 3, clustering is part of Gaussian mixture models that have played an important role in statistical analysis of data due to their flexibility for modelling a wide variety of random phenomena about data.

This thesis adopts a data driven clustering approach where by formulation of time series data streams into states is accomplished by using the clustering principles in (Fraley & Raftery , 2002). This is a model based clustering technique that allows data driven identification of differing clusters within a specific time period.

This approach highlights the different states data streams can transition to in a given time period, therefore, multiple clusters depict multiple states.

This is a special type of unsupervised learning process that considers data as a mixture of density as follows.

Let $D = (N\ X\ V)$ be a data matrix and each observation $x_i \in D$ is a V dimensional vector of random variables $(x_i, \dots, x_v)$. A model based clustering applied to D assumes that elements $x_i \in D$ forms a finite mixture of K components that corresponds to groups within the D. To this respect, the density of each component is modelled as a multivariate Gaussian distribution as follows;

$$f(x_i|\varphi) = \sum_{=1}^{K} T_k \omega(x_i \mid \mu_k \sum_k) \quad (13)$$

Where T_k is the mixing proportions and $\sum_{k=1}^{K} T$ = 1 and T_i > 0 and $\omega(.)$ is the multivariate Gaussian density comprised of mean vector μ_k and covariance matriz $\sum_k$. $\varphi = (T_1, \dots, T_{k-1}, \mu_1, \dots, \mu_k, \sum_k, \dots, \sum_k)$ is the vector of model parameters.

The component densities represent the group or cluster each observation x_i belongs to and this is represented by a variable Z_{ik}, such the $Z_{ik} = 1$ if x_i belongs to the *k* component otherwise $Z_{ik} = 0$. Fraley et al. (2012) notes that clusters are ellipsoidal centered at the mean vector μ_k with other geometric features such as shape, volume and orientation that is determined by the covariance matrix $\sum_k$. Parameterization of the covariance matrices are obtained using eigen decompositions of the form $\sum_k = \delta_k D_k A_k D_k^T$, where δ_k is a scalar that controls the volume of the ellipsoid, A_k is the diagonal matrix that controls the shape, D_k is

the orthogonal matrix that determines the orientation and D_k^T is the transpose of the orthogonal matrix. Estimating model parameters is completed by utilizing the expectation-maximization algorithm which uses a hierarchical approach to building clusters.

To accomplish similar clustering, this thesis adopts the Mclust package in R that implemented a model based clustering as described in (Fraley, Adrian, Murphy, & Scrucca, 2012). Mclust uses the maximum likelihood to fit all the models with different covariance matrix parametrization, the best model is then selected using the log likelihood and Bayesian Information Criterion (BIC). A higher BIC score indicates strong evidence of the corresponding model resulting in realistic data driven clusters and the integrated complete-data likelihood (ICL) shows the performance of the model in selecting the number of clusters. A demonstration of this process in this thesis is described next.

A sample time series data D with observations captured at some time interval and represented by a N x V matrix, where N = 10,632 and V = 4 attributes {HRVALUE, RRVAVLUE, SPO2VALUE PULSEPVALUE} is provided as input to the Mclust algorithm.

Figure 5.7 shows a pictorial representation of this process. In particular, part (a) in Fig 5.7 shows the content of a subset of the data series while (b) shows the graphical representation of the entire data series D. Results from Mclust algorithm are shown in (c) where data driven 9 clusters are generated and presented in columns labelled [1] to [9]. Entries in each of these columns represents the cluster means (centroids) for each of the attributes within the clusters and the last row represents the volume count in each cluster.

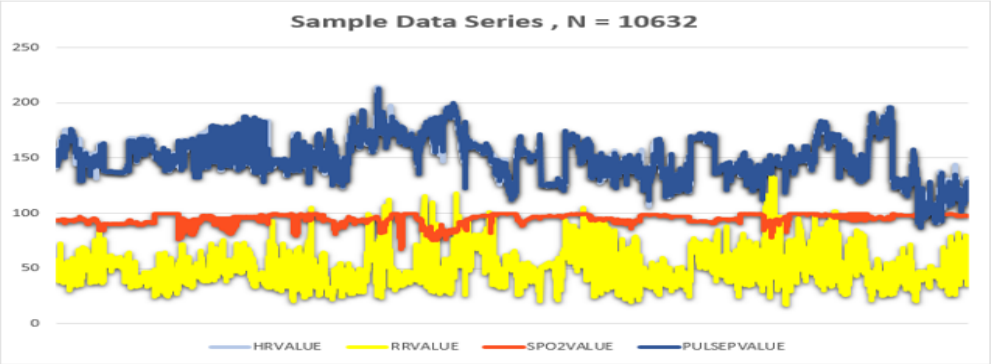
For example, cluster 1 contains {HRVALUE = 146.2943, RRVALUE = 47.63646, SPO2VALUE = 99.6871, PULSEPVALUE = 146.0482} and the cluster contains 2031 data series.

Figure 5. 7 : Clustering Derived from Data Driven Clustering Algorithm

(a) Subset of sample data series

READINGTIME	HRVALUE	RRVALUE	SPO2VALUE	PULSEPVALUE
1537833984	147	49	94	152
1537833985	152	52	94	143
1537833986	152	58	94	143
1537833987	153	57	94	144
1537833988	154	57	94	144
1537833984	155	52	94	144
1537833985	158	50	94	149
1537833986	158	45	94	155
1537833987	158	45	94	155
1537833988	158	50	94	149
1537833984	162	73	93	155
1537833985	163	67	93	157
1537833986	164	67	93	157
1537833987	158	50	94	149
1537833988	166	56	93	157
1537833984	173	42	95	164
1537833985	173	40	94	164
1537833986	174	40	94	164
1537833987	175	41	94	164
1537833988	176	42	94	166

(b) Sample data series captured a some time interval



(C) Mclust generates the best data driven ellipsoidal model of 9 Clusters with varying, volume, shape. The first column specify the attribute of data represented in the other columns.

CLUSTER	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
HRVALUE	146.2943	160.7077	127.2967	157.0434	171.8994	145.1212	168.5639	175.0189	142.2973
RRVALUE	47.63646	60.75539	49.37286	48.65457	60.46248	37.53138	40.98759	53.69699	56.82528
SPO2VALUE	99.6871	94.91669	97.09372	92.71802	97.3497	97.03908	93.53688	84.63732	92.87913
PULSEPVALUE	146.0482	160.5263	126.9923	155.5435	171.4645	143.4871	167.2148	174.5015	141.9628
N	2031	972	1312	751	822	1304	963	919	1558

The derived clusters facilitate characterization of the many states a data series may transition to in a given period of time, such that clusters (CL) form the states (S) and the number of clusters (ck) are the total number of states. For example, using the results in figure

5.7, as $(ck = 9)$ then $(CL = c_1, \ldots, c_9)$ then there are 9 possible states the data series D may transition and therefore, $S = (s_1, \ldots, s_9)$ and $(CL = S)$.

The states generated from clustering are the foundation for formulating the Markov Chains thus enabling the time series data to be modelled as a fine state machine as described in section 5.2.

5.3.1.2 Step 2: Application of Hidden Markov Models

The model based clustering approach adopted in this thesis uses the EM algorithm for estimating probabilities when assigning an object to a cluster (Fraley, Adrian, Murphy, & Scrucca, 2012). The same EM algorithm is applied in HMM (Baum, 1972) for estimating transition probabilities, as such the use of predefined states which are derived from data driven clustering as input for HMM is therefore considered suitable in this thesis.

The clustering process described in Step 1 allows the identification of the states a data series may transition in a given time period, but it does not provide the order or likelihood of that transition. To accomplish this process, this thesis adopts the use of HMM which allow formulation of the transition of a process from one state to the next with weighted measures represented by probabilities.

As described in section 5.2, this thesis formulates the problem space as a HMM process in that, given time series observations $O = \left(o_1, o_2, .., o_i, .., o_j, .., o_n\right)$ and a set of possible states $S = (S_1, S_2, \ldots, S_m)$ one can discover all possible hidden state sequences in O for a given

time interval T. Each o_i represents a data streams recorded at some t_i to t_j and $i < j$ and if *i = 1,* then 0_{i1} is the first set of observation recorded at time t_1

This approach allows multiple observations to be processed concurrently thus generating several sets of output as described next.

Transition probabilities

The key property of markov principles as described in chapter 3 is to be able to describe the transition between states, and the probability of moving from one state, as dependent only on the previous state. When handling patient data as presented in chapter 4, the potential for the trajectory of a patient health status can change very quickly and can be affected by an instant change of events as noted in section 4.2, heart rate is regulated by the body's nervous system or other factors. For example, a patient's heart rate may increase or decrease suddenly after the administration of some medication. Therefore, the effect of such change is not dependent on what may have happened hours before the administration of medication but due to the sudden administration of the medication. Such events can be characterised as states and the change from one state to the next can be seen as following the markov principles whereby the occurrence of an event is dependent only on the previous event. In this thesis, characterization of the likelihood of transitioning from one state to the next is facilitated by modelling patient data using HMM. This process allows quantification of changes in patients physiological data with probabilities.

Providing as input to the HMM algorithm is a data series similar to data depicted in Fig 5.6 (b) and the set of states $S = (s_1, \dots, s_9)$ derived from clustering as described in section

5.3.1.1 Step 1 generates several outputs. One of these outputs is a matrix of probabilities (transition matrix) X such that $X = x_{01}, x_{02}, \dots, x_{n1}, \dots, x_{nn}$ and each $x_{ij} \in X$ represent the probability of transitioning from state s_i to s_j, such that $\sum_{j=1}^{n} x_{ij} = 1, \ \forall i$. The transition matrix in Table 5.1 represents the probabilities of transitioning from each of the derived states S where $S1$ is the state start and $S9$ is the end state. These probabilities are represented with percentages.

A sum of all the probabilities in each of the rows equals to 1 (100%). For example, in Table 5.1, the second row (in blue) shows the probabilities of transitioning from state s2 to 4 other states i.e. s2 to s2 = 92.4%, s2 to s3 = 3.3%, s2 to s6 = 0.4% s2 to s8 = 0.8% and s2 to s9 = 3.2% and a total of the probabilities = 100%. The transition matrix also indicates that there is 0% probability of transitioning directly from state s2 to s1. Each column represents the probability of the process transitioning from a state S_i given the process is already in state S_j. For example, the fourth entry in the first column (s4s1) $= 5.2\%$ indicates the probability of transitioning to a state s1 if the process was already in state s4.

Table 5. 1: A Sample Transition Matrix

There are 9 states that a process may transition to and their associated probabilities. To ease the display, red indicates transition greater that 0 but less than 5% for all state other than the current state. For example, a process can be in stat es6 with a 96.2% probability and can transition directly from s6 to 5 states (s2, s3, s5, s8, s9).

	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	94.8%	0.1%	0.1%	2.6%	1.4%	0.2%	0.3%	0.3%	0.1%
s2	0.0%	92.4%	3.3%	0.0%	0.0%	0.4%	0.0%	0.8%	3.2%
s3	0.1%	2.8%	92.7%	0.3%	0.1%	2.4%	0.0%	0.9%	0.6%
s4	5.2%	0.3%	0.4%	88.1%	0.2%	0.5%	0.1%	0.9%	4.3%
s5	1.8%	0.1%	0.4%	0.2%	93.3%	0.1%	3.2%	0.8%	0.1%
s6	0.0%	0.4%	1.8%	0.5%	0.0%	96.2%	0.0%	0.9%	0.1%
s7	0.2%	0.0%	0.1%	0.2%	3.2%	0.0%	96.3%	0.0%	0.0%
s8	1.1%	2.1%	2.0%	1.4%	0.4%	0.6%	0.0%	91.5%	0.9%
s9	0.2%	3.0%	0.4%	3.8%	0.3%	0.0%	0.0%	1.5%	90.8%

(i) Temporal Transactions

There are two kinds of temporal transactions that can be derived from clustering and HMM models. First, a discussion of the temporal transactions within the HMM paradigm is presented.

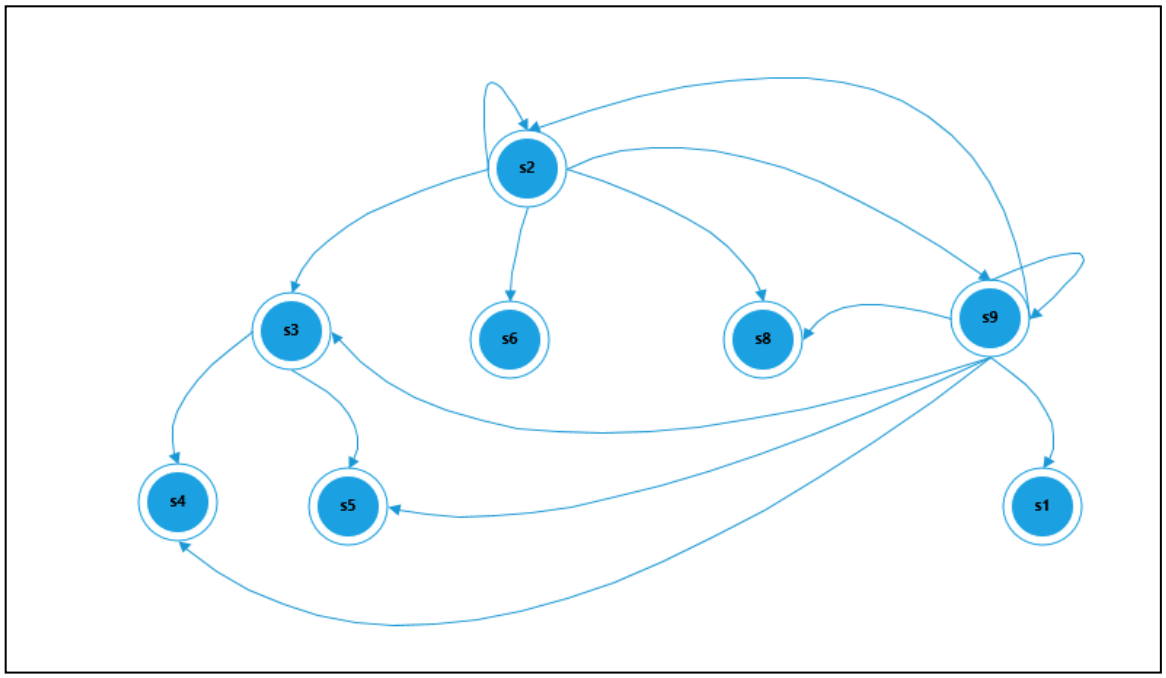
HMM Based Temporal Transactions

As a data stream can transit to multiple states, i.e., $\{(s_o, s_i), (s_0, s_j), \ldots, (s_i, s_j), .. (s_i, s_n), .., (s_n, s_j), \in S$, with hidden probabilities $A = (x_{1i}, \ldots, x_{ij}, \ldots, x_{nn})$.

With the transition probabilities we are able to identify the many state transitions that are possible, this formulates a path like structure quantified by probabilities. For example, if there is a transition from a state S_i to a state S_v and then S_v transitions to state S_j then transition probabilities $x_{iv} > 0$ and $x_{vj} > 0$ indicates that there is a path from state S_i to state S_j passing through state S_v quantified by transition probabilities x_{iv} and x_{vj}. A combination of

such paths forms the temporal transactions. A tree like structure allow the identification of the sequence of states a process can transition to. Figure 5.8 shows a sample of some of the potential states a process can transition to starting from state S2.

Figure 5. 8 : A Sample Temporal Sequence Starting from State s2
This path like structure is derived from some of the contents from the transition matrix in Table 5.1



Temporal Relationships

Further evaluation of temporal transactions allows identification of relationships among different states within a given time interval, these are termed as temporal relationships. In particular, there are two relationships that can be derived in this process as follows.

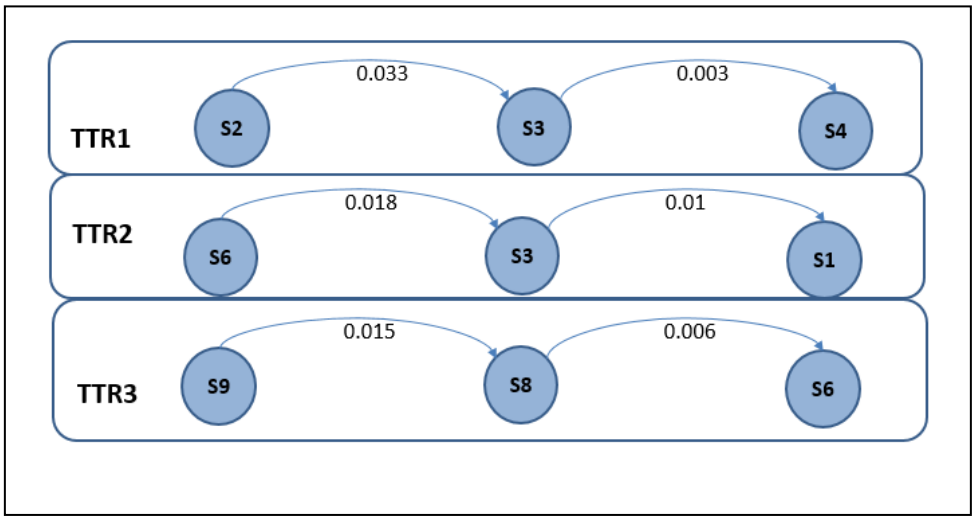
Transitive Temporal Relationship (TTR):

From the transition matrix in Table 5.1 then there are several transition relationships that can be derived which are represented as set TTR = {TTR1,TTR2,…,TTRn} where TTR1 =

$\{s_i, s_l, s_j\}$, if there is a path from s_i to s_j through s_l while a TTR2 = $\{s_j, s_m, s_i\}$ if there is a path from s_j to , $s_i$through state s_m.

Figure 5.9 shows three TTR relationships based on the transition matrix in Table 5.1. TTR1 represents the probabilities of transitioning from state s2 to s3 through s4, TTR2 represents the transition from s6 to s3 through s1 and TTR3 represents the transition from s9 to s6 through s8.

Figure 5. 9 : Transitive Temporal Relationships
TTR1 indicates a relationship between S2 to s4 through S3



Equivalence Temporal Relationship (ETR)

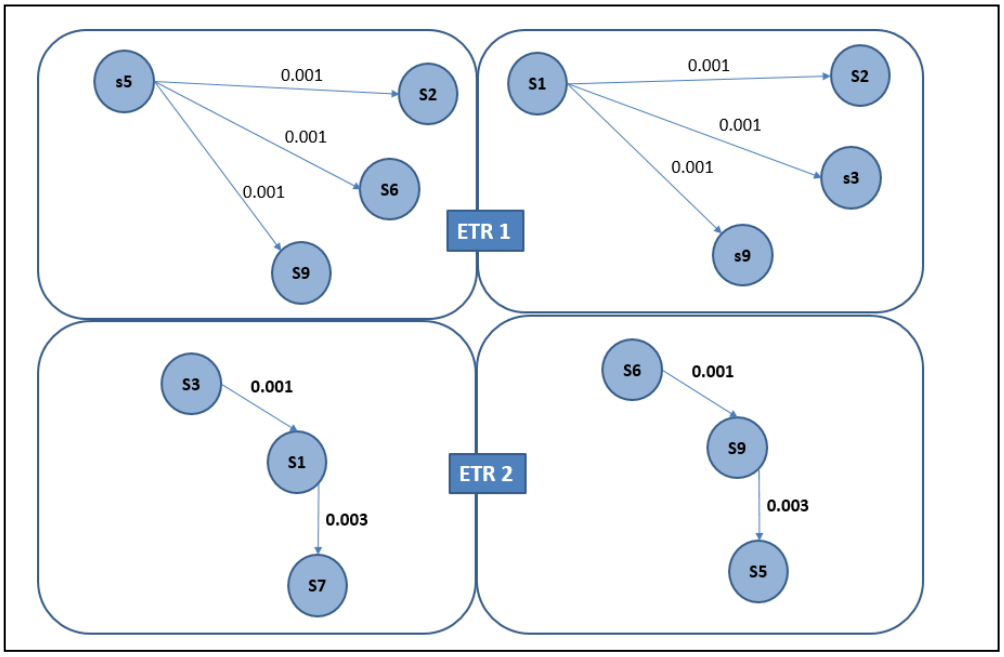
Equivalence relationship is the second temporal relationships identified where if the probability of transitioning from a state s_i to s_j is equal to the probability of transitioning from s_j to s_k, then there exists an equivalence relationship { s_i, s_j } and {s_j, s_k}. A transitive relation is also termed as an equivalence relationship if there are states s_x, s_y, s_z and s_m, s_n,

s_p and the probability from s_x to s_z through s_y is equal to the probability of transitioning from s_m to s_p through s_n.

Figure 5.10 shows 2 equivalence relationships ETR1 and ETR2 whereby in ETR1 the probability of transitioning from state *s8* to *s6* other states is the same as the probability of transitioning from a state *s4* to state *s6* or *s8*.

Figure 5. 10 : Equivalence Transition Relations

ETR1 indicates the equivalence probabilities of transitioning from state S5 to 3 other states and from s1 to 2 states while ETR2 indicates an equivalence of transitioning from s3 to s7 through s1 and from s6 to s5 through s9.



(ii) Clustering Based Temporal Transactions

A second type of temporal transaction is formulated using the data driven clustering adopted in this thesis. This process allows grouping data streams into unique clusters which are then identified as states a process can transition to. Each cluster has a centroid value (cluster mean) that can be used to represent the elements within the cluster. With multiple

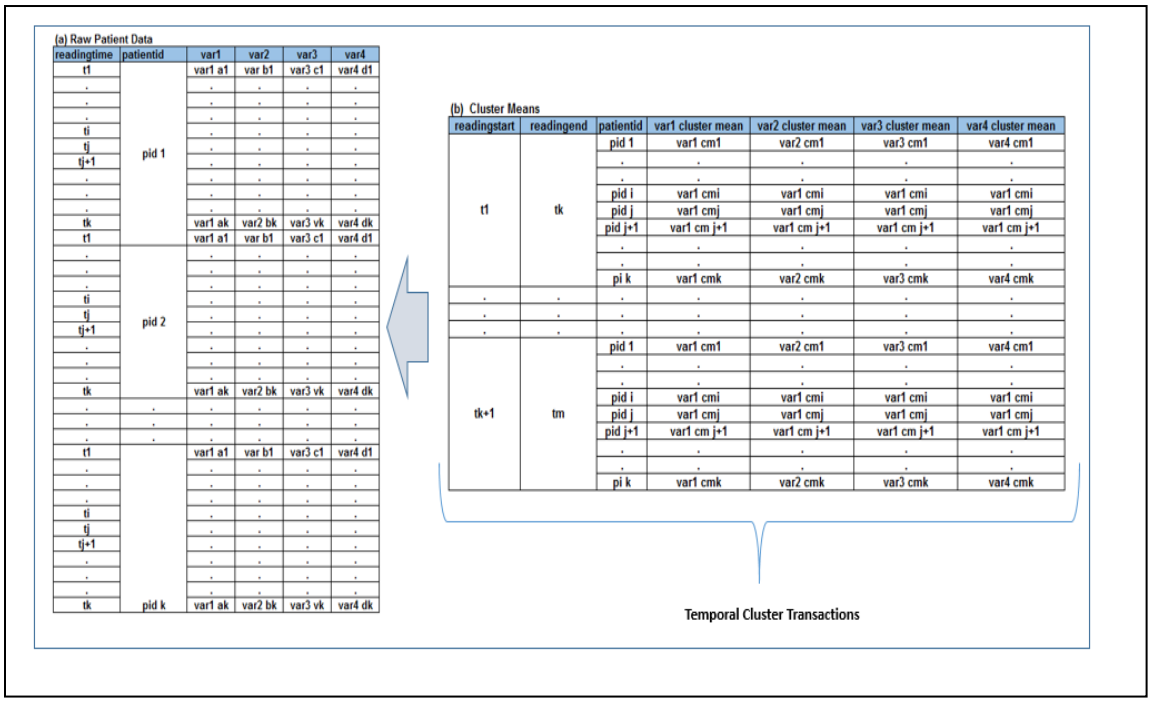
data streams captured from different sources i.e. different patients, then we can represent the underlying data streams with their respective cluster centroids. In a specific time interval a patient can have multiple data streams, these data streams if represented with their underlying cluster centroids resulting to temporal transactions.

Given a set of data streams as shown in Figure 5.7(a), replace each of the records with their associated cluster means such, the resulting data set is the temporal transactions per second. A combination of all the transactions at some time interval is referred to as temporal cluster transactions (TCT).

For example, Figure 5.11 (a) shows a sample of raw data with 4 variables (*var1, ver2, var3, var4*) each containing data for a unique patient identified with patientID with values { pid 1, …, pid k}. This data is fed into the clustering algorithm while Figure 5.11 (b) shows the cluster means generated from the derived clusters. This high level representation of large volumes of data points is a dimension reduction strategy that make it feasible for further data processing as detailed in the next section.

Figure 5. 11 : Temporal Cluster Transactions

(a) Shows a sample data (4 variables) for multiple patients pid 1 to pid k. Each patient has a reading for each of the four variables at each second. Clustering generates a cluster mean for each variable, therefore (b) shows the potential cluster means that are used to represent the original patient raw data. For a specific time interval ti to tk, then the resulting datasets forms temporal cluster transactions.



Quantification of Temporal Relationships

It is possible to quantify the temporal relationships identified. One approach can be the use of some predefined risk or no risk, and with this risk, there can be a set of characteristics introduced about the data based on the transition probabilities and the scores associated with a specific state. For example, if state *a* has higher risk than a state *b*, then the transition from state *a* to *c* and state *c* to state *b* is a more risky path than the latter. Such quantification highlights more details about relationships in data streams that can be associated to some patient risk if processing healthcare data domain or a malicious cyber attack when processing cybersecurity domains.

5.3.1.3 Step 3: Identification of Frequent Temporal Patterns

Temporal Patterns: There are two types of temporal patterns that result from a combination of all the generated temporal transactions. Probabilistic temporal patterns derived from the temporal transaction TTR and ETR and cluster temporal patterns based on the temporal cluster transactions such that a Temporal Patterns TP = {{TTR1, TTR2,…,TTRn} U {ETR1, ETR2,…, ETRn} U { TCT1, TCT2,…,TCTn}}.

After deriving temporal patterns in a given time period, next is to identify the frequent patterns from the generated temporal patterns. To facilitate this process the Apriori algorithm as described in (Agrawal & Srikant, 1995) is adopted. In the Apriori algorithm input datasets are formulated as transactions as described in section 5.2.5 and the algorithm counts the number or times items in each transactions are found in the dataset. The temporal cluster transactions are formulated as itemsets for feeding as input to the Apriori algorithm. Association rules are then applied to generate the frequent cluster means and associated states in a given time period. These forms the frequent temporal patterns.

A further processing of the temporal cluster transactions in Figure 5.11b derives the data in Table 5.2 where each record in a specific time interval is termed as a transaction set and data values in each transactions are termed as itemsets. Table 5.2 shows a sample transaction sets 1 to n and each transaction can have multiple itemsets. Applying association rule algorithm generates frequent and infrequent itemsets.

Table 5. 2 : Temporal Transactions

Trans1 is the first record set in a given time period, starting from 1st minute to the kth minute, trans k contains the last data captured in the kth minute.

transactions	itemsets
trans1	item1, item4item3,item1,item6
.	.
.	.
trans i	item1, item2,item3,item1,item3,item9
trans j	item1, item9,item7,item2,item3, item6
trans j+1	item1, item6,item7,item2,item6, item2
.	.
.	.
.	.
.	.
.	.
.	.
tranks k	item1, item2,item3,item1,item3

In handling multiple data streams, even in a short time period, a massive data set is processed and feeding raw data similar to the sample dataset in Figure 5.7 to an Apriori like algorithm would generate hundreds of rules making it very hard to discern which are important. To address this problem in this thesis is accomplished by use of temporal cluster transactions as described in section 5.3.1.2 where a higher level presentation of the raw data is completed.

In particular, clustering allows effective scaling of the amount of data streams fed to the mining algorithm as a set of input data streams is represented by their cluster centroid. This process is accomplished as follows.

Frequent Temporal Patterns

The output of the mining algorithm is a set of frequent temporal patterns. In particular given a streaming data Transactions $T = \{T_1, T_2, \dots, T_n\}$ where each $T_j \epsilon T$, for all $j = \{1, \dots, n\}$ consists of set of items $T_j = \{x_1, x_2, \dots, x_k\}$ and each x_i is recorded at some time t_i. A temporal data window TW_{ij} contains all transactions $\{T_i, T_{i+1}, \dots, T_j\}$ i <j.

An itemset $P \in T_i$ is defined by the number of items x_i it contains. Mining frequent patterns in data streams finds all the frequent pattern P contained in TW_{ij} resulting to association rules. For example, if there is a rule X → Y generated from mining TW_{ij}, where X contains items (x_1 = 90 $\wedge$ x_2 = 90 $\wedge$ x_3= 90 $\wedge$ x_4 =120) and *Y* contains {x_6=3}, then support of the rule X →Y is the number of times transactions includes *X* and *Y* appears in TW_{ij}.

Using the data from Figure 5.7, one of the rules generated is of form (HRvalue= {104,139} ^ RRValue = {40.7, 54.3} ^ spo2 = {87, 95.3} ^ Pulsepleth = {104, 139} → cluster = {6,9}). Other measures that can be derived from R includes the confidence and lift as described later in section 6.1.6 and more results are shown in section 7.

5.3.1.4 Step 4: Classification with Temporal Association Rules:

A set of rules $R = \{R_1, R_2, \dots, R_m\}$ and 1 < m generated from a temporal window TW_{ij} can be grouped based on some domain context $DC = \{DC_1, DC_2, \dots, DC_j\}$, $DC_i \neq DC_j$, for all $(i, j) \epsilon \mathbb{R}$. A ruleset *RS* $\subseteq R_i \in$ R is associated with some $DC_i \in DC$ if there are some identified context in DC_i that contains the rules in R_i. A combination of $RS_i, RS_j \in$ R where $RS_i \neq RS_j$ forms a class $CL = \{RS_1, RS_2, \dots, RS_k\}$. A new rule R_x is categorized as belonging to class CL based on some similarities with patterns in CL. A fusion of all classes *CL* forms $TCS = \{CL_1, CL_2, \dots, CL_3\}$ the overall classification system.

For example, given a data domain such as clinical care where high level abstraction of some patient physiological data streams is provided i.e. heart rate (80, 100, 180) is categorized as abnormal low, normal, abnormal high, scoring scheme can then be created which can score patterns that are normal or abnormal in the associated patient.

In this approach a classification system can then be built which contains all the frequent patterns based on some threshold on containing normal/ abnormal values. This process changes data to information that could then be utilized by healthcare providers when making decisions about care of patients.

Demonstration of utilizing TPRMine algorithm in healthcare domain is discussed in chapter6.

5.4 Conclusions

This chapter has presented an innovative method to detect and represent relationships that may exist in temporal abstraction and temporal patterns derived from time oriented data. The development of this method involves extensions to the following frameworks, CRISP-TDM described in (Catley, 2009) and $STDM_0^n$ detailed in (McGregor, 2011). New components have been added to the existing frameworks in order to support the knowledge discovery process leading to research contributions to computer science and health informatics. The contributions in computer science involves advancement of knowledge representation and discovery of patterns using a phased fusion of machine learning principles in pattern recognition, dimension reduction and data mining integrated with frequent pattern mining and classification. This approach is facilitated by introduction of the TPR Process and an associated TPRMine algorithm where time series data is modelled as a finite state machine and the goal is to understand the behavior of transitions the process may transition to at any given moment in time. Demonstration of application of the developed method in contribution to health informatics by extending the $STDM_0^n$ in (McGregor, 2011) is detailed next in chapter 6.

Chapter 6

6 Application to Health Informatics and Medicine

6.1 Application to Health Informatics

This chapter presents details on contributions to health Informatics completed in this thesis. This contribution involves extending the patented multi-dimensional temporal data mining framework presented by (McGregor, 2011) and integrating these enhancements to the Artemis Platform (McGregor et al, 2013) for discovery of temporal relationships in abstractions and generating temporal patterns from time oriented physiological data.

As noted in chapter 2, most of the systems developed that have attempted knowledge discovery in clinical care have been able to generate temporal abstractions but have not provided techniques to detect and represent relationships that may exist among temporal abstractions and temporal patterns. In addition, existing systems are not able to maintain the temporal nature of developed patterns and could result in loss of important information of time and context on events in real-time data streams.

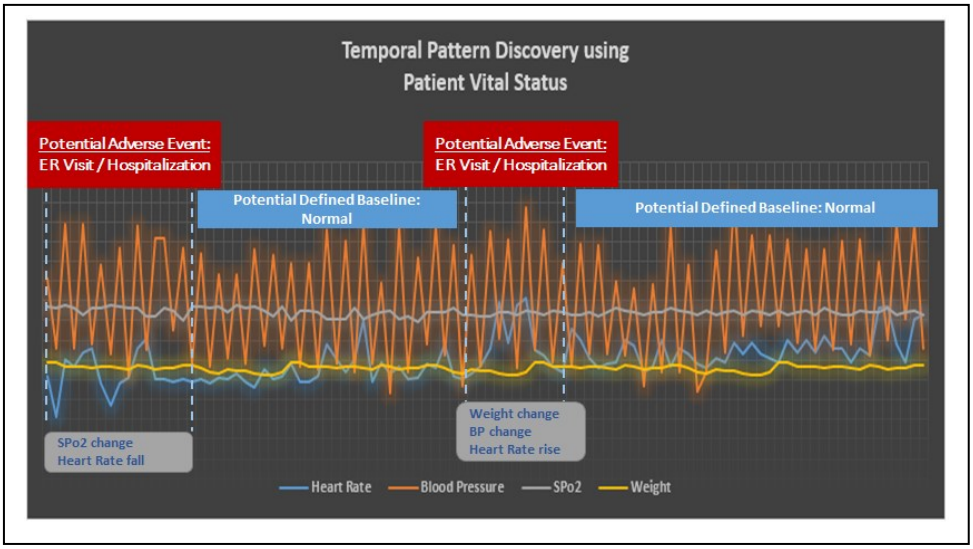
To address this problem, this thesis proposes an innovative method utilizing the temporal pattern recognition process developed in the computer science contribution of this thesis and presented in Chapter 5 for application to physiological data streams. The goal is to demonstrate the potential for knowledge discovery through detection of temporal relationships and the subsequent temporal patterns among physiological data streams at any

given period of time. Discovery of such patterns can aid in clinical decision support on appropriate care for patients.

To facilitate this process, this research provides a framework for real-time knowledge discovery using physiological data streams and classification of temporal events and episodes in a clinical context by (a) extending the multi-dimensional temporal data mining ($STDM_0^n$) framework presented by (McGregor, 2011) and (b) instantiating these enhancements within Artemis Platform (McGregor et al, 2013).

Preliminary research described in (Inibhunu & McGregor, 2017) demonstrated the potential to quantify medical events using physiological data collected from a remote patient monitoring service. As presented in Figure 6.1 potential adverse events are shown to occur in a period of time when; (a) there is a change in SpO_2 and heart rate falls or (b) there is weight change, blood pressure change and rise in heart rate. The ability to capture when such physiological changes happen, how long this takes place and the effect one has over another is key to understanding temporal patterns in data. Detection of such patterns using small datasets might be feasible but this is a computationally expensive process when faced with multiple physiological data flowing from monitors.

Figure 6. 1: Potential Adverse Event Detection with Temporal Patterns



6.1.1 Enhancements to $STDM_0^n$

This thesis proposes to extend the ($STDM_0^n$) framework by (McGregor, 2011). These enhancements are characterized in purple in Figure 6.2. In particular, this research proposes extensions to; data preparation and modelling phases, rules ontology and data management layers. This is facilitated using the 4 components proposed in the computer science contribution discussed in chapter 5.

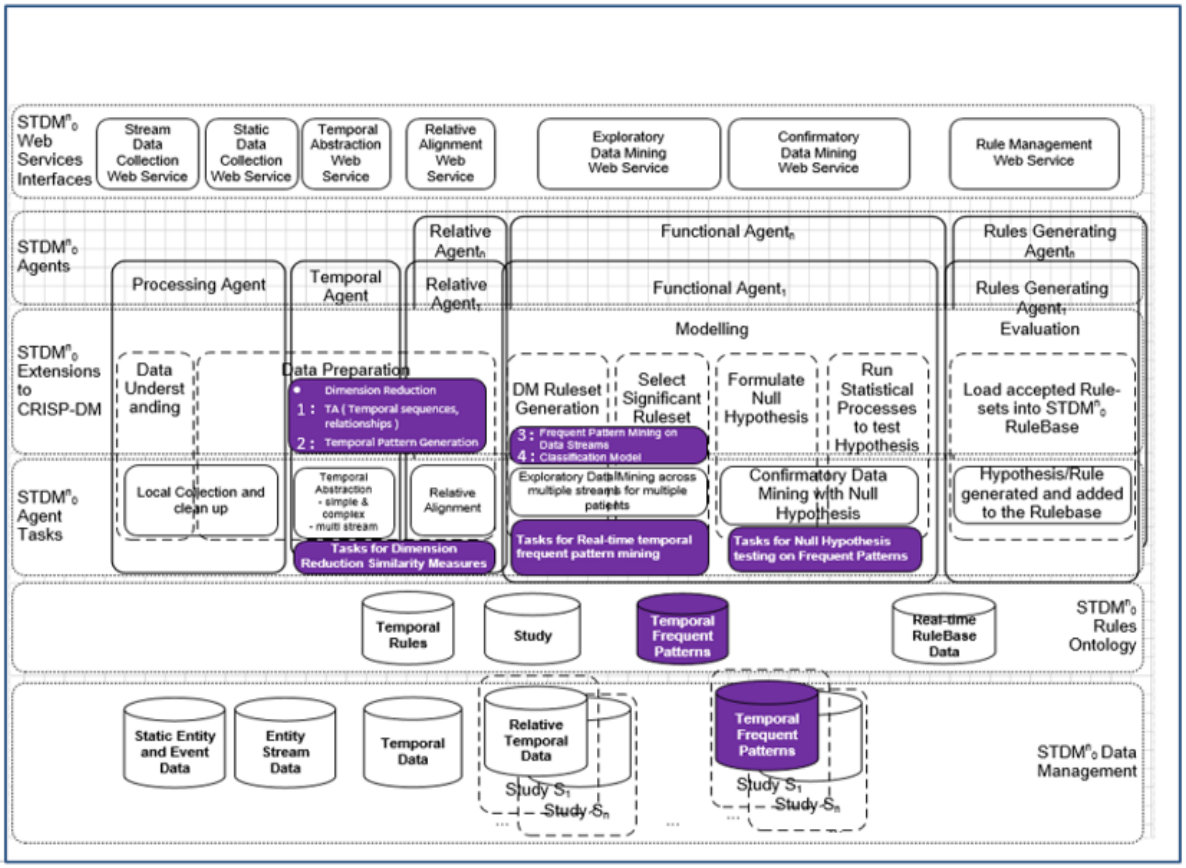
With respect to data preparation two components labelled 1 and 2 in Figure 6.2 are added to support dimension reduction, generation of temporal abstractions using temporal sequences, understanding relationships among sequences and generation of temporal patterns. To support these components, an extra task for dimension reduction and similarity measures is added in $STDM_0^n$ Agent Tasks. This also involves adding more functions to the temporal and relative agents which will need to perform the tasks in the added component.

With respect to modelling, two components labeled 3 and 4 in Figure 6.2 for frequent pattern mining and building a classification model are added and an enhancement to include more tasks in the $STDM_0^n$ Agent has been added. This allows the functional agent to facilitate real-time temporal frequent pattern mining including the ability to support null hypothesis testing on frequent patterns.

Regarding the $STDM_0^n$ Rules Ontology, the discovered temporal frequent patterns forms an extra set of rules that are dynamically updated as new temporal patterns are generated.

The data management layer is also enhanced to include an extra set of data comprised of unique temporal frequent patterns and their quantification metrics, this would be vital for classification of study cohorts based similarity measures. To facilitate this process, this thesis incorporates application of similar principles attempted in temporal databases for holding and updating frequent temporal patterns by incorporating their associated timelines (Combi et al, 2010).

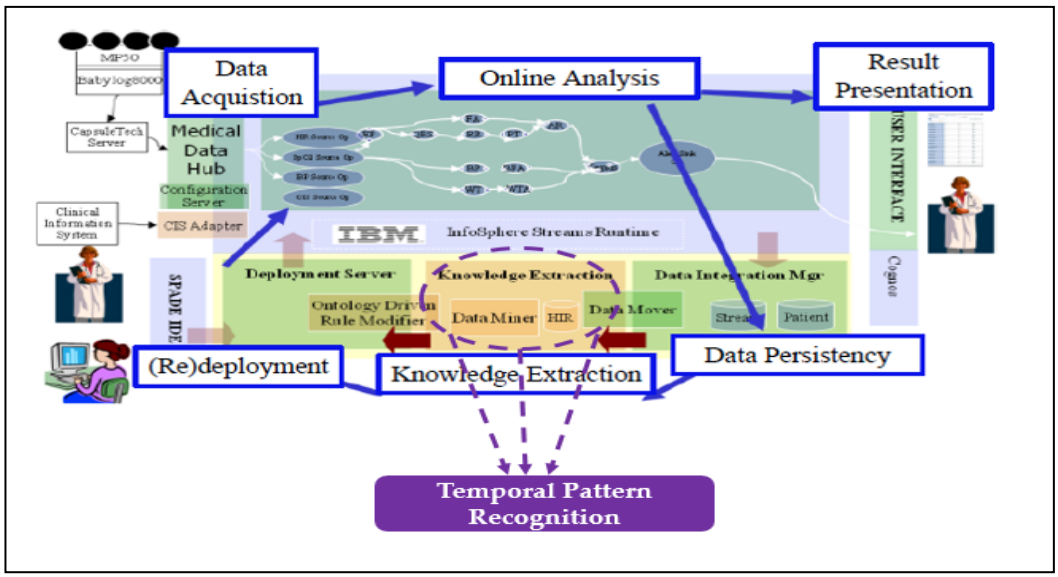
Figure 6. 2 : Proposed Advancement to the STDM Framework (McGregor, 2011)



6.1.2. Enhancement to Artemis Framework

In continuation to creating a robust decision support system for clinical application, this thesis enhances the knowledge extraction component in Artemis (McGregor, 2009). The goal is to create frequent pattern mining algorithms that generate and quantify temporal patterns to support development of a classification system with which to classify normal and abnormal patterns from time series data as shown in Figure 6.3 and depicted in purple. As a result the existing instances of Artemis at collaborating NICUs is utilized. This is facilitated by utilizing the data preparation and modelling components proposed in the enhancement of the $STDM_0^n$ framework as mentioned in section 6.1.1.

Figure 6. 3 : Proposed Enhancement to the Artemis Framework



6.2 TPR Process Applied to Physiological Data Streams

Application of the proposed enhancement to ($STDM_0^n$) and the Artemis framework in this research utilizes physiological data streams captured through the Artemis Big Data analytics platform (Inibhunu, et al., 2019). This process is facilitated under ethics approved in (HiREB 3859-D) and Ontario Tech (REB #14136) in addition to Late Neonatal Sepsis Study under (HiREB 4833-C) and Ontario Tech (#15536). Every second, a single patient at an NICU at McMaster Children's has physiological data generated from multiple sensors at the bedside and displayed in Phillips Intellivue monitors. The same data is simultaneously captured in Artemis cloud using the real time data streaming modules described in (Inibhunu, et al., 2019). This data is comprised of multiple physiological features that includes: heart rate (HR) which is a measure of the speed of number of contractions of a heart per minute (HR), respiratory rate (RR) which measures the phases of inspiration and expiration, blood oxygen

saturation (SpO_2) which is a percentage of hemoglobin molecules in the arterial blood that are saturated with oxygen, pulse rate from Plethysmogram (PulsePleth) which is the distance between the start and end of a Photoplethysmography signal that measures the blood flow or blood volume in the body, multiple blood pressure measurements, electrocardiogram (ECG) waves among other features.

This thesis utilises the work already established in (McGregor et al, 2010 - 2013; Inibhunu, et al., 2019) and seeks to build on the knowledge within Artemis and in particular, utilizes 4 physiological data streams, HR, RR, SpO_2, and PulsePleth to contribute to the health informatics domain. The variable names that correspond to each of these 4 data streams are as follows (HRVALUE = HR, RRVALUE = RR, SPO2VALUE = SpO_2, PULSEPVALUE= PulsePleth).

Each of these 4 data streams have a tuple generated approximately every second (1024 ms) and therefore these are termed as high frequency time series data streams. Values in each time series data stream are independent of each other. For example if a data stream contains HR, then a HR value is not dependent on another HR value in the time series given that HR is regulated by the response of the nervous system as noted in section 4.2. The same case applies to the other 3 data steams i.e. RR, SPO2 and PulsePleth.

Within 1 hr, approx. 3518 independent tuples are generated within the HR, RR, SpO_2 and PulsePleth streams resulting in a total of 14,400 records for a patient. As there are 51 NICU beds in McMaster, if all these beds are occupied by neonatal patients, then there are have a total of 792,000 records captured per hour for HR, RR, SpO_2 and PulsePleth. As such, research remains open for development of methods that can detect changes and relationships in

patterns that may be exhibited in patients data. Detection of hidden relationships or patterns in patients may lead to discovery of unknown diseases or condition pathophysiology.

To facilitate this process, the proposed TPRMine algorithm described in section 5.3 is applied to physiological data as it allows a scaled approach to processing high frequency time series data streams with temporal data windows. In particular, using patient physiological data streams which are generated approximately every second, processing of temporal windows over some length of time i.e. hours can derive hierarchies of information which would be instrumental to building a knowledge base about the underlying patients.

To enable the creation of temporal data windows, a fixed window approach is demonstrated that adopts similar principles in mining data streams as (Lee, Jin, & Agrawal, 2014) for efficient processing of the high frequency data streams that continue to be generated at the bedside and captured within Artemis.

This thesis argues that TPRMine algorithm can be applied to any given subset of temporal data windows. However, there can be consequences on the time complexity the entire process takes to complete depending on the size of the temporal window. As noted in the evaluation detailed in chapter 7 (figure 7.3), providing as input different temporal windows of data to the TPRMine can result into exponential execution times as data size increases. To this respect, this thesis uses fixed temporal windows comprised of small data size that still enable a two order of magnitude reduction in the size of the data. Such an approach overcomes exponential computation times in processing large data volumes as shown in chapter 7(figure 7.3) at the same time deriving meaningful information.

This is a form of down sampling that reduces the order of magnitude of the data stream being analysed by a factor of 2.

Utilizing the TPRMine algorithm is accomplished in a step wise manner as described next and the pictorial process flow is shown in Figure 6.17.

6.2.1 Step 1: Formulation of Temporal Data Windows

As defined in section 5.2, TPRMine process is applied to a subset of data such that,

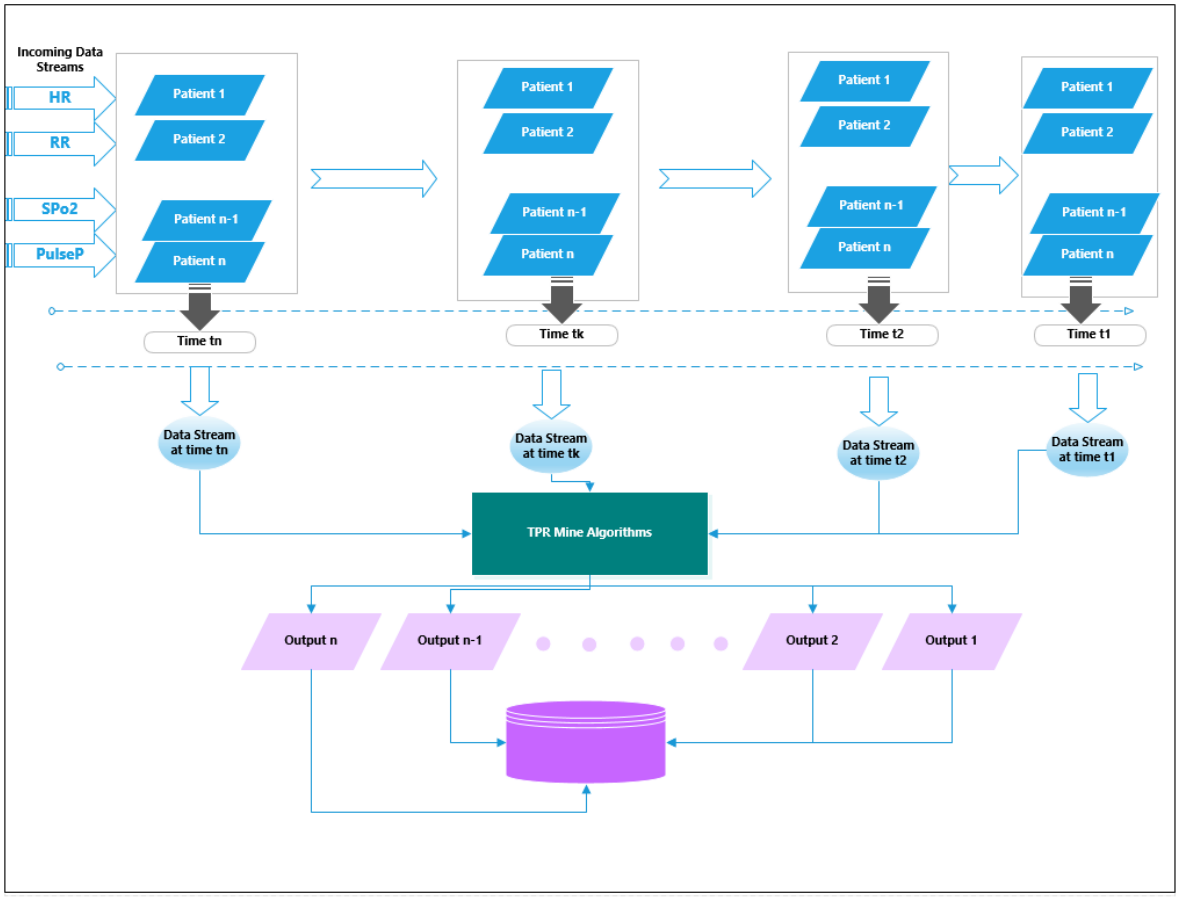
$$ts_i(SO) = \left\{\left\{o_{i,t_1}, o_{i,t_2}, \dots, o_{i,t_k}\right\}, \left\{o_{j,t_1}, o_{j,t_2}, \dots, o_{j,t_k}\right\}, \dots, \left\{o_{m,t_1}, o_{mt_2}, \dots, o_{m,t_k}\right\}\right\} \quad (14)$$

captured in a select time segment $ts_i = \{t_1, .., t_k\}$. The value of k can be dynamically determined or based on some fixed number. When utilizing patients physiological data streams that are captured every second, the value of k can be a fixed window. However, as noted in section 6.2 and demonstrated in the evaluation section in chapter 7 (figure 7.3), the value of k can have a significant impact on both the data size and the time complexity needed for execution of TPRMine components. To this respect, within the health informatics application of TPRMine, this thesis refers $ts_i(SO)$ as a temporal data window (TW_i) captured in a fixed time segment such that $k = 5$ and $ts_i = \{t_1, .., t_5\}$.

As such, elements in each *TW* are multiple physiological data streams captured within a 5 minute time interval. In this thesis, formulation of each *TW* is comprised of 4 data streams HR, RR, SpO_2, PulsePleth per patient and the respective times when the physiological streams were captured within the interval. The resulting temporal data windows are fed as input to the TPRMine Algorithm.

Figure 6.4 shows a sample representation of temporal data windows where data stream at time t_1 is processed first, then data stream at time t_2 all the way to time t_n.

Figure 6. 4 : Temporal Data Windows
Data Streams comprised of 5 minutes Intervals of Data
Time t1 is the first set of incoming data streams processed, followed by time t2 till time tn.



Advantages of Windowing in Data Processing

This thesis has adopted 5 minutes data windows for processing with a premise that, each 5 minute data window is a unique problem space and applying TPRMine to such a problem would generate unique characteristics about the associated temporal data window.

The evaluation described in chapter 7 demonstrates how effective this approach is with respect to amount of time it takes for TPRMine algorithm to process large vs small temporal data windows.

The use of 5 minutes interval is a unique contribution in this research as it allows; (a) minimal data processing, (b) unique identification of patterns in shorter time intervals instead of waiting for an hour to have data processed, (c) The change in cluster transitions among patients highlight any uncertainty in a specific patient and (d) Clusters are high level representation of low level data making it easier to interpret patient state transitions in a given period of time.

After formulation of the temporal widow, next is discovery of the states a patient data may transition to.

6.2.2 Step 2: Identification of State Transitions

To be able to discover the different states the underlying data streams may transition to within a temporal window where the data is captured, data driven clustering described in section 5.3 is adopted. With 4 physiological streams, clustering allows grouping multiple data streams into clusters. As the data is time based, it's clear that a patient cannot be at two different clusters in a unit time within the 5 minutes. However, a patient can transition from one state to another over the course of the time.

The use of clusters also makes it's easier to abstract the underlying data streams to a higher level representation. This is because for each cluster c_i generated, there is a cluster

centroid c_{id}, that can represent all the physiological data streams within the cluster c_i. This centroid is represented by the cluster means.

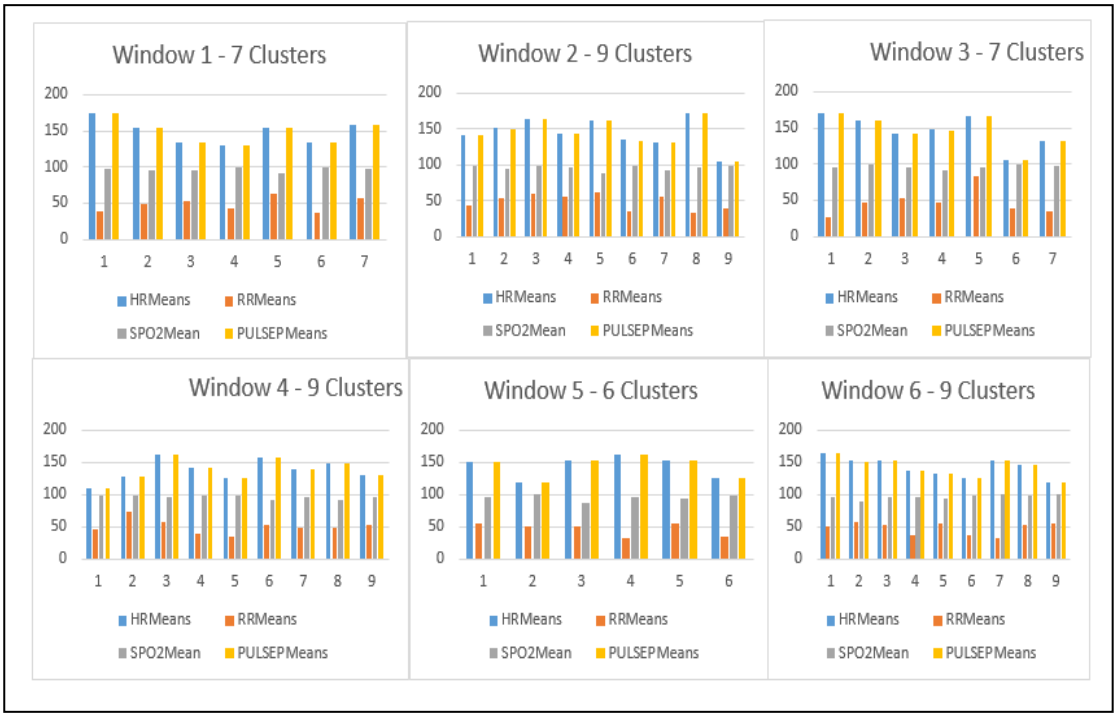
With 4 data streams in a cluster, there are 4 cluster centroids each representing a unique data stream, and the overall cluster is the state the 4 data streams are in a particular time. The data driven clustering approach adopted is a non supervised method that allows the data to dictate the discovery of clusters within a specific time interval as described in section 5.3.1.

Module for Clustering:

1. *For each of the set of physiological data streams in a temporal window*
2. *Generate Data Driven Clusters*
3. *Identify the cluster centroids*

After running clustering on 6 temporal windows using 4 physiological features on patients in NICU, Figure 6.5 shows the different clusters generated for each window.

Figure 6. 5 : Data Driven Clusters Generated from 6 Widows of Time
Y-axis shows the cluster centroids (means) while x-axis represent the cluster



6.2.3 Step 3: Transition Probabilities

The clusters derived in step 2 (Figure 6.5) forms the input for the Hidden Markov Model state. In particular, given a patient P with the following physiological data streams (HR, RR, SpO_2, PulsePleth), what is the probability of transitioning from one state to the next?

Applying the TPRMine Hidden Markov module it's possible to then determine the likelihood of the data streams HR, RR, SpO_2 and PulseP transitioning from one state to the next. This is facilitated by assigning the identified clusters as the potential states to the HMM module. The result is a matrix set of temporal probabilities similar to the results in Table 6.1. Data streams containing (HR, RR, SpO_2 and PulseP) captured at 3 different time windows (periods) generates three matrices which represents the probability of moving from state to state in

each time window. TPRMine has been applied to 11 hours of patient data, a sample of 15 minutes (3 time windows) of data is presented to show there are 9 derived states i.e. s1, s2 … s9 at each of the time windows. This process follows the following procedure.

Module for Modelling HMM

1. *For each 5 minutes Interval*
2. *Get Input data and the potential states*
3. *Prepare data for modeling*
4. *Apply HMM*
5. *Identify initial probability*
6. *Generate Transition probabilities*
7. *Predict the next probable state*

After processing data in this module, each 5 minute interval processed has a one step transition matrix generated that forms the temporal transaction in the given time widow, a sample is shown in Table 6.1.

Table 6. 1 : Transition Matrices on 3 Consecutive Periods

Each Period represents a time window i.e., Period 1 represents time window 1

Period 1: Transition Matrix, Initial State = 6

	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	93.2%	0.0%	0.2%	0.7%	0.5%	0.0%	1.9%	3.4%	0.1%
s2	0.5%	93.4%	0.0%	0.0%	2.8%	3.0%	0.0%	0.3%	0.0%
s3	0.0%	0.1%	94.6%	0.2%	0.1%	0.5%	1.1%	2.4%	0.9%
s4	1.3%	0.1%	0.2%	91.9%	3.2%	0.3%	3.1%	0.0%	0.0%
s5	0.3%	1.2%	0.3%	3.8%	92.1%	1.5%	0.2%	0.3%	0.4%
s6	0.1%	1.6%	0.9%	0.4%	1.6%	94.1%	0.3%	0.2%	0.9%
s7	3.0%	0.0%	1.1%	2.4%	0.3%	0.2%	92.3%	0.4%	0.2%
s8	4.1%	0.1%	2.1%	0.1%	0.3%	0.0%	0.4%	93.0%	0.0%
s9	0.1%	0.3%	2.3%	0.0%	0.6%	1.0%	0.9%	0.3%	94.5%

Period 2: Transition Matrix, Initial State = 8

	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	91.1%	0.1%	3.2%	4.4%	0.5%	0.6%	0.1%	0.0%	0.0%
s2	0.3%	92.6%	2.9%	0.2%	0.2%	0.1%	3.2%	0.0%	0.4%
s3	2.3%	3.6%	90.9%	1.3%	0.1%	0.8%	0.7%	0.2%	0.2%
s4	2.9%	0.4%	2.2%	89.9%	0.1%	0.1%	0.2%	0.6%	3.6%
s5	0.1%	0.3%	0.1%	0.2%	97.4%	0.7%	1.1%	0.1%	0.0%
s6	0.3%	0.1%	0.6%	0.4%	1.3%	93.2%	0.2%	1.2%	2.7%
s7	0.6%	3.9%	0.5%	0.2%	2.5%	0.2%	91.8%	0.3%	0.0%
s8	0.1%	0.0%	0.3%	0.4%	0.2%	1.0%	0.1%	96.4%	1.5%
s9	0.0%	0.3%	0.8%	5.3%	0.2%	2.8%	0.1%	1.0%	89.5%

Period 3: Transition Matrix, Initial State = 3

	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	93.3%	0.1%	0.3%	2.5%	1.6%	0.3%	1.4%	0.1%	0.5%
s2	0.0%	96.6%	0.1%	0.0%	0.9%	0.4%	0.9%	0.2%	1.0%
s3	0.4%	0.0%	94.5%	2.1%	0.1%	2.3%	0.1%	0.4%	0.1%
s4	3.6%	0.1%	2.2%	92.8%	0.1%	0.4%	0.3%	0.1%	0.4%
s5	1.4%	1.0%	0.0%	0.0%	93.7%	0.4%	3.1%	0.2%	0.3%
s6	0.3%	0.5%	2.4%	0.1%	0.3%	94.5%	0.0%	1.7%	0.2%
s7	1.0%	1.5%	0.0%	0.9%	2.0%	0.2%	94.0%	0.2%	0.2%
s8	0.4%	0.5%	0.3%	0.1%	0.0%	2.5%	0.2%	93.8%	2.3%
s9	1.3%	2.6%	0.2%	0.3%	0.6%	0.4%	0.0%	4.0%	90.6%

From the transition matrices in Table 6.1, it's possible to derive all the possible transitions from one state to the next. It's also possible to understand where there is no likelihood of transitioning from a state x to a state y.

The number of temporal transactions is dependant on the number of states in the underlying data. Given that states are determined by the number of data driven clusters generated from

the input data streams, it's possible to have different number of temporal transactions in different windows of time.

For example, the Table 6.2 shows a transition matrix for periods 1 and 2 and period 1 has 7 states while period 2 has 9 derived states.

Table 6. 2 : Different Data Driven States in 2 Different Periods

Period 1: Transition Matrix, Initial State = 6

	s1	s2	s3	s4	s5	s6	s7
s1	92.3%	2.1%	0.4%	4.5%	0.0%	0.0%	0.8%
s2	1.3%	94.3%	0.3%	0.7%	2.8%	0.3%	0.3%
s3	0.0%	0.2%	94.3%	5.1%	0.0%	0.2%	0.1%
s4	3.7%	0.9%	3.1%	91.5%	0.5%	0.2%	0.2%
s5	0.2%	2.8%	0.0%	0.3%	94.3%	0.2%	2.2%
s6	0.2%	0.7%	0.0%	0.2%	0.2%	96.4%	2.3%
s7	0.4%	0.7%	0.3%	0.0%	1.8%	1.4%	95.5%

Period 2: Transition Matrix, Initial State = 3

	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	95.0%	0.1%	0.3%	0.0%	2.2%	0.4%	1.6%	0.4%	0.0%
s2	0.1%	92.4%	2.8%	2.6%	0.3%	0.1%	1.5%	0.0%	0.2%
s3	0.1%	2.5%	94.8%	0.0%	1.0%	1.0%	0.0%	0.6%	0.0%
s4	0.0%	2.3%	0.1%	94.9%	0.2%	0.9%	0.1%	1.2%	0.3%
s5	1.7%	0.4%	1.1%	0.0%	95.6%	0.0%	1.1%	0.1%	0.0%
s6	0.3%	0.9%	0.8%	1.0%	0.2%	94.0%	0.1%	0.5%	2.1%
s7	1.3%	2.2%	0.2%	0.3%	1.9%	0.0%	93.5%	0.5%	0.1%
s8	1.0%	0.1%	1.0%	1.9%	0.7%	0.6%	0.1%	94.5%	0.1%
s9	0.3%	0.9%	0.0%	0.6%	0.0%	2.8%	0.0%	0.2%	95.2%

Temporal Relationships

The transition matrix helps explain the relationship among different states. A quantification of temporal transactions is performed to highlight probabilistic relationship among different states.

Transitive Relations

There are several transitive relationships exhibited in the sample transition matrix in Table 6.2, for example using the matrix in period 2, if a patient is in state s9, then it's likely

they can transition to s4 through s6, additionally if a patient is in state s8, they can transition to s1 though s7. These are some of the transitive relations that can be derived in a transition matrix that is generated from patient data.

Equivalence Relations

The probability to transition from a state a to a state b is the same at the transition from a state c to state d, in this case state ad, and cd are termed as equivalent paths. However the underlying data might not be equivalent. Using the sample in period 1 Table 6.2, the probability of transitioning from states s3 to s4 and s5 to s2 are equal which indicates equivalence relationships among the 4 states.

6.2.4 Step 4: Predicting the Next Probably State

Predict the possibility of patient P remaining in the same state S_k at time T_{k+1}, $1<k<K+1 \in \mathbb{R}$.

In a given time window, a prediction can be generated for a patient state in the next time period. There is potential to have more than one state predicted as the final output. To select the most probably state, a ranking algorithm is used by building a ranking vector V, such that, $= \{V_1, V_2, \dots, V_n\}$, frequency $F = \{F_1, F_2, \dots, F_k\}$ where V_i represents the predicted state and F_i is the count of how many times V_i is predicted in the time period. V_i with the highest frequency count is determined as the final predicted state. If after some quantification of the predicted state based on some domain context i.e. abnormal or normal state, then it's possible to predict the likelihood of a patient being in a risky state in the next time period. This approach can then be used to generate quantifiable alerts if the number of times a

patient is predicted to be in an abnormal state reaches some prescribed threshold. Further details on methods utilized in quantification of a state is detailed in section 6.3.6.

Given that in TPRMine, data is processed from multiple patients simultaneously, there is potential that there would be multiple patterns exhibited in the data. Deriving which of these patterns are frequent is described next.

Association Rule Mining Paradigm

Utilizing patients physiological data streams as input to a mining algorithm, frequent patterns can be determined. To accomplish this, patient data is modeled as transaction sets and applying frequent mining process in that data generates patterns and their associated support and confidence as discussed in section 5.2.5 and 5.2.6. A rank of the resulting rules shows the prevalence of patterns in the data. One can then select which are the top n rules based on some user defined threshold.

As noted in the literature review in Chapter 2 and 3, there are challenges to utilizing the current algorithms to generated frequent patterns in that (a) if the amount of data is too large, this results in to many frequent rules generated that makes it very hard to quantify those that are important. (B) Huge volume of data means that there are many iterations needed to process the massive data.

This thesis proposes that a process should be incorporated to downsample a data stream and in so doing introduce information about the behaviour of the stream. These downsampled streams representing higher abstractions of behaviours are provided as input

to the mining algorithm in order to overcome the challenges mentioned above. To this respect, this thesis has utilized a process to mining frequent patterns on data captured for a specific period of time. This is accomplished using the following module.

6.2.5 Step 5: Identify the Frequent Patterns

Following the principles detailed in section 5.3.1, a module for frequent pattern identification is implemented as follows;

Module

1. For each period of time
2. Replace the physiological stream value with it's corresponding cluster centroid
3. Generate the transactions
4. Identify the frequent items sets from the generated transactions
5. Generate the support and confidence of each frequent sets

End.

This approach allows generation of rules of the form *P(X, Period, W)* and *Q (X, Y)* $\rightarrow$ *State (X, Z)*

Where *X* is the set of patients, *P* and *Q* are the predicate variables representing the physiological dimensions that can be instantiated with relevant data values *W, Y, Z*.

For example, if there is a set of data captured in a specific time window, if a subset of that data comprised of all the physiological variables i.e. HR, RR, SpO_2, PulsePleth (PulseP), States

(generated from clustering) and corresponding data that is captured 5 minutes into the hour, then we can identify the following sample rules.

(i) $HR(100,120,150)\text{^}RR(20,40,60)\text{^}SpO_2(90,92,91,93)\text{^}PulseP(120,140,110) \rightarrow States(4,5,6)$

(ii) $HR(40,70,88)\text{^}RR(20,40,60)\text{^}SpO_2(90,92,91,93)\text{^}PulseP(110,130,150) \rightarrow States(1,3)$.

6.2.6 Step 6: Quantification of Temporal Relationships

Within the health informatics application, an additional component is added to quantify the clinical meaning of temporal relationship identified. A scoring approach similar to the work in (McGregor C. , James, Eklund, & et al., 2013) is adopted.

To facilitate this process a higher level abstraction of the cluster means of HR, RR and SpO_2 can be completed using some clinical guidelines similar to the work in (Thommandram, Pugh, Eklund, McGregor, & James, 2013) and (Baker, et al., 2012). The researchers provided the following clinical guidelines on what is considered normal or abnormal values (if values are high or low). For HR normal range (100 to 160), Low (< 100), high (> 160), for RR normal range (30 – 60), low (<30), high (>60) and for SpO_2 normal range (88 -92), low (<=85), high (>=94). Utilizing this clinical abstraction of physiological data enables quantification of the states a patient transitions to in a given period.

Applying abstraction to the sample data in Figure 5.7(a) results to the data similar to the results in Table 6.3 where new variables are derived {HRSTATUS, RRSTATUS, SPO2STATUS} to represent the clinical abstraction of each of {HRVALUE, RRVALUE, SPO2VALUE}.

For example the red colored row represents an abstraction values in cluster 6 such that; HRVALUE = 112.4 is between 100 – 160 then HRSTATUS = normal, RRVALUE = 63.6 is > 60 then RRSTATUS is abnormal and SPO2VALUE = 99.9 which is > 94 then SPO2STATUS is abnormal. In this thesis, if a cluster contains an abnormal abstraction, then the cluster is identified as containing some risk. The number of risky or non-risky abstractions of a given cluster could determine if a path associated with a cluster (state) are risky or not.

This allows identifying for a given time period, what are the more risky paths and the premise in this thesis is that, if a patient is seen with these paths then they are at high risk of some medical conditions.

Table 6. 3: A sample Clinical Abstractions of Temporal Clusters

STARTTIME	ENDTIME	HRVALUE	RRVALUE	SPO2VALUE	PULSEPVALUE	CLUSTER	HRSTATUS	RRSTATUS	SPO2STATUS
1537844400	1537844759	155.8996318	49.56939298	92.2980799	154.5058732	1	NORMAL	NORMAL	ABNORMAL
1537844400	1537844759	123.5106248	42.58595386	98.00582096	123.4529103	2	NORMAL	NORMAL	ABNORMAL
1537844400	1537844759	139.3871747	36.39869079	98.49359277	139.1405029	3	NORMAL	NORMAL	ABNORMAL
1537844400	1537844759	167.0630645	37.70931072	98.21318691	165.2604679	4	NORMAL	NORMAL	ABNORMAL
1537844400	1537844759	153.7965764	49.40339362	97.95238589	153.2158818	5	NORMAL	NORMAL	ABNORMAL
1537844400	1537844759	112.3698433	63.59936915	99.93739404	112.1377488	6	NORMAL	ABNORMAL	ABNORMAL
1537844400	1537844759	164.729402	32.5316984	96.14700002	164.5290156	7	NORMAL	NORMAL	ABNORMAL
1537844400	1537844759	141.8825802	57.88152108	95.49109203	141.6370383	8	NORMAL	NORMAL	ABNORMAL
1537844400	1537844759	165.1016759	57.87013353	87.65886825	164.6124394	9	NORMAL	NORMAL	NORMAL

6.2.7 Step 7: Classification with Frequent Temporal Clusters

Classification has been adopted in many studies where objects with similar characteristics are placed in groups. With respect to health informatics, similar principles can be adopted, where patients with similar physiological features can be characterised into similar groups.

To facilitate this kind of characterization a scoring approach is applied to classification of physiological data streams as follows.

Temporal Abstraction Scoring

The TPRMine algorithm first step generated data driven clusters, in this module, a quantification of the variability of those clusters is completed by using a clinical abstraction score. This process works by introducing clinical thresholds on some physiological data streams based on the clinical guidelines presented earlier and described next.

In particular, the clinical threshold for normal or abnormal vital status are adopted based on the work in (Pugh, Thommandram, Ng, & McGregor, 2013).

A ranking mechanism is adopted to quantify the risk level for each of the 3 physiological streams (HR, RR, SpO_2) using scores. This thesis utilizes four levels of vital scoring as flows. If all the 3 features (HR, RR and SpO_2) are in normal range, then the vital score is 0, if either one of these is abnormal then the score is 1, if either 2 are abnormal then the score is 2 and if all three are abnormal then the score is 3. The red colored row in Table 6.4 indicates the states where vital scores are equal to 2. A similar scoring mechanism was detailed in (McGregor et al., 2013) where researchers utilized 4 different data streams and assigned 4 levels based on some quantified threshold to each raw data streams separately. In this thesis, a score is assigned based on combination of the elements in a cluster (state) and not to raw data streams. As a state can contain multiple data streams, the scoring per state can simultaneously incorporate all the underlying context in the data.

Note that this is a scalable approach where different rules can be integrated based on the physiological data streams utilized.

The temporal abstraction scoring results for the sample data set is shown in Table 6.4 as a matrix containing an added variable (VITAL_SCORE) to represent the ranked score of the specific cluster.

Table 6. 4 : State Quantification with Clinical Temporal Abstractions

STARTTIME	ENDTIME	HRVALUE	RRVALUE	SPO2VALUE	PULSEPVALUE	CLUSTER	HRSTATUS	RRSTATUS	SPO2STATUS	VITALSCORE
1537837200	1537837559	175.2147507	38.66383421	97.03794455	174.9420776	1	NORMAL	NORMAL	ABNORMAL	1
1537837200	1537837559	153.9800086	49.51901407	95.58620506	153.8132803	2	NORMAL	NORMAL	ABNORMAL	1
1537837200	1537837559	134.2854856	52.45581483	95.0971569	134.1403892	3	NORMAL	NORMAL	ABNORMAL	1
1537837200	1537837559	130.4085026	42.81288894	99.76924678	129.9842361	4	NORMAL	NORMAL	ABNORMAL	1
1537837200	1537837559	153.9218798	63.93266209	91.75905022	153.6131991	5	NORMAL	ABNORMAL	NORMAL	1
1537837200	1537837559	133.9814355	37.64027068	98.78239853	133.8291852	6	NORMAL	NORMAL	ABNORMAL	1
1537837200	1537837559	158.7721863	56.5356854	97.49599043	158.5871394	7	NORMAL	NORMAL	ABNORMAL	1
1537837560	1537837859	142.4994752	42.19848762	98.45825337	141.8768183	1	NORMAL	NORMAL	ABNORMAL	1
1537837560	1537837859	151.9256865	54.3640605	94.24332684	150.6434428	2	NORMAL	NORMAL	ABNORMAL	1
1537837560	1537837859	163.6977883	60.15946746	97.48127211	163.4652984	3	NORMAL	ABNORMAL	ABNORMAL	2
1537837560	1537837859	144.4572422	55.66810481	96.00190663	144.1892143	4	NORMAL	NORMAL	ABNORMAL	1
1537837560	1537837859	161.7979885	60.80295637	87.88403143	161.0225316	5	NORMAL	ABNORMAL	NORMAL	1
1537837560	1537837859	134.5590251	35.47808227	98.41288939	134.3345179	6	NORMAL	NORMAL	ABNORMAL	1
1537837560	1537837859	131.8499706	54.972592	91.80355685	131.6702217	7	NORMAL	NORMAL	NORMAL	0
1537837560	1537837859	171.5816492	32.29472465	96.50983439	171.4284906	8	NORMAL	NORMAL	ABNORMAL	1
1537837560	1537837859	104.9628193	38.93064303	98.5229589	104.7680628	9	NORMAL	NORMAL	ABNORMAL	1
1537837860	1537838159	171.1555788	27.39945727	95.89902907	171.0086298	1	NORMAL	ABNORMAL	ABNORMAL	2
1537837860	1537838159	160.1165932	46.1057092	98.87526967	159.8357446	2	NORMAL	NORMAL	ABNORMAL	1
1537837860	1537838159	143.0596426	53.74040198	95.31717099	142.6781697	3	NORMAL	NORMAL	ABNORMAL	1
1537837860	1537838159	148.1844026	46.92554234	91.21635242	147.1123793	4	NORMAL	NORMAL	NORMAL	0
1537837860	1537838159	166.9882772	82.62816101	94.71971702	166.6230139	5	NORMAL	ABNORMAL	ABNORMAL	2
1537837860	1537838159	106.3340603	38.81596085	98.57781001	106.2012601	6	NORMAL	NORMAL	ABNORMAL	1
1537837860	1537838159	131.1230335	34.14248875	98.33637567	130.9890668	7	NORMAL	NORMAL	ABNORMAL	1
1537838160	1537838459	108.8527838	44.87199327	99.17473934	108.6831154	1	NORMAL	NORMAL	ABNORMAL	1
1537838160	1537838459	127.0272495	73.2009649	97.81426097	127.2049282	2	NORMAL	ABNORMAL	ABNORMAL	2
1537838160	1537838459	161.4950448	56.47477379	96.11072669	161.1753915	3	NORMAL	NORMAL	ABNORMAL	1
1537838160	1537838459	142.1031721	38.29245059	98.88034447	141.4036877	4	NORMAL	NORMAL	ABNORMAL	1
1537838160	1537838459	126.456598	33.88113375	98.32010154	126.3848678	5	NORMAL	NORMAL	ABNORMAL	1
1537838160	1537838459	158.1893789	52.03122212	90.50428043	157.8648821	6	NORMAL	NORMAL	NORMAL	0
1537838160	1537838459	139.7787972	48.54573432	95.24609346	139.4087477	7	NORMAL	NORMAL	ABNORMAL	1
1537838160	1537838459	148.3426042	47.98991148	91.03809913	147.7384934	8	NORMAL	NORMAL	NORMAL	0
1537838160	1537838459	130.7403297	52.24029959	95.18211825	130.6438607	9	NORMAL	NORMAL	ABNORMAL	1

Next is to use the vitals scoring of states for classifying a patient, the following patient scoring module is introduced.

Module for Patient Scoring

Begin:

For each period hour (PeriodHR)

Identify the clinical abstractions of the cluster means for each of the physiological data streams

For each cluster set combination

Assign the Vital score to each cluster set

For each patient

Identify the number times the patient is in each vital score → VTScore

If VTScore for patient P is beyond a threshold K

Flag Patient P is high risk other wise low risk

End.

The TPRMine algorithm is applied to data for 10 unique patients (P_1 … P_{10}). After the patient scoring process is completed, three of the patients (P_1, P_3 and P_7) have the highest risk scores while the rest of the patients have low risk scores. Then using frequent pattern mining, all the association rules that have the states where the high risk patient transitioned to are identified in the mined rules. This results to 2 sets of rules which can be inferred as high risk or low risks.

With this approach it is then possible to understand the percentage of times a patient is in a specific vital score within a given period of time i.e. 1hr. Table 6.5 below shows 15 different patients level of scoring in 6 consecutive hours.

Table 6. 5 : Patient Vital Scoring for a Period of 6 Hours

PeriodHr	VitalScore	1001	1002	1003	1004	1005	1006	1007	1008	1009	1010	1011	1012	1013	1014	1015
1	0	16.22%	23.91%	30.43%	8.70%	3.03%	16.18%	12.00%	23.81%	11.36%	20.00%	21.15%	22.45%	19.64%	24.44%	17.24%
	1	81.08%	67.39%	60.87%	86.96%	87.88%	67.65%	78.00%	69.05%	75.00%	73.33%	65.38%	65.31%	71.43%	62.22%	65.52%
	2	2.70%	8.70%	8.70%	4.35%	9.09%	16.18%	10.00%	7.14%	13.64%	6.67%	13.46%	12.24%	8.93%	13.33%	17.24%
2	0	0.00%	13.64%	18.75%	10.00%	1.96%	10.81%	5.88%	13.21%	8.33%	10.87%	12.07%	13.21%	7.94%	12.07%	13.11%
	1	92.68%	81.82%	78.13%	90.00%	88.24%	83.78%	80.39%	83.02%	84.72%	82.61%	82.76%	79.25%	87.30%	77.59%	78.69%
	2	7.32%	4.55%	3.13%	0.00%	9.80%	5.41%	13.73%	3.77%	6.94%	6.52%	5.17%	7.55%	4.76%	10.34%	8.20%
3	0	6.90%	10.17%	14.29%	10.34%	0.00%	8.14%	3.77%	10.29%	8.75%	11.76%	5.77%	6.56%	8.97%	21.05%	10.67%
	1	89.66%	84.75%	85.71%	87.93%	100.00%	87.21%	86.79%	86.76%	85.00%	85.29%	86.54%	85.25%	87.18%	78.95%	82.67%
	2	3.45%	5.08%	0.00%	1.72%	0.00%	4.65%	9.43%	2.94%	6.25%	2.94%	7.69%	8.20%	3.85%	0.00%	6.67%
4	0	9.52%	16.42%	35.71%	13.11%	2.44%	14.77%	13.33%	13.89%	7.79%	18.52%	9.46%	11.84%	12.64%	16.36%	14.67%
	1	80.95%	73.13%	50.00%	78.69%	92.68%	72.73%	65.00%	76.39%	81.82%	66.67%	82.43%	73.68%	78.16%	61.82%	76.00%
	2	9.52%	10.45%	14.29%	8.20%	4.88%	12.50%	21.67%	9.72%	10.39%	14.81%	8.11%	13.16%	9.20%	20.00%	9.33%
	3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.32%	0.00%	1.82%	0.00%
5	0	20.69%	13.04%	19.23%	10.87%	2.86%	17.33%	15.25%	23.19%	16.18%	18.18%	10.20%	20.31%	20.00%	20.75%	17.19%
	1	65.52%	80.43%	65.38%	76.09%	97.14%	70.67%	69.49%	68.12%	73.53%	71.21%	85.71%	68.75%	73.33%	66.04%	71.88%
	2	13.79%	6.52%	15.38%	13.04%	0.00%	12.00%	15.25%	8.70%	10.29%	10.61%	4.08%	10.94%	6.67%	13.21%	10.94%
6	0	12.70%	16.81%	47.06%	23.30%	4.76%	14.20%	15.75%	15.79%	14.17%	16.81%	16.81%	16.28%	14.40%	17.05%	18.46%
	1	82.54%	77.88%	45.10%	71.84%	90.48%	80.86%	79.53%	78.95%	79.17%	76.11%	78.99%	77.52%	79.20%	76.74%	75.38%
	2	4.76%	5.31%	7.84%	4.85%	4.76%	4.94%	4.72%	5.26%	6.67%	7.08%	4.20%	6.20%	6.40%	6.20%	6.15%

Finally a patient knowledge base can be created that includes: the state transitions, cluster means, frequent patterns, the vital scores, and percentage of time a patient stays at a risky or non risky state. With the frequent patterns, it's possible to derive rules that are quantified by the associated state score. This becomes a classification for identifying patterns that are frequent based on the underlying clinical threshold.

This process described in section 6.2 fits into the enhancement on Artemis as a knowledge discovery and representation framework with respect to physiological data streams from a diverse number of health care domains. To support this process several data structures have been developed to allow further analytical questions that can be formulated in addition to easily querying the data structured generated. To support this process, the flow chart in Figure 6.6 shows the sequence of steps adopted.

6.2.8 Additional Results

Although a lot of the details outlined in the TPRMine outputs have been discussed within each of the steps above, this section presents additional results generated within specific modules during the TPRMine process flow that can support retrospective knowledge discovery.

Clustering Generated from Select Time Intervals

Data captured at each temporal window is provided as input to the Mclust algorithm which then generates the model derived clusters based on data content. As earlier described in section 5.3.1, the use of model driven clustering allows grouping independent data into components referred to as clusters. In this thesis, as the physiological data processed has no

predefined categories or classes, the log likelihood, BIC, and ICL generated from Mclust are utilized for selecting the best model for a given input data (Fraley, Adrian, Murphy, & Scrucca, 2012).

Figure 6.6 presents results of Mclust algorithm after processing physiological data captured in a given temporal window. The best model derived has 9 unique clusters with log likelihood -147265, BIC is -295772.3 and ICL = -300593.5. Figure 6.7 shows the best model selection using BIC and ICL while the resulting classification using the model is presented in Figure 6.8.

Figure 6. 6: Results of Mclust Algorithm
Provided as input for modelling are 4 Physiological Data Streams captured from 36 unique patients, N = 10632 generating 9 clusters

```
Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 9
components:

 log.likelihood     n  df       BIC       ICL
        -147265 10632 134 -295772.3 -300593.5

Clustering table:
   1    2    3    4    5    6    7    8    9
2031  972 1312  751  822 1304  963  919 1558

Mixing probabilities:
         1          2          3          4          5          6          7
0.17143806 0.08771836 0.11609416 0.08844637 0.07782699 0.14085000 0.07965154
         8          9
0.09020890 0.14776560

Means:
                  [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
HRVALUE      146.29431 160.70766 127.29668 157.04335 171.89941 145.12117
RRVALUE       47.63646  60.75539  49.37286  48.65457  60.46248  37.53138
SPO2VALUE     99.68710  94.91669  97.09372  92.71802  97.34970  97.03908
PULSEPVALUE  146.04820 160.52634 126.99229 155.54351 171.46447 143.48709
                  [,7]      [,8]      [,9]
HRVALUE      168.56388 175.01888 142.29732
RRVALUE       40.98759  53.69699  56.82528
SPO2VALUE     93.53688  84.63732  92.87913
PULSEPVALUE  167.21476 174.50149 141.96283
```

Figure 6. 7: Plots of BIC and ICL Best Model Selection Criteria

(a) Represents the model selected using ICL and (b) represents the model selection with BIC.

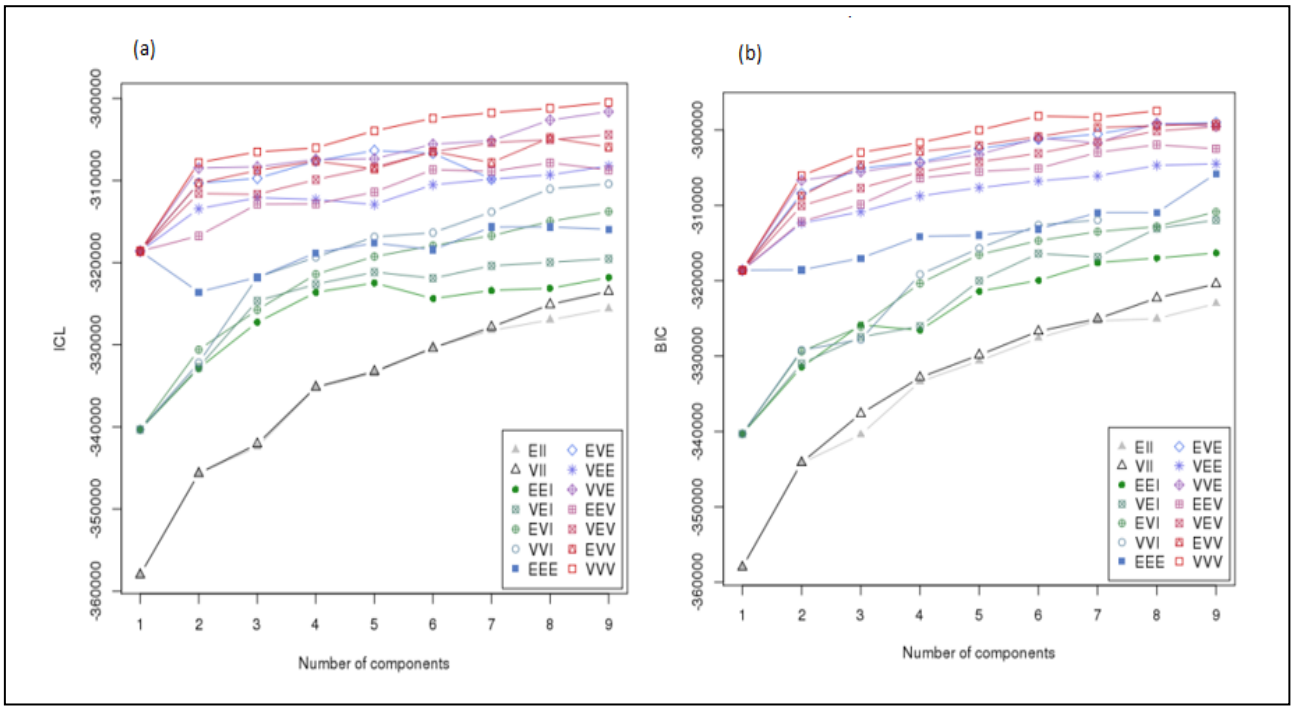
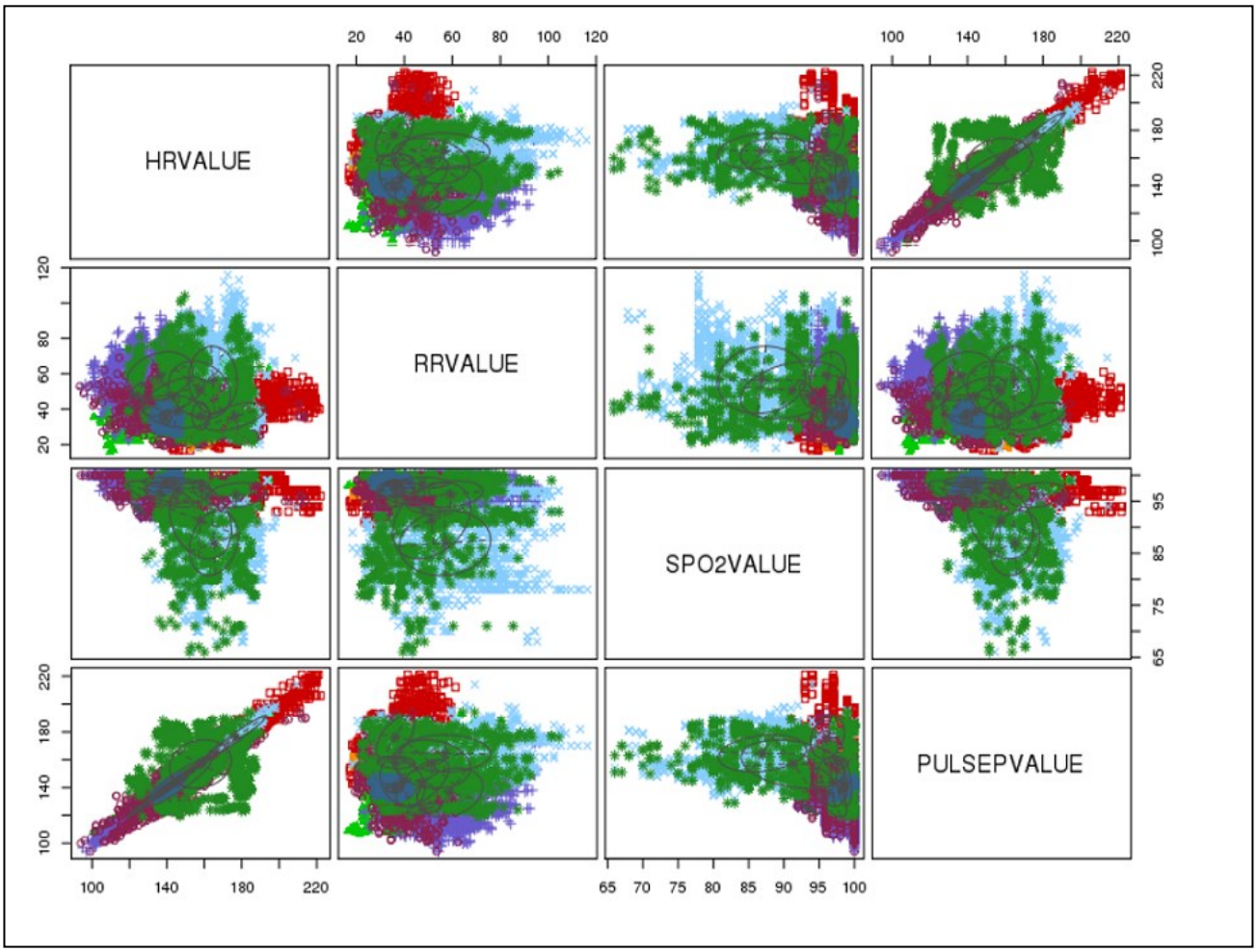


Figure 6. 8: Classification Results on 4 Variables in a Specific Time Interval



After clustering, it's possible to understand the many states patients can transition to in any given period of time. Figure 6.9 shows the state transitions for 6 randomly selected patients on data streams captured for 6 hours. Note that each hour a patient can transition to multiple states.

Figure 6. 9: State Transitions for 6 Patients in a 6 Hour Period.

Each of the patients are identified as P1 to P6. Y-Axis presents the states generated and X-Axis represents the time window of a specific state transition.

As it's not easy to quantify what is different about patients using the state transitions as shown in Figure 6.9, further break down of these results generates the details in Figure 6.10 (a) and (b). In this case it's possible to see the total number of state transitions for a given period for all the patients i.e. every 5, 10, 15.., 60 minute intervals. (b) Shows the total no. of state transitions for in the 1st hour. It is clear patients do vary in the number of states they transition to in a given time interval. Figure 6.11 shows results after processing data for the 6th hour.

Figure 6. 10: A Breakdown of State Transitions of 6 Patients in Hour 1.
Each 5 Minute intervals shows the changes in the number of state transition per patient over an hour. There is also a difference on number of transitions within each interval among the 6 patients. Patient P6 has the highest number of state transitions within the hour while P1 has the lowest.

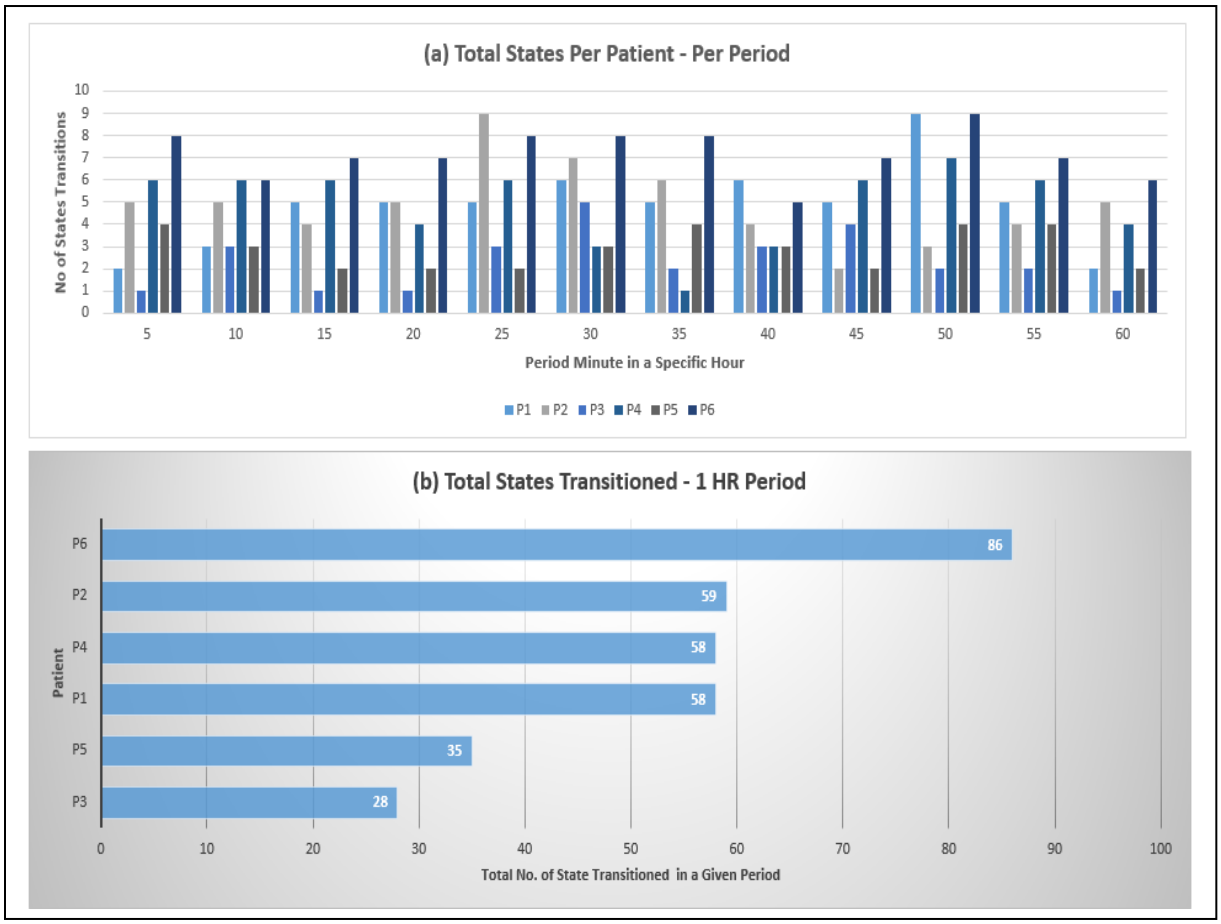
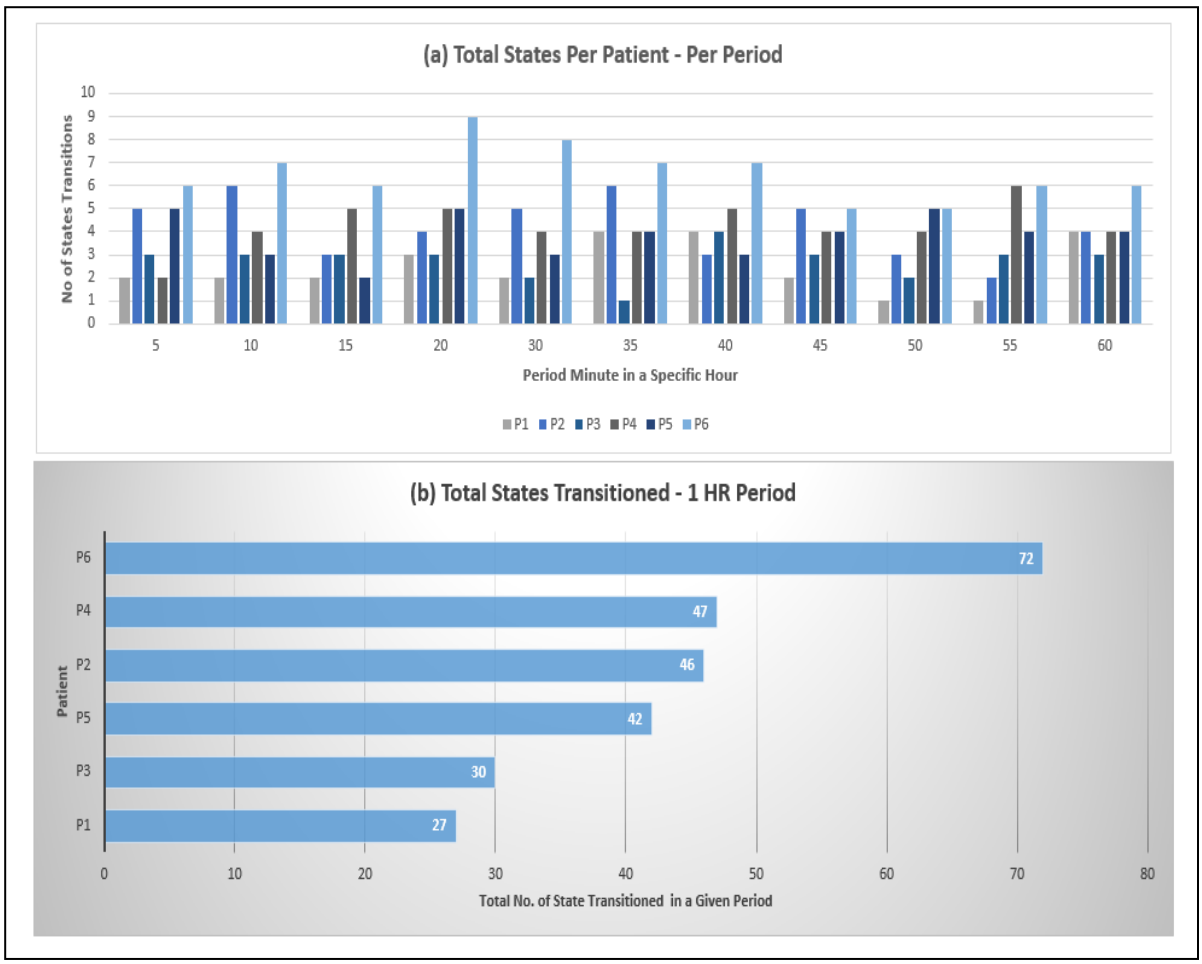


Figure 6. 11: A Breakdown of State Transitions of 6 Patients after 6 Hours.
Each 5 Minute intervals shows the changes in the number of state transition per patient in the hour. There is also a difference on number of transitions within each interval among the 6 patient with P6 having the highest number of state transitions in the hour while P1 has the lowest transitions.



Identification of the Transition Probabilities

TPRMine allows the generation of transition matrices that allows identification of what is the likely hood of transitioning from a state S1 to S2. Figure 6.12 shows the possible state transitions from a state s6 in a select time interval.

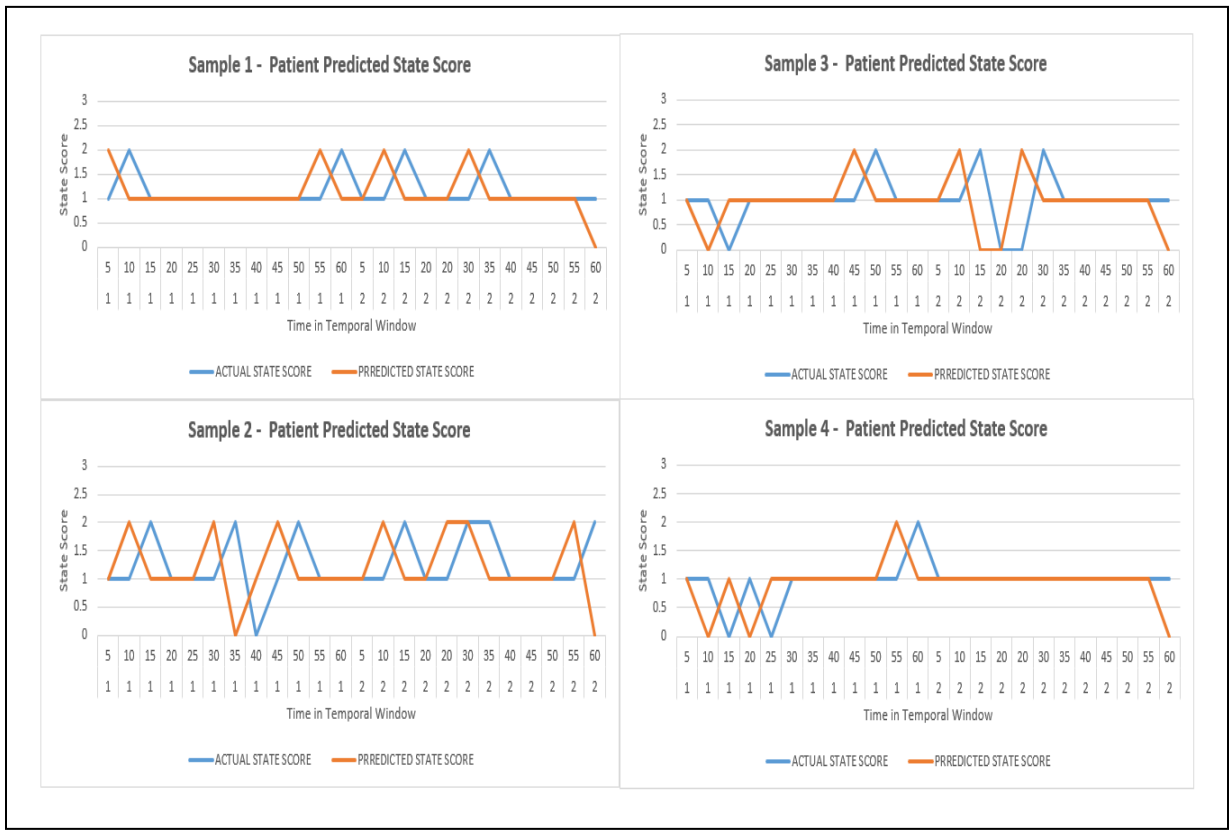
Figure 6. 12: Possible State Transition from a State in a Select Temporal Window

	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	94.8%	0.1%	0.1%	2.6%	1.4%	0.2%	0.3%	0.3%	0.1%
s2	0.0%	92.4%	3.3%	0.0%	0.0%	0.4%	0.0%	0.8%	3.2%
s3	0.1%	2.8%	92.7%	0.3%	0.1%	2.4%	0.0%	0.9%	0.6%
s4	5.2%	0.3%	0.4%	88.1%	0.2%	0.5%	0.1%	0.9%	4.3%
s5	1.8%	0.1%	0.4%	0.2%	93.3%	0.1%	3.2%	0.8%	0.1%
s6	0.0%	0.4%	1.8%	0.5%	0.0%	96.2%	0.0%	0.9%	0.1%
s7	0.2%	0.0%	0.1%	0.2%	3.2%	0.0%	96.3%	0.0%	0.0%
s8	1.1%	2.1%	2.0%	1.4%	0.4%	0.6%	0.0%	91.5%	0.9%
s9	0.2%	3.0%	0.4%	3.8%	0.3%	0.0%	0.0%	1.5%	90.8%

Predicting the Next Probable State

After identification of possible state transitions, next is to predict the next probable state. TPRMine is able to utilize the data in a current temporal window to predict the probable state of each patient in the following temporal window. As described in section 6.2.4, a ranking algorithm is used by building a ranking vector V, such that $= \{V_1, V_2, \dots, V_n\}$, frequency $F = \{F_1, F_2, \dots, F_k\}$ where V_i represents the predicted state and F_i is the count of how many times V_i is predicted in the time period. V_i with the highest frequency count is determined as the final predicted state. Application of this ranking algorithm shows the method is able to accurately predict 80% of the next probable state after processing 11 hours of data captured from NICU 36 patients with a 96% confidence interval. A random sample of this prediction is shown in Figure 6.13 where a patient's actual state score and predicted state score are displayed. Based on quantification of states with scores, the higher the score the higher the risk. With this, it possible to count how many times a patient is predicted to be in a risky state.

Figure 6. 13: Predicted Probably Next State on Sample Patients



Identification of the Frequent Patterns Within the Time Interval

For each time period, a set of frequent and infrequent patterns are generated. Tables 6.6, 6.7 and 6.8 shows the frequent patterns in a time frame with varying lift, confidence and support levels which are calculated using the technique detailed in section 5.2.6. In particular each row represents a rule R of the form $J \rightarrow L$. For example, using the first row, is a rule represented with the first 2 variables {RHS=Right Hand Side, LHS= Left Hand Side) therefore J = LHS and L = RHS forming the rule $(LHS \rightarrow RHS)$=([HRVALUE $= \{134,153\}$, SPO2VALUE $= \{95,97\}$, Cluster $= \{3,5\}$] $\rightarrow$ [PULESPLETH $= \{134,153\}$]).

The support of the rule indicates the percentage of the transactions containing the elements in the rule. As described in section 5.2.6, the confidence and lift are as follows;

$$\text{Confidence } ((J \rightarrow L) = \text{probability } Pr(L|J) = \frac{support\ count\ (J \cup L)}{Support\ Count\ (J)} = \frac{support\ count\ (LHS \cup RHS)}{Support\ Count\ (LHS)} \quad (15)$$

The correlation between the LHS and RHS is referred to as a lift (lf) and calculated using probabilities such that for a rule $(J \rightarrow L)$, $lf_{J \rightarrow L} = \left(\frac{Pr(J \cup L)}{pr(J)P(L)}\right)$. If $lf_{J \rightarrow L} < 1$ then occurrence of *J* is negatively correlated with occurrence of *L*, if $lf_{J \rightarrow L} > 1$ then *J* and *L* are positively correlated and if $lf_{J \rightarrow L} = 1$ indicates *J* and *L* occurs independent of each other (Han & Kamber, 2001). The rule in row 1 has a lift = 7 which indicate that the LHS and RHS are positively correlated.

Table 6. 6: Rules Sets with a Specified Lift and Confidence Level.
A Subset of Rules with > 66 % Confidence

LHS	RHS	SUPPORT	CONFIDENT	LIFT
{HRVALUE=[134,153),SPO2VALUE=[95,97),Cluster=[3,5)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{SPO2VALUE=[95,97),PULSEPVALUE=[134,153),Cluster=[3,5)}	{HRVALUE=[134,153)}	0.142857143	1	7
{HRVALUE=[134,153),RRVALUE=[52,63],SPO2VALUE=[95,97)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],SPO2VALUE=[95,97),PULSEPVALUE=[134,153)}	{HRVALUE=[134,153)}	0.142857143	1	7
{HRVALUE=[134,153),RRVALUE=[52,63],Cluster=[3,5)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],PULSEPVALUE=[134,153),Cluster=[3,5)}	{HRVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],SPO2VALUE=[95,97),Cluster=[3,5)}	{HRVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],SPO2VALUE=[95,97),Cluster=[3,5)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{HRVALUE=[134,153),RRVALUE=[52,63],SPO2VALUE=[95,97),Cluster=[3,5)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],SPO2VALUE=[95,97),PULSEPVALUE=[134,153),Cluster=[3,5)}	{HRVALUE=[134,153)}	0.142857143	1	7
{SPO2VALUE=[97,99],Cluster=[1,3)}	{RRVALUE=[37,42)}	0.142857143	1	3.5
{RRVALUE=[37,42),PULSEPVALUE=[153,174]}	{Cluster=[1,3)}	0.142857143	1	3.5
{HRVALUE=[153,175],RRVALUE=[37,42)}	{Cluster=[1,3)}	0.142857143	1	3.5
{HRVALUE=[130,134),RRVALUE=[37,42)}	{PULSEPVALUE=[129,134)}	0.142857143	1	3.5
{RRVALUE=[37,42),PULSEPVALUE=[129,134)}	{HRVALUE=[130,134)}	0.142857143	1	3.5
{RRVALUE=[37,42),Cluster=[5,7]}	{HRVALUE=[130,134)}	0.142857143	1	3.5
{HRVALUE=[130,134),Cluster=[5,7]}	{RRVALUE=[37,42)}	0.142857143	1	3.5
{RRVALUE=[37,42),Cluster=[5,7]}	{PULSEPVALUE=[129,134)}	0.142857143	1	3.5
{SPO2VALUE=[95,97),PULSEPVALUE=[134,153)}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{PULSEPVALUE=[134,153),Cluster=[3,5)}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{RRVALUE=[52,63],SPO2VALUE=[91,95)}	{Cluster=[5,7]}	0.142857143	1	2.333333
{SPO2VALUE=[91,95),Cluster=[5,7]}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{SPO2VALUE=[91,95),PULSEPVALUE=[153,174]}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{HRVALUE=[153,175],SPO2VALUE=[91,95)}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{SPO2VALUE=[91,95),PULSEPVALUE=[153,174]}	{Cluster=[5,7]}	0.142857143	1	2.333333
{HRVALUE=[153,175],SPO2VALUE=[91,95)}	{Cluster=[5,7]}	0.142857143	1	2.333333
{RRVALUE=[52,63]}	{Cluster=[5,7]}	0.285714286	0.666666667	1.555556
{Cluster=[5,7]}	{RRVALUE=[52,63]}	0.285714286	0.666666667	1.555556
{RRVALUE=[52,63]}	{PULSEPVALUE=[153,174]}	0.285714286	0.666666667	1.166667
{RRVALUE=[52,63]}	{HRVALUE=[153,175]}	0.285714286	0.666666667	1.166667
{Cluster=[5,7]}	{SPO2VALUE=[97,99]}	0.285714286	0.666666667	1.166667
{Cluster=[5,7]}	{PULSEPVALUE=[153,174]}	0.285714286	0.666666667	1.166667
{Cluster=[5,7]}	{HRVALUE=[153,175]}	0.285714286	0.666666667	1.166667

Table 6. 7: Rules Sets with a Specified Lift and Confidence Level.
A Subset of Rules with > 66 % Confidence

LHS	RHS	SUPPORT	CONFIDENT	LIFT
{HRVALUE=[134,153),SPO2VALUE=[95,97),Cluster=[3,5)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{SPO2VALUE=[95,97),PULSEPVALUE=[134,153),Cluster=[3,5)}	{HRVALUE=[134,153)}	0.142857143	1	7
{HRVALUE=[134,153),RRVALUE=[52,63],SPO2VALUE=[95,97)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],SPO2VALUE=[95,97),PULSEPVALUE=[134,153)}	{HRVALUE=[134,153)}	0.142857143	1	7
{HRVALUE=[134,153),RRVALUE=[52,63],Cluster=[3,5)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],PULSEPVALUE=[134,153),Cluster=[3,5)}	{HRVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],SPO2VALUE=[95,97),Cluster=[3,5)}	{HRVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],SPO2VALUE=[95,97),Cluster=[3,5)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{HRVALUE=[134,153),RRVALUE=[52,63],SPO2VALUE=[95,97),Cluster=[3,5)}	{PULSEPVALUE=[134,153)}	0.142857143	1	7
{RRVALUE=[52,63],SPO2VALUE=[95,97),PULSEPVALUE=[134,153),Cluster=[3,5)}	{HRVALUE=[134,153)}	0.142857143	1	7
{SPO2VALUE=[97,99],Cluster=[1,3)}	{RRVALUE=[37,42)}	0.142857143	1	3.5
{RRVALUE=[37,42),PULSEPVALUE=[153,174]}	{Cluster=[1,3)}	0.142857143	1	3.5
{HRVALUE=[153,175],RRVALUE=[37,42)}	{Cluster=[1,3)}	0.142857143	1	3.5
{HRVALUE=[130,134),RRVALUE=[37,42)}	{PULSEPVALUE=[129,134)}	0.142857143	1	3.5
{RRVALUE=[37,42),PULSEPVALUE=[129,134)}	{HRVALUE=[130,134)}	0.142857143	1	3.5
{RRVALUE=[37,42),Cluster=[5,7]}	{HRVALUE=[130,134)}	0.142857143	1	3.5
{HRVALUE=[130,134),Cluster=[5,7]}	{RRVALUE=[37,42)}	0.142857143	1	3.5
{RRVALUE=[37,42),Cluster=[5,7]}	{PULSEPVALUE=[129,134)}	0.142857143	1	3.5
{SPO2VALUE=[95,97),PULSEPVALUE=[134,153)}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{PULSEPVALUE=[134,153),Cluster=[3,5)}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{RRVALUE=[52,63],SPO2VALUE=[91,95)}	{Cluster=[5,7]}	0.142857143	1	2.333333
{SPO2VALUE=[91,95),Cluster=[5,7]}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{SPO2VALUE=[91,95),PULSEPVALUE=[153,174]}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{HRVALUE=[153,175],SPO2VALUE=[91,95)}	{RRVALUE=[52,63]}	0.142857143	1	2.333333
{SPO2VALUE=[91,95),PULSEPVALUE=[153,174]}	{Cluster=[5,7]}	0.142857143	1	2.333333
{HRVALUE=[153,175],SPO2VALUE=[91,95)}	{Cluster=[5,7]}	0.142857143	1	2.333333
{RRVALUE=[52,63]}	{Cluster=[5,7]}	0.285714286	0.666666667	1.555556
{Cluster=[5,7]}	{RRVALUE=[52,63]}	0.285714286	0.666666667	1.555556
{RRVALUE=[52,63]}	{PULSEPVALUE=[153,174]}	0.285714286	0.666666667	1.166667
{RRVALUE=[52,63]}	{HRVALUE=[153,175]}	0.285714286	0.666666667	1.166667
{Cluster=[5,7]}	{SPO2VALUE=[97,99]}	0.285714286	0.666666667	1.166667
{Cluster=[5,7]}	{PULSEPVALUE=[153,174]}	0.285714286	0.666666667	1.166667
{Cluster=[5,7]}	{HRVALUE=[153,175]}	0.285714286	0.666666667	1.166667

Table 6. 8: Rules sets with a Confidence Level > 67% and Support < 22 %

LHS	RHS	SUPPORT	CONFIDENT	LIFT
{HRVALUE=[139,154),SPO2VALUE=[95.3,97.3),PULSEPVALUE=[139,154),Cluster=[4,6)}	{RRVALUE=[54.3,60]}	0.111111111	1	3
{HRVALUE=[139,154),RRVALUE=[54.3,60],SPO2VALUE=[95.3,97.3),Cluster=[4,6)}	{PULSEPVALUE=[139,154]}	0.111111111	1	3
{RRVALUE=[54.3,60],SPO2VALUE=[95.3,97.3),PULSEPVALUE=[139,154),Cluster=[4,6)}	{HRVALUE=[139,154)}	0.111111111	1	3
{HRVALUE=[104,139),RRVALUE=[40.7,54.3),SPO2VALUE=[87,95.3),PULSEPVALUE=[104,139)}	{Cluster=[6,9]}	0.111111111	1	3
{HRVALUE=[104,139),RRVALUE=[40.7,54.3),SPO2VALUE=[87,95.3),Cluster=[6,9]}	{PULSEPVALUE=[104,139)}	0.111111111	1	3
{HRVALUE=[104,139),RRVALUE=[32,40.7),SPO2VALUE=[97.3,98]}	{PULSEPVALUE=[104,139)}	0.222222222	1	3
{HRVALUE=[104,139),SPO2VALUE=[97.3,98],PULSEPVALUE=[104,139)}	{RRVALUE=[32,40.7)}	0.222222222	1	3
{RRVALUE=[32,40.7),SPO2VALUE=[97.3,98],PULSEPVALUE=[104,139)}	{HRVALUE=[104,139)}	0.222222222	1	3
{HRVALUE=[104,139),RRVALUE=[32,40.7),PULSEPVALUE=[104,139)}	{SPO2VALUE=[97.3,98]}	0.222222222	1	3
{SPO2VALUE=[95.3,97.3)}	{HRVALUE=[154,171]}	0.222222222	0.666666667	2
{RRVALUE=[54.3,60]}	{SPO2VALUE=[95.3,97.3)}	0.222222222	0.666666667	2
{PULSEPVALUE=[154,171]}	{RRVALUE=[54.3,60]}	0.222222222	0.666666667	2
{RRVALUE=[54.3,60]}	{Cluster=[4,6)}	0.222222222	0.666666667	2
{HRVALUE=[139,154),SPO2VALUE=[95.3,97.3),PULSEPVALUE=[139,154),Cluster=[4,6)}	{RRVALUE=[54.3,60]}	0.111111111	1	3
{HRVALUE=[139,154),RRVALUE=[54.3,60],SPO2VALUE=[95.3,97.3),Cluster=[4,6)}	{PULSEPVALUE=[139,154)}	0.111111111	1	3
{RRVALUE=[54.3,60],SPO2VALUE=[95.3,97.3),PULSEPVALUE=[139,154),Cluster=[4,6)}	{HRVALUE=[139,154)}	0.111111111	1	3
{HRVALUE=[104,139),RRVALUE=[40.7,54.3),SPO2VALUE=[87,95.3),PULSEPVALUE=[104,139)}	{Cluster=[6,9]}	0.111111111	1	3
{HRVALUE=[104,139),RRVALUE=[40.7,54.3),SPO2VALUE=[87,95.3),Cluster=[6,9]}	{PULSEPVALUE=[104,139)}	0.111111111	1	3
{Cluster=[6,9]}	{PULSEPVALUE=[104,139)}	0.222222222	0.666666667	2
{HRVALUE=[139,154),PULSEPVALUE=[139,154)}	{RRVALUE=[40.7,54.3)}	0.222222222	0.666666667	2
{HRVALUE=[139,154),PULSEPVALUE=[139,154)}	{Cluster=[1,4)}	0.222222222	0.666666667	2
{HRVALUE=[104,139),PULSEPVALUE=[104,139)}	{SPO2VALUE=[97.3,98]}	0.222222222	0.666666667	2
{HRVALUE=[154,171],PULSEPVALUE=[154,171]}	{RRVALUE=[54.3,60]}	0.222222222	0.666666667	2

Table 6. 9: Rules sets with a Specified Support < 15% and Confidence Level 100%

LHS	RHS	SUPPORT	CONFIDENT	LIFT
{RRVALUE=[57.7,63],SPO2VALUE=[85,94)}	{HRVALUE=[158,176]}	0.125	1	2.666666667
{HRVALUE=[158,176],SPO2VALUE=[85,94)}	{RRVALUE=[57.7,63]}	0.125	1	2.666666667
{RRVALUE=[57.7,63],SPO2VALUE=[85,94)}	{PULSEPVALUE=[157,176]}	0.125	1	2.666666667
{SPO2VALUE=[85,94),PULSEPVALUE=[157,176]}	{RRVALUE=[57.7,63]}	0.125	1	2.666666667
{RRVALUE=[57.7,63],SPO2VALUE=[85,94)}	{Cluster=[1,3)}	0.125	1	2.666666667
{SPO2VALUE=[85,94),Cluster=[1,3)}	{RRVALUE=[57.7,63]}	0.125	1	2.666666667
{HRVALUE=[158,176],SPO2VALUE=[85,94)}	{PULSEPVALUE=[157,176]}	0.125	1	2.666666667
{SPO2VALUE=[85,94),PULSEPVALUE=[157,176]}	{HRVALUE=[158,176]}	0.125	1	2.666666667
{HRVALUE=[158,176],SPO2VALUE=[85,94)}	{Cluster=[1,3)}	0.125	1	2.666666667
{SPO2VALUE=[85,94),Cluster=[1,3)}	{HRVALUE=[158,176]}	0.125	1	2.666666667
{SPO2VALUE=[85,94),PULSEPVALUE=[157,176]}	{Cluster=[1,3)}	0.125	1	2.666666667
{SPO2VALUE=[85,94),Cluster=[1,3)}	{PULSEPVALUE=[157,176]}	0.125	1	2.666666667
{RRVALUE=[39.7,57.7),Cluster=[3,5.67)}	{HRVALUE=[145,158)}	0.125	1	4
{HRVALUE=[153,175],SPO2VALUE=[91,95)}	{Cluster=[5,7]}	0.142857	1	2.333333333
{SPO2VALUE=[91,95),PULSEPVALUE=[153,174]}	{HRVALUE=[153,175]}	0.142857	1	1.75
{HRVALUE=[153,175],SPO2VALUE=[91,95)}	{PULSEPVALUE=[153,174]}	0.142857	1	1.75
{RRVALUE=[37,42),Cluster=[1,3)}	{SPO2VALUE=[97,99]}	0.142857	1	1.75
{SPO2VALUE=[97,99],Cluster=[1,3)}	{RRVALUE=[37,42)}	0.142857	1	3.5
{RRVALUE=[37,42),Cluster=[1,3)}	{PULSEPVALUE=[153,174]}	0.142857	1	1.75
{RRVALUE=[37,42),PULSEPVALUE=[153,174]}	{Cluster=[1,3)}	0.142857	1	3.5
{RRVALUE=[33,38.3),PULSEPVALUE=[140,157)}	{SPO2VALUE=[96,99]}	0.125	1	1.6
{RRVALUE=[33,38.3),SPO2VALUE=[89,94.7)}	{HRVALUE=[157,173]}	0.125	1	2.666666667
{RRVALUE=[33,38.3),PULSEPVALUE=[157,173]}	{HRVALUE=[157,173]}	0.125	1	2.666666667
{RRVALUE=[33,38.3),SPO2VALUE=[89,94.7)}	{Cluster=[5.67,8]}	0.125	1	2.666666667

Results on Patient Risk Scores

Results of a patients categorised by the vital scores in 6 different periods is shown in Table 6.9. This shows the assigning a patient a score based on the clinical abstraction detailed in section 6.2.7. This data is then used to generate several other reports as shown in Figure 6.14, 6.15 and 6.16 which quantifies the percentage of time a patient stays in risky or none risky state based on associated vital scores.

Table 6. 10: Patient Risk Score.
A Sample representing one patient's calculated vital scores in 6 time intervals. The vital scores ranges from 0 to 2 with majority of vital score = 1.

PATIENTID	CLUSTER	HRMEANS	RRMEANS	SPO2MEANS	PULSEPMEANS	HRSTATUS	RRSTATUS	SPO2STATUS	VITALSCORE
1	1	175.2147507	38.66383421	97.03794455	174.9420776	NORMAL	NORMAL	ABNORMAL	1
1	2	153.9800086	49.51901407	95.58620506	153.8132803	NORMAL	NORMAL	ABNORMAL	1
1	3	134.2854856	52.45581483	95.0971569	134.1403892	NORMAL	NORMAL	ABNORMAL	1
1	5	153.9218798	63.93266209	91.75905022	153.6131991	NORMAL	ABNORMAL	NORMAL	1
1	7	158.7721863	56.5356854	97.49599043	158.5871394	NORMAL	NORMAL	ABNORMAL	1
1	1	142.4994752	42.19848762	98.45825337	141.8768183	NORMAL	NORMAL	ABNORMAL	1
1	2	151.9256865	54.3640605	94.24332684	150.6434428	NORMAL	NORMAL	ABNORMAL	1
1	4	144.4572422	55.66810481	96.00190663	144.1892143	NORMAL	NORMAL	ABNORMAL	1
1	5	161.7979885	60.80295637	87.88403143	161.0225316	NORMAL	ABNORMAL	NORMAL	1
1	2	160.1165932	46.1057092	98.87526967	159.8357446	NORMAL	NORMAL	ABNORMAL	1
1	3	143.0596426	53.74040198	95.31717099	142.6781697	NORMAL	NORMAL	ABNORMAL	1
1	4	148.1844026	46.92554234	91.21635242	147.1123793	NORMAL	NORMAL	NORMAL	0
1	5	166.9882772	82.62816101	94.71971702	166.6230139	NORMAL	ABNORMAL	ABNORMAL	2
1	3	161.4950448	56.47477379	96.11072669	161.1753915	NORMAL	NORMAL	ABNORMAL	1
1	4	142.1031721	38.29245059	98.88034447	141.4036877	NORMAL	NORMAL	ABNORMAL	1
1	6	158.1893789	52.03122212	90.50428043	157.8648821	NORMAL	NORMAL	NORMAL	0
1	8	148.3426042	47.98991148	91.03809913	147.7384934	NORMAL	NORMAL	NORMAL	0
1	1	151.3301136	53.84826457	96.17815295	151.0259445	NORMAL	NORMAL	ABNORMAL	1
1	3	153.0624522	51.20663945	87.53743466	152.410938	NORMAL	NORMAL	NORMAL	0
1	4	162.3415642	33.23974443	96.9800224	162.0734588	NORMAL	NORMAL	ABNORMAL	1
1	5	153.4982486	55.7153637	94.63164175	152.9012429	NORMAL	NORMAL	ABNORMAL	1
1	1	165.1578261	50.18108613	95.66985503	164.754446	NORMAL	NORMAL	ABNORMAL	1
1	3	152.3894432	53.04091963	95.35784378	152.1831551	NORMAL	NORMAL	ABNORMAL	1
1	4	136.5081652	36.03189173	95.16643303	136.0304823	NORMAL	NORMAL	ABNORMAL	1

Figure 6. 14 : Percentage of the Vital Score (Risk Scores) per Hour.

This shows the area view of which hours patients had risk scores 0 to 3. The fourth hour had patients who had risk scores 0 to 3 while the rest of the hours had risk scores 0 to 2.

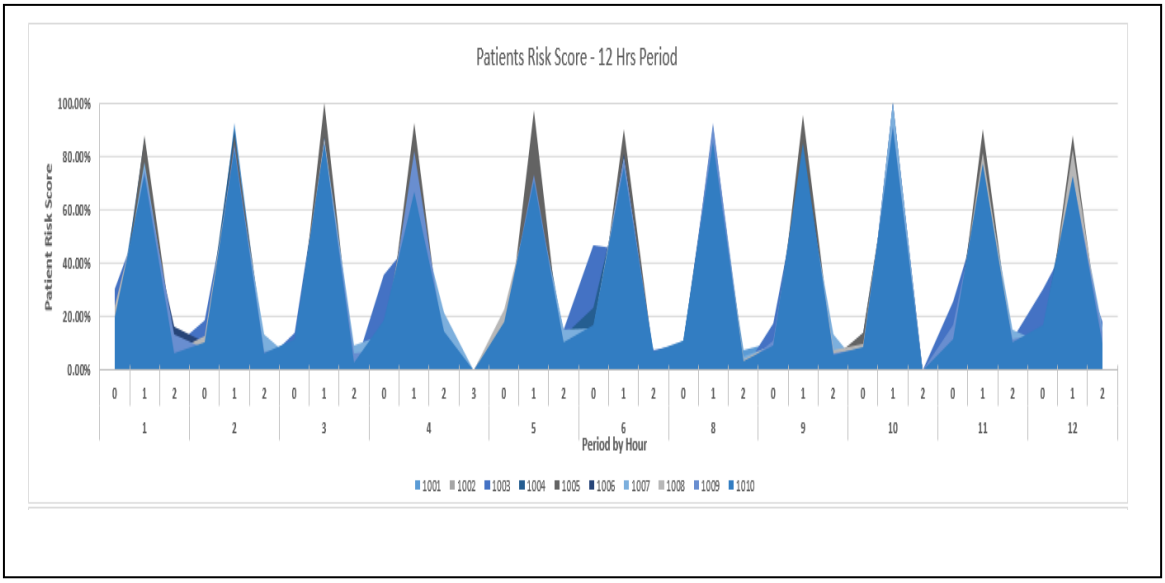


Figure 6. 15: Patient Risk Scores for 12 Hours on 10 Patients.

This figure identifies the patients with their risk scores. This is a complement of Figure thus clearly shows which patients had the risky scores at each hour.

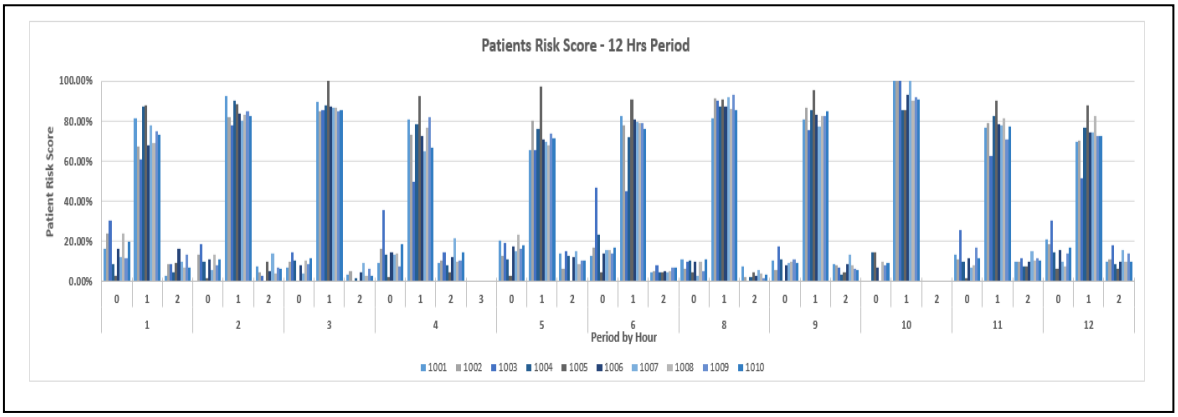
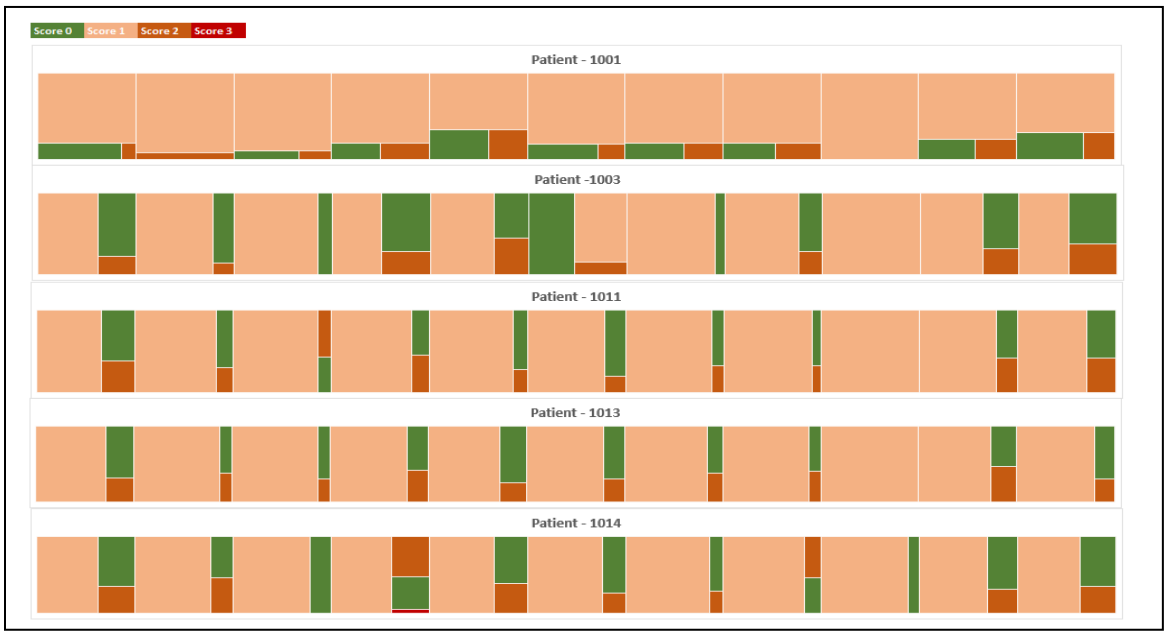


Figure 6. 16: Visual Map on 4 Patients Vital Scores

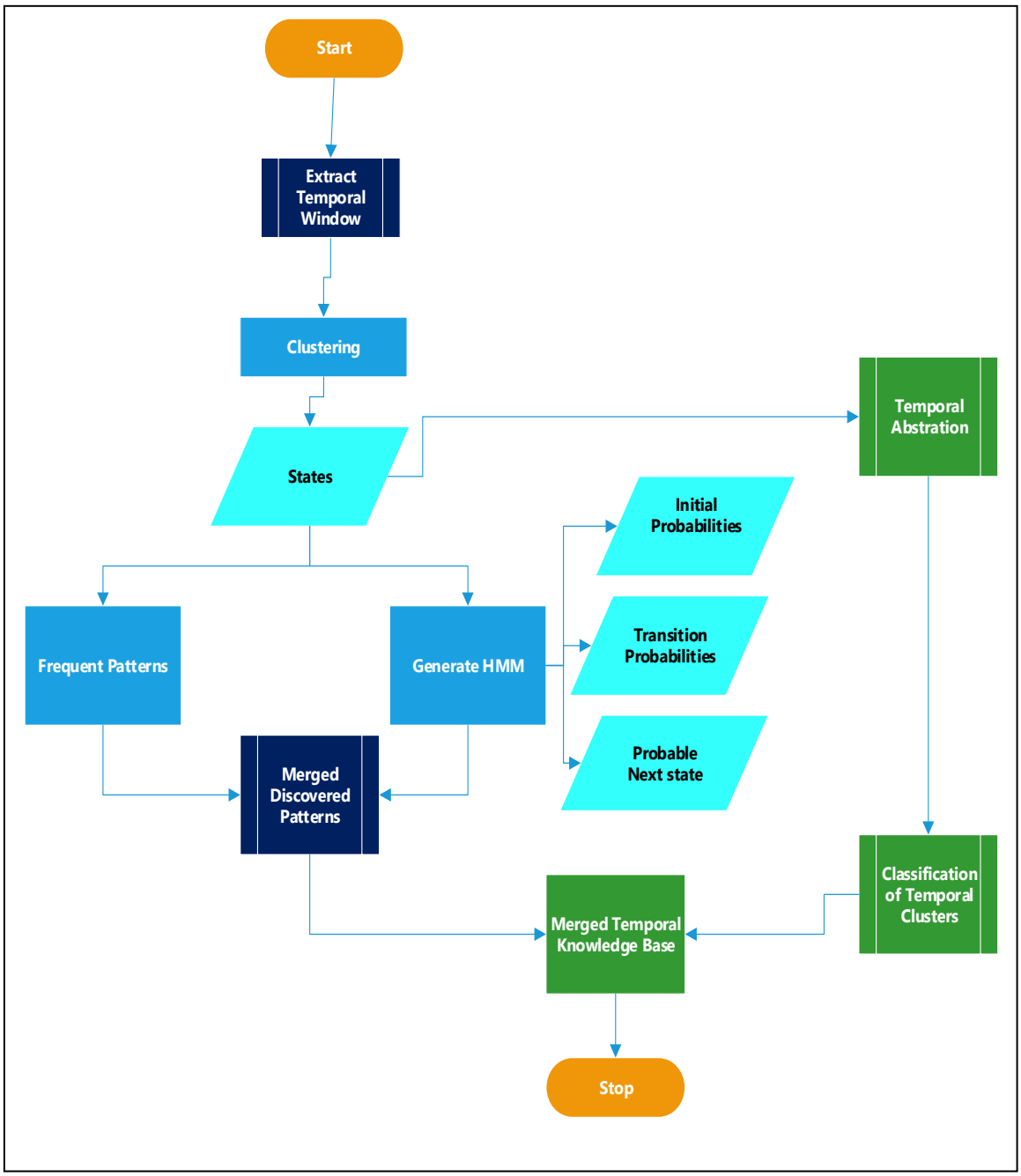
Displays shows analysis of 11 hours of each patient physiological data streams



6.2.9 Summary

Application of TPRMine is demonstrated in section 6.2 by processing patient physiological data streams thus facilitating the discovery of temporal patterns among different patient cohort. First, data driven clustering allows identification of the probable states a patients may transition to in a given time interval. Application of HMM allows quantification of the possible transitions from one state to the next and predictions of the next probable state. Next is the quantification of the states by using clinical abstraction which allows categorizing how risky a state is. With this risk, a vital score is generated for each patient. Additionally, it's possible to identify which are frequent temporal patterns in a given time window and the derived association rules. Over a period of time, this process is able to identify temporal patterns among different cohort of patients and this is knowledge that can be utilized by healthcare providers when making decisions about care of patients.

Figure 6. 17 : Process Flow in Application to Health Informatics



6.3 Application to Medicine

6.3.1 Case Study Application to Late Onset Neonatal Sepsis (LONS) in Neonatology

Neonatal sepsis is a bacterial infection in infants admitted to neonatal intensive care units which can cause severe morbidity or mortality. The infections may occur in the uterus, during birth or post natal. Very low birth weight (VLBW) infants especially those born premature have a higher factor of neonatal infection where incidents of infection are 3 to 10 times in preterm infants compared to full term infants (Griffin & Moorman , 2001). Late-onset neonatal sepsis (LONS) occurs after 3 days of life in 25% of very VLBW (1500 g) infants and leads to a more than doubling of mortality and a 50% increase in hospital stay (Griffin, et al., 2003).

Diagnosis of neonatal sepsis is usually difficult to establish. Griffin and Moorman (2011) hypothesis that abnormal characteristics of heart rate could precede clinical diagnosis of neonatal sepsis, if such information is captured early it could be utilized in addition to standard clinical parameters for care of patients. Analysis of heart rate variability (HRV) in preterm infants can help in tracking changes in subclinical signatures of a disease (Joshi, et al., 2019).

Premature infants are very susceptible to infectious pathogens and diagnosis of conditions such as neonatal sepsis is a challenging problem because the signs of the condition are often nonspecific. Late onset neonatal sepsis (LONS), which is the focus of this study, occurs in approximately 10% of neonates and 25% of very low birth weight infants hospitalized in NICUs. Slight changes in physiological data, which may not be easy to detect

through manual recordings at regular intervals, can be important in discovering the onset of sepsis in neonates as noted in (McGregor, Catley, & James, 2012).

An algorithm for detection of LONS by quantifying the variability of the HR and the RR was developed and detailed in (McGregor, Catley, & James, 2012). In prior work they found a significant increase in the probability of LONS when HR variability (HRV) was less that 38% and when respiration variability (RRV) greater than 38%. In that work, a sampling rate of 2 readings a minute was used due to limitations of data available and as a result patterns that could be derived from a higher sampling rate were not possible.

Application of the TPRMine algorithm seeks to understand if continuous analysis of temporal changes in patient heart rate, respiratory rate, SpO2 and PulpsePleth can lead to discovery of pathophysiological patterns in those physiological data streams for patients at risk of LONS.

To facilitate the process, this thesis enhances the threshold scoring approach in (McGregor, 2012) by integrating a more granular data representation to demonstrate that application of TPRMine algorithm to multiple physiological data streams from patients in NICU could enable discovery of temporal patterns within the hour of high or low threshold by;

- A combination of the frequent patterns generated for each period forms a classification system

- A scoring system is then derived by combining the HRV and RRV thresholds and the temporal patterns derived from the patient states.
- Identification of frequent patterns exhibited by patients based on the associated scoring thresholds.

Supplying patient physiological data streams as input to TPRMine algorithm generates data driven clusters including artefacts such as cluster means and cluster variances which can be utilized for retrospective analysis. Further more,

a) Data driven clusters form the states that a patient can transition to in a given period. These states are then fed to hidden markov models to quantify the possibly of transitioning from a state to another and the probability a patient will remain in the same state in the next time period.

b) The identified state transition form the temporal patterns for each patient

c) Now utilizing the HRV/RRV Threshold, a scoring mechanism is developed that highlight the potential of a patient physiological features becoming abnormal and potential association with conditions such as LONS

d) Identification of frequent features from patients who have different HRV RRV thresholds allows us to generate frequent patterns for different cohorts of patient forming a classification system

e) The classification system can then be used to formulate further hypotheses

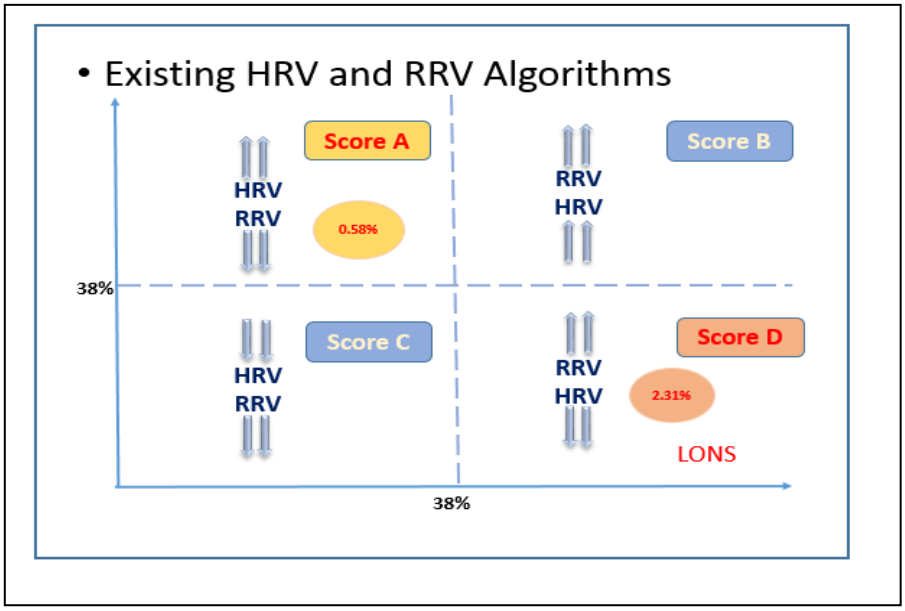
A pictorial adoption of TPRMine in algorithm to neonatology is shown in Figure 6.26.

6.3.1.1 An Overview of HRV and RRV Thresholds

The Existing HRV and RRV algorithms based on research work described in McGregor, Catley, & James (2012) calculates the variability of a patient's heart and respiratory rates and generates a percentage threshold as shown in Figure 6.18. A patient is then given a threshold score based on the percentage of HR and RR variability.

These algorithms have been able to detect potential LONS on those patients with a HRV < 38% and RRV > 38%. These are the patients who would fall in the score D region, see Figure 6.19.

Figure 6. 18 : Threshold Scores Based on HRV and RRV Algorithms



Additionally, utilizing the TPRMine algorithm allows the addition of more knowledge base about patients in NICU.

To support this process, implementation of the HRV and RRV algorithms is completed in order to gather the scoring threshold per patient. The algorithm has a score after every hour

for each patient, and with TPRMine, it is possible to gather more facts about a patient before the hour is completed. This is accomplished by the data driven clustering that allows identification of the many states a patient can transition to in a given hour. Of the patients identified to fall in each of the 4 score regions, we can now use TPR Mine algorithm to gather more knowledge of the underlying patient.

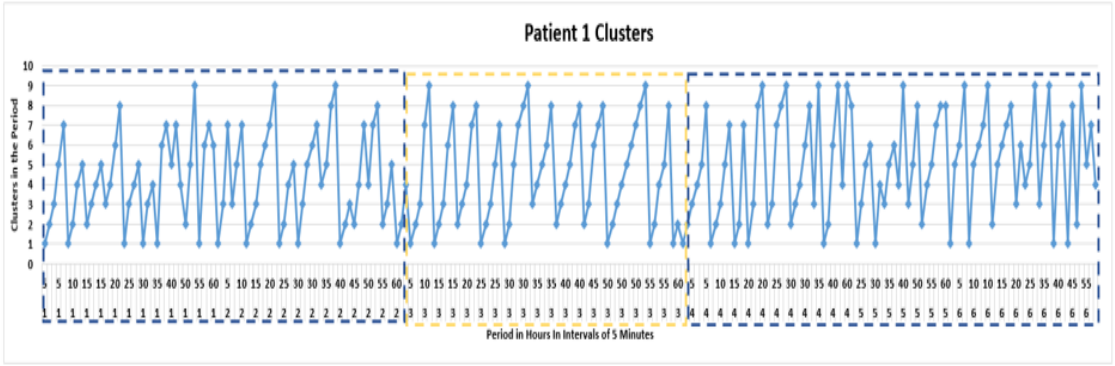
First is the identification of the states that patients transition to in a given period as shown in Figure 6.19, where 6hrs of patient data has been processed. Now incorporating the HRV and RRV threshold scores we can see the state transitions for a specific threshold score. Within the 6 hour period, Patient 1 in Figure 6.19 , (a) shows the threshold scores based on the HRV, RRV algorithms where the patient has been in two scoring regions a and B and in (b) shows the TPRMine results where in those 6 hours patient has transitioned between several states.

Figure 6. 19 : A Patient Threshold Scores Combined with TPRMine State Transitions.
In (b) the y-axis represented the states and x-axis represents the temporal windows and the hour. Yellow indicate period when patient is in SCORE_A while the rest of the time the patient is in SCORE_B

(a)

PERIOD HR	HRV_PERCENT	RRV_PERCENT	THRESHOLD
1	61	83	SCORE_B
2	54	53	SCORE_B
3	50	22	SCORE_A
4	73	64	SCORE_B
5	42	52	SCORE_B
6	57	43	SCORE_B

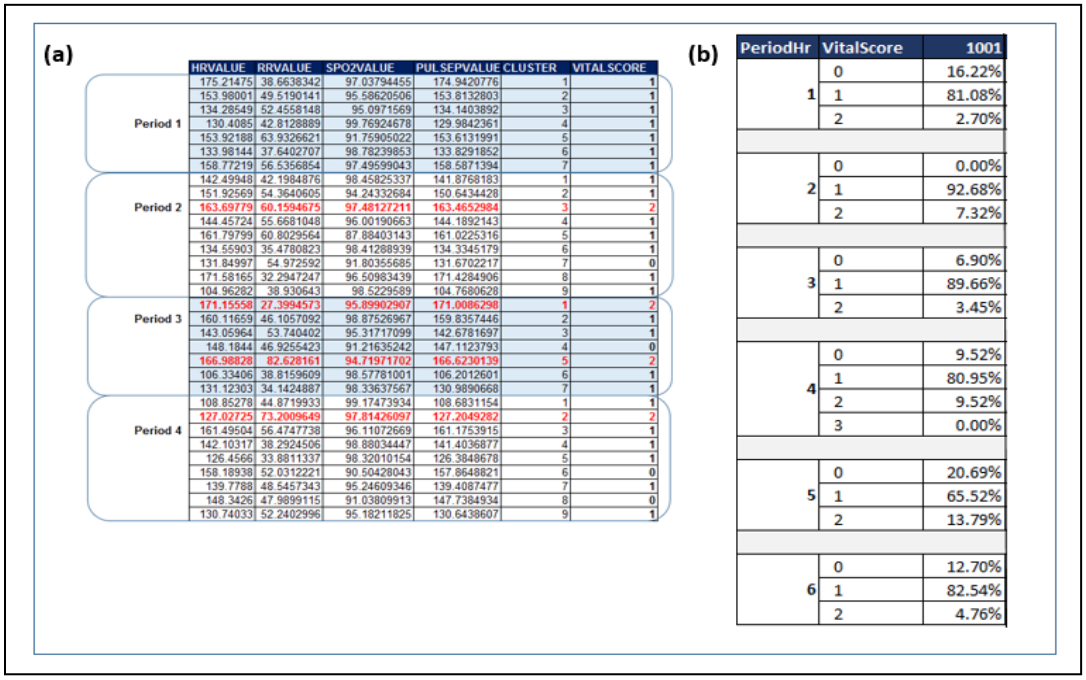
(b)



6.3.1.2 Patient Scoring with TPRMine Algorithm

Next, the patient is scored using the TPRMine algorithm as shown in Figure 6.20 where (a) identifies several facts about a specific period of time, the cluster, the cluster centroids and the vital score of the particular cluster. This is then used to determine of the 6 hrs, what is the percentage of times a patient stayed at a specific vital score, for example in the PeriodHr 2, the patient HR, RR or SpO_2 were abnormal the entire hour as the patient had 0% of vital score 0, 92.68% in vital score 1 and 7.32% in vital score 2.

Figure 6. 20 : Patient Scoring with Clusters Vital Scores



(a)

	HRVALUE	RRVALUE	SPO2VALUE	PULSEPVALUE	CLUSTER	VITALSCORE
Period 1	175.21475	38.6638342	97.03794455	174.9420776	1	1
	153.98001	49.5190141	95.58620506	153.8132803	2	1
	134.28549	52.4558148	95.0971569	134.1403892	3	1
	130.4085	42.8128889	99.76924678	129.9842361	4	1
	153.92188	63.9326621	91.75905022	153.6131991	5	1
	133.98144	37.6402707	98.78239853	133.8291852	6	1
	158.77219	56.5356854	97.49599043	158.5871394	7	1
Period 2	142.49948	42.1984876	98.45825337	141.8768183	1	1
	151.92569	54.3640605	94.24332684	150.6434428	2	1
	163.69779	60.1594675	97.48127211	163.4652984	3	2
	144.45724	55.6681048	96.00190663	144.1892143	4	1
	161.79799	60.8029564	87.88403143	161.0225316	5	1
	134.55903	35.4780823	98.41288939	134.3345179	6	1
	131.84997	54.972592	91.80355685	131.6702217	7	0
	171.58165	32.2947247	96.50983439	171.4284906	8	1
	104.96282	38.930643	98.5229589	104.7680628	9	1
Period 3	171.15558	27.3994573	95.89902907	171.0086298	1	2
	160.11659	46.1057092	98.87526967	159.8357446	2	1
	143.05964	53.740402	95.31717099	142.6781697	3	1
	148.1844	46.9255423	91.21635242	147.1123793	4	0
	166.98828	82.628161	94.71971702	166.6230139	5	2
	106.33406	38.8159609	98.57781001	106.2012601	6	1
	131.12303	34.1424887	98.33637567	130.9890668	7	1
Period 4	108.85278	44.8719933	99.17473934	108.6831154	1	1
	127.02725	73.2009649	97.81426097	127.2049282	2	2
	161.49504	56.4747738	96.11072669	161.1753915	3	1
	142.10317	38.2924506	98.88034447	141.4036877	4	1
	126.4566	33.8811337	98.32010154	126.3848678	5	1
	158.18938	52.0312221	90.50428043	157.8648821	6	0
	139.7788	48.5457343	95.24609346	139.4087477	7	1
	148.3426	47.9899115	91.03809913	147.7384934	8	0
	130.74033	52.2402996	95.18211825	130.6438607	9	1

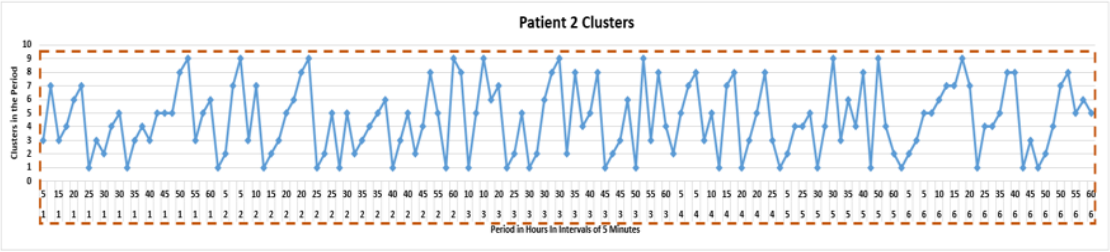
(b)

PeriodHr	VitalScore	1001
1	0	16.22%
	1	81.08%
	2	2.70%
2	0	0.00%
	1	92.68%
	2	7.32%
3	0	6.90%
	1	89.66%
	2	3.45%
4	0	9.52%
	1	80.95%
	2	9.52%
	3	0.00%
5	0	20.69%
	1	65.52%
	2	13.79%
6	0	12.70%
	1	82.54%
	2	4.76%

Figure 6.21 shows a different patient (Patient 2) whose HRV and RRV thresholds were at score D region which indicates a potential risk for LONS, McGregor et al. TPRMine identifies the many states the patient transitions to within the 6 hours.

Figure 6. 21 : State Transitions with TPRMine and Threshold Score from HRV /RRV Algorithms

PERIOD HR	HRV_PERCENT	RRV_PERCENT	THRESHOLD
1	2	69	SCORE_D
2	5	82	SCORE_D
3	10	85	SCORE_D
4	0	85	SCORE_D
5	0	74	SCORE_D
6	8	53	SCORE_D



6.3.1.3 Comparison of Two Patients Based of the HRV RRV Threshold Algorithm

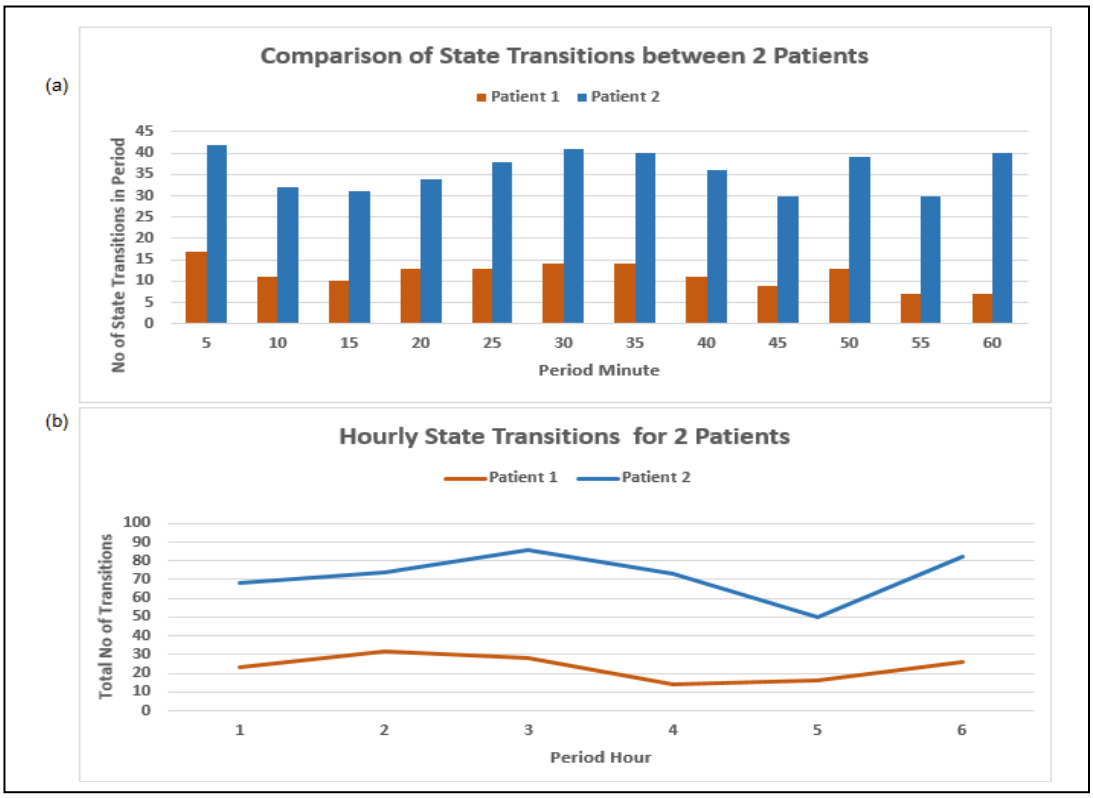
As there seems to be a clear difference between the 2 patients in Figures 6.20 and 6.21, as shown in the threshold scores produced by the HRV and RRV algorithms based on the underlying patient's data, next is to understand if the 2 patients exhibit differing temporal state transitions from the TPRMine algorithms. A comparison on thresholds between the 2 patients is presented in Figure 6.22 using data captured in 11 hours excluding the 7th hour due to data quality issues discovered during testing.

Figure 6. 22 : A Comparison on Patient 1 and 2 Based on States, HRV and RRV Algorithms

PATIENT_ID	PERIOD HR	HRV_PERCENT	RRV_PERCENT	THRESHOLD
1	1	2	69	SCORE_D
2	1	53	81	SCORE_B
1	2	5	82	SCORE_D
2	2	42	3	SCORE_A
1	3	10	85	SCORE_D
2	3	57	28	SCORE_A
1	4	0	85	SCORE_D
2	4	73	56	SCORE_B
1	5	0	74	SCORE_D
2	5	73	50	SCORE_B
1	6	8	53	SCORE_D
2	6	68	55	SCORE_B
1	8	17	76	SCORE_D
2	8	71	78	SCORE_B
1	9	22	83	SCORE_D
2	9	59	52	SCORE_B
1	10	23	88	SCORE_D
2	10	52	60	SCORE_B
1	11	25	83	SCORE_D
2	11	72	76	SCORE_B
1	12	7	83	SCORE_D
2	12	69	77	SCORE_B

Using TPR Mine, further comparison is performed of the state transitions between the two patients as shown in Figure 6.23 (a) shows in each period minute, the number of states the patient transitions to, patient 2 has higher number state transitions in each period hour compared to patient 1.Figure 6.23 (b) shows the pattern of transition in 6hr period between patient 1 and 2. The patient with the score D is seen as having lower number of state transitions compare to the patient with score A and B.

Figure 6. 23 : Number of State Transitions in a 6 Hour Period Between 2 Patients



6.3.1.4 Comparison of the Entire Patient Population and Formulation of Hypothesis Tests

After comparing two different patients, an analysis of the entire patient population is performed to understand the variation on the number of state transitions across all the cohorts. In an 11 hour period, there were 36 patients in the data analyzed. Of the 36, 3 were

found to have a threshold SCORE_D while the 32 had a threshold SCORE_B and 1 had a threshold SCORE_A. Two groups are then formed, those with threshold SCORE_D in group B while the rest are placed in group A. This enables understanding the overall state transitions between the two groups of patients. The results in Figure 6.24 shows the hourly total transitions on 11 hours of data from the two groups of patients.

Figure 6.24 (a) shows results for group A while b shows results for group B. Difference were noted in the number of state transitions between the two groups. To better understand this difference, statistical principles for group analysis is completed by calculating hourly average transitions and standard deviation between the two patient groups. The results presented in Figure 6.25 shows the hourly average in Group A is consistently higher that 50 compared to that of Group B in each of the 6 hours analyzed.

Figure 6. 24 : Total State Transitions by 36 Patients
(a) Group A shows 33 patients with HRV/RRV Threshold Score A or B, (b) Group B shows 3 patients with HRV/RRV Threshold Score D

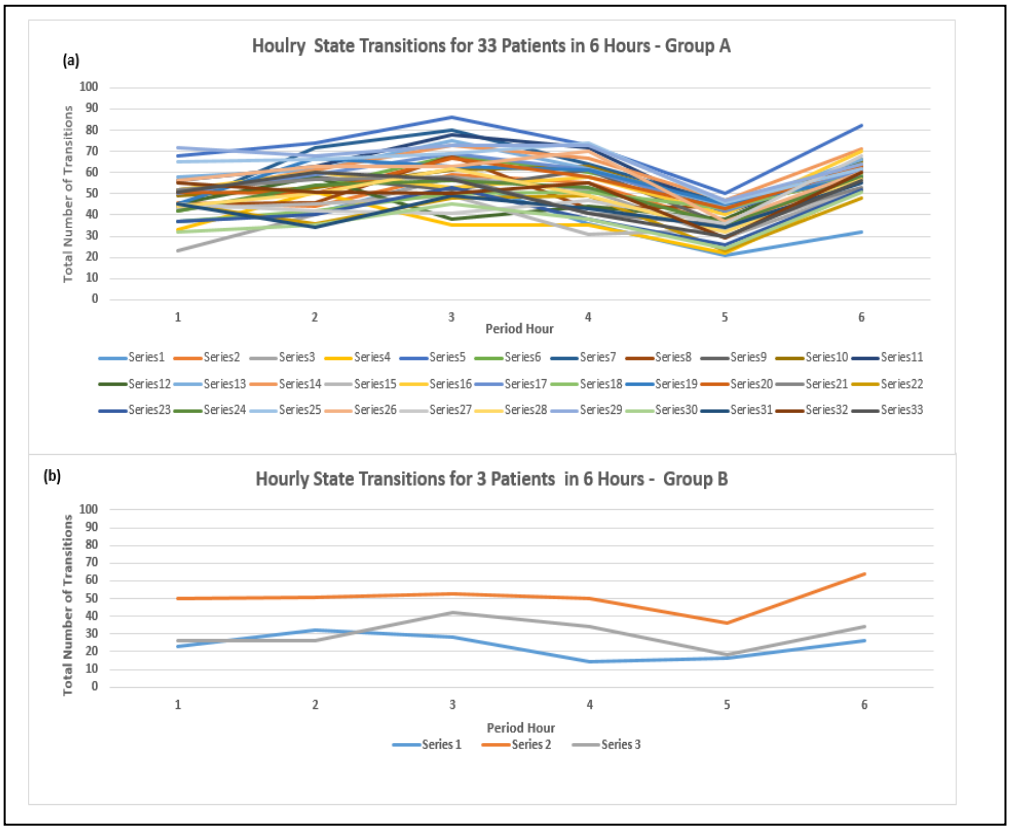
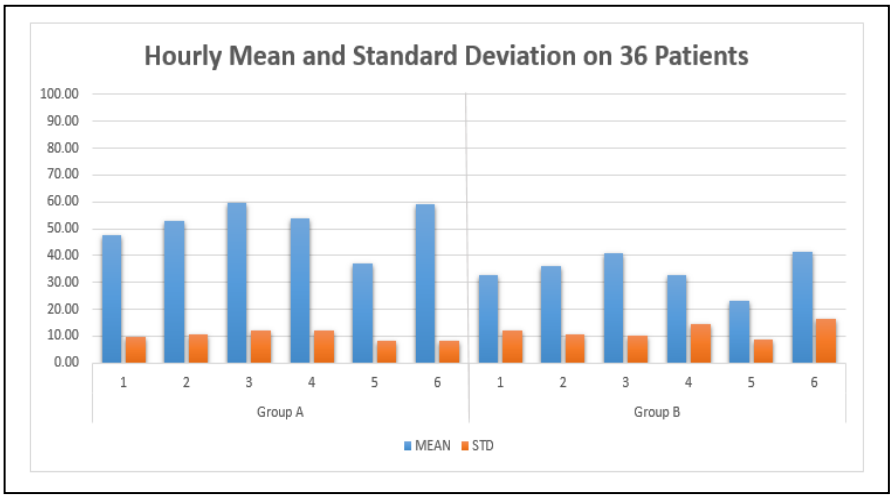


Figure 6. 25 : Comparison of Means and Standard Deviation of Hourly Transitions by 36 Patients



The results shown in Figure 6.25 indicates a difference in the average transition between the two groups of patients facilitating formulation of null hypothesis tests. One hypothesis test formulated in this thesis shows that the average number of state transitions in the group of patients with a threshold SCORE_D (Cohort B) is less that the transitions in patients with other threshold scores (Cohort A). This is the alternative hypothesis (H_1) while the null hypothesis (H_0) states that there is no difference in the two averages. If the average hourly state transitions in Cohort A is represented as MeanA, and average hourly state transitions in Cohort B is represented as MeanB, then the hypothesis test is written as follows;

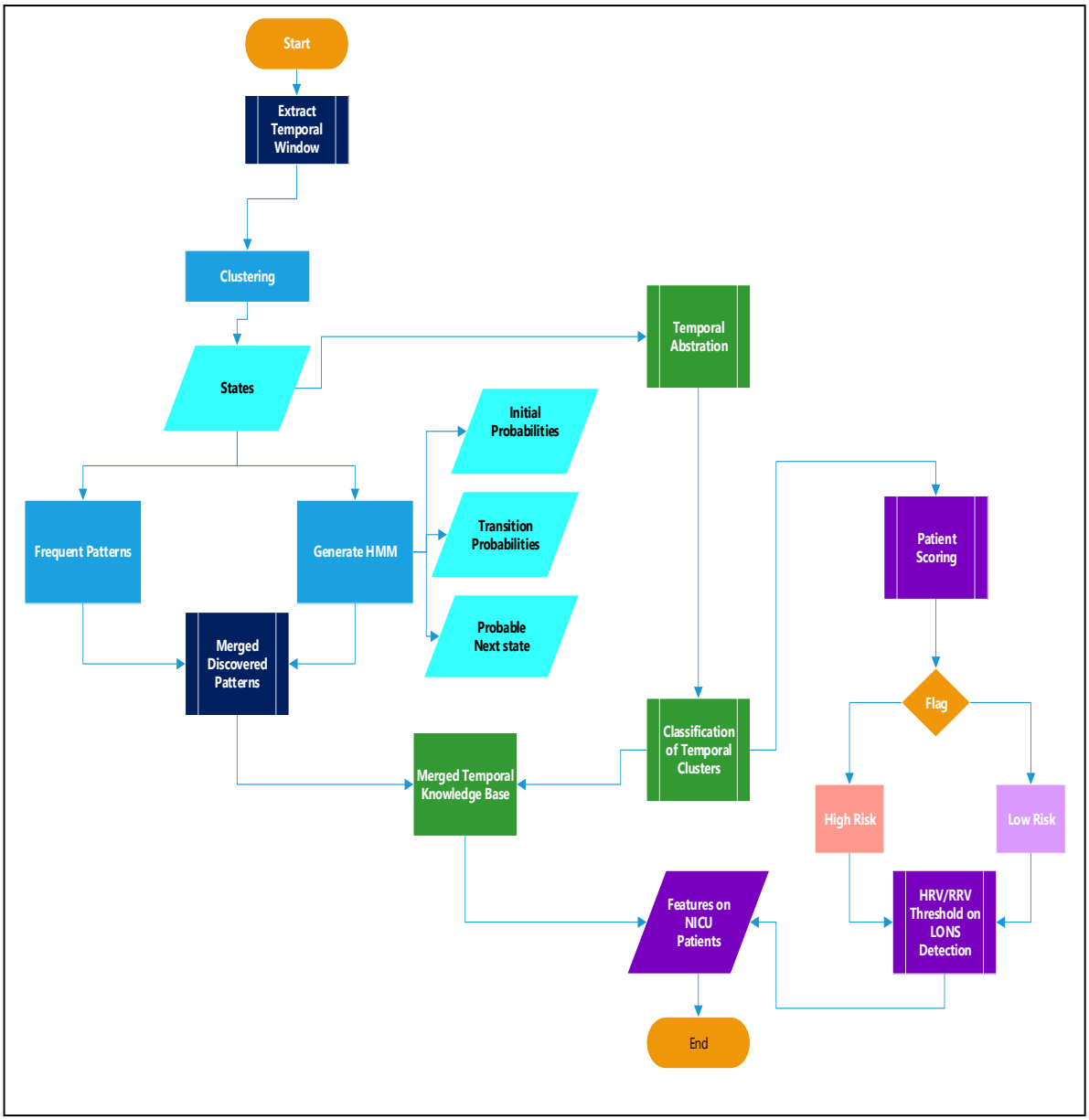
H_0: MeanA = MeanB

H_1: MeanA ≠ MeanB

A random sample of 22 patients from Cohort A is selected using data from hour 3 to test this hypothesis. As the hypothesis is interested in testing if the two means (MeanA, MeanB) are different, the t-test is performed (Mendenhall, Beaver, & Beaver, 1999). This generates the following results; *t*-statistic = 3.8951, df = 9.5016, p-value = 0.003288, alternative hypothesis is true which indicates the difference in means between the 2 cohorts is not equal to 0 with a 95 percent confidence interval, as such the null hypothesis is rejected.

The results of such hypothesis tests demonstrates the ability of the proposed TPRMine process to facilitate quantification of statistical difference in cohorts of patients and can be utilized in classification of those patients that may be at risk of neonatal infections such as LONS.

Figure 6. 26 : The process for Application of TPRMine Algorithm to Neonatology



6.3.2 Application of TPRMine Algorithm to Chronic Care of Elderly Patients

This section presents an application of the method proposed in this thesis to a second case study as follows.

The costs of lengthy hospital admissions and multiple emergency room visits (ER Visits) from patients with conditions such as heart failure (HF) and chronic obstructive pulmonary disease (COPD) can place a significant burden on healthcare systems (Ward et al, 2014). A report generated from National Ambulatory Care Reporting System (NACRS) in Canada shows that in 2014-15 the leading conditions for which patients were admitted from emergency departments (EDs) were Chronic Obstructive Pulmonary Disease (COPD), Heart Failure (HF) and Pneumonia (CIHI, 2015). In 2012, the direct costs associated with caring for patients with such conditions was estimated at over $50 billon in USA (Gotz, Wang, & Perer, 2014).

Additionally Gorst et al notes that escalation of HF and COPD is associated with an in increase in the aging population. It is expected that by 2036, 1 in 4 Canadians will be over 65 and about 85% will have chronic conditions (Ward, Schiller, & Goodman, 2014). Understanding the various factors contributing to lengthy hospitalization and multiple ER visits could aid in cost-effective management in the delivery of services leading to potential improvement on quality of life for patients.

To facilitate this case study utilizing data collected from patients participating in a remote patient monitoring program was used. This is a collaboration with Alaya Care, We Care, Southlake and University Of Ontario Institute Of Technology under OCE Grant (MIS#23823) and ethics approval was received from both Southlake Regional Health Centre (SRHC REB

#0087-1516) and the University Of Ontario Institute Of Technology (UOIT REB #14136) for the research study entitled “Reducing Hospital Admissions and Emergency Department Visits for Chronically Ill Patients using Remote Patient Monitoring and Telehealth Tools”.

Initial models developed in this research applied correlation analysis and stepwise Bayesian predictive modelling on static anonymized patient data provided from Alaya Care monitoring system. Preliminary results show that gender, past medical history and vital status are key factors to hospital admissions and ER Visits. Additionally, when a factor to indicate the period before, during and after an ER Visits was included, the resulting model shows a very high likelihood ratio and improved p values on all vital status. This work was detailed in (Inibhunu, Schauer, Redwood, Clifford & McGregor, 2017) and (Inibhunu & McGregor, 2018).

Current results shows that Bayesian models are effective in providing statistical significance and understanding hypothesis formulation on key contributors to lengthy hospitalization and multiple ER Visits (Inibhunu et al, 2017). However these are not enough to explain any hidden relationships in data. As such, more research is needed to fully understand the temporal patterns among variables before, during and after a hospital admission or ER visit. This research facilitates this process by investigating the use of temporal abstraction and pattern recognition models to fully quantify the temporal patterns exhibited in patient data before, during and after an adverse event.

In continuation with application of the proposed temporal pattern discovery framework to telehealth services, another research paper was described in (Inibhunu & McGregor , 2018)

This thesis premise is that understanding the various factors contributing to lengthy hospitalization and multiple ER visits could aid in cost-effective management in the delivery of services leading to potential improvement on quality of life for patients.

6.3.2.1 Case Study Application in Chronic Care of Elderly Patients

To facilitate this process, the TPRMine algorithm is applied to understand the temporal patterns among a select set of variables before, during and after a hospital admission or ER visit.

Research was described within publications related to this thesis in (Inibhunu et al, 2017, 2018) where data captured periodically about patient vital status (blood pressure, pulse rate, blood oxygen saturation and weight) was utilized to understand key contributors to lengthy hospital admissions and multiple emergency room visits. The results indicated physiological features such as average pulse rate, SpO_2 and patient's weight were some of the key contributors to adverse events. However, it was not clear at what point in time these features exhibited abnormal levels. The data contains dates when there was an ER Visit which facilities the calculation of days before the ER Visit and Days after the ER Visit.

With the TPRMine algorithm, the state transition models have processed the data on those patients who had an adverse event that resulted to a visit in ER using 3 steps. Step 1: Process Data for 1 Day before ER Visit occurred, Step 2: Process Data for the day when ER Visit occurred and Step 3: Process Data for 1 day after ER Visit occurred. In each of these steps TPRMine algorithm is applied resulting to several analytical datasets. Each process calculates the total number of states and the probability of transitioning from one state to the next. Table 6.10 shows 3 periods of time before, during and after a visit to ER and the associated

transition matrix while Table 6.11 shows the cluster means for each of the respective time periods.

Table 6. 11: Transition Matrices on 3 Periods Before, During and After an Adverse Event

Transition Matrix 1: 1 Day Before ER

	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	94.5%	0.7%	0.0%	0.3%	0.3%	0.7%	1.7%	1.7%	0.0%
s2	0.5%	93.9%	0.3%	0.5%	2.1%	1.1%	0.5%	0.8%	0.3%
s3	0.5%	0.5%	94.4%	0.0%	0.9%	0.9%	0.5%	0.9%	1.4%
s4	0.7%	0.7%	0.5%	95.8%	0.5%	0.2%	0.7%	0.2%	0.7%
s5	0.3%	2.0%	0.3%	0.7%	92.3%	0.7%	2.7%	0.7%	0.3%
s6	0.3%	1.2%	0.3%	1.2%	0.3%	94.1%	1.2%	0.6%	0.6%
s7	0.9%	0.2%	0.7%	0.9%	1.1%	0.9%	93.0%	0.9%	1.4%
s8	1.1%	1.4%	0.5%	0.0%	0.3%	0.5%	1.1%	94.3%	0.8%
s9	0.0%	0.5%	0.5%	1.1%	0.8%	0.5%	1.1%	0.5%	94.8%

Transition Matrix 2: Day of ER

	s1	s2	s3	s4	s5	s6	s7	s8	s9
s1	94.9%	0.0%	0.3%	0.9%	1.6%	1.3%	0.3%	0.6%	0.0%
s2	1.1%	93.6%	0.4%	0.4%	0.0%	1.4%	1.1%	1.1%	1.1%
s3	0.2%	0.6%	95.1%	0.4%	0.9%	0.4%	0.2%	0.9%	1.3%
s4	0.7%	0.3%	1.0%	92.2%	2.0%	1.0%	1.4%	0.7%	0.7%
s5	0.5%	0.3%	2.5%	1.4%	94.2%	0.5%	0.3%	0.0%	0.3%
s6	0.3%	1.6%	0.6%	0.3%	0.0%	93.9%	1.3%	0.3%	1.6%
s7	0.7%	0.0%	0.4%	2.6%	0.7%	0.4%	93.7%	0.0%	1.5%
s8	0.5%	1.0%	1.0%	0.5%	0.3%	0.3%	0.5%	95.9%	0.0%
s9	0.7%	0.9%	0.7%	0.5%	0.7%	0.2%	0.2%	0.9%	95.0%

Transition Matrix 3: Day of ER: 1 Day After ER

	s1	s2	s3	s4	s5	s6	s7	s8
s1	93.5%	1.3%	1.7%	0.9%	0.4%	0.4%	0.9%	0.9%
s2	0.4%	96.4%	0.0%	0.0%	0.7%	0.7%	0.4%	1.4%
s3	1.5%	0.7%	96.0%	0.4%	0.7%	0.4%	0.4%	0.0%
s4	2.2%	1.1%	0.0%	91.4%	1.1%	4.3%	0.0%	0.0%
s5	2.0%	0.0%	1.3%	1.3%	91.9%	1.3%	1.3%	0.7%
s6	0.4%	0.4%	1.2%	0.4%	1.2%	94.5%	0.8%	1.2%
s7	0.7%	0.7%	0.7%	0.7%	0.7%	2.0%	94.6%	0.0%
s8	2.4%	1.2%	0.6%	0.6%	0.6%	0.6%	0.0%	94.0%

Table 6. 12: Cluster Means for 3 Periods

Each row represents the cluster means in each state for the variables identified in the first column.

Cluster Means 1: 1 Day Before ER

	s1	s2	s3	s4	s5	s6	s7	s8	s9
ave_weight	87.4344	62.4564	70.637	103.917	93.8623	68.9085	61.7465	36.8663	91.2987
ave_pulse	85.51	92.7001	68.4551	70.2015	80.9251	70.174	85.9908	88.0365	69.0001
ave_spo2	92.4195	95.5889	93.2124	91.0428	93.6488	95.7322	92.6914	88.2651	94.9999
ave_bpmin	79.6296	76.3638	67.6896	75.6716	73.0444	66.139	72.0283	66.3006	138.997
ave_bpmax	130.558	128.985	126.873	139.664	135.754	107.75	135.111	108.419	70.0013

Cluster Means 2: Day of ER

	s1	s2	s3	s4	s5	s6	s7	s8	s9
ave_weight	61.6821	98.1506	41.3843	76.672	62.7554	93.8425	60.5143	81.5521	68.3202
ave_pulse	96.0065	67.8959	85.9093	76.2542	80.7403	80.4798	86.7387	86.8592	79.6403
ave_spo2	95.565	91.1241	89.2445	91.8775	94.7998	93.6293	92.8774	90.8169	92.5007
ave_bpmin	63.1167	81.5134	72.6984	77.7116	73.7128	72.997	67.6891	81.0877	73.3683
ave_bpmax	133.289	149.528	102.49	134.404	123.473	135.442	120.767	131.105	126.206

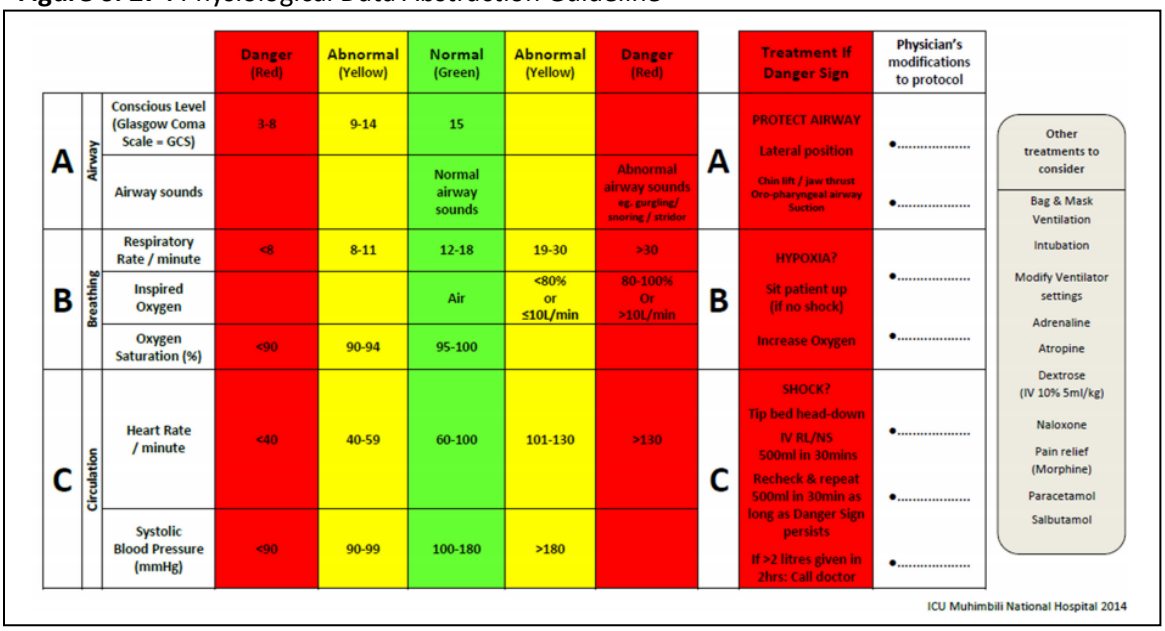
Cluster Means 3: 1 Day Post ER

	s1	s2	s3	s4	s5	s6	s7	s8
ave_weight	48.1715	64.8232	91.2758	93.6392	59.3537	99.5019	69.2012	65.2493
ave_pulse	91.0383	68.0116	69	78.3895	82.6387	78.562	92.4015	84.4968
ave_spo2	92.8198	90.9741	95.0013	93.502	91.8174	93.1963	94.0373	93.3892
ave_bpmin	64.8779	74.8744	138.946	75.4529	108.168	72.538	79.933	68.8962
ave_bpmax	120.138	133.325	70.0114	134.718	151.263	139.285	122.887	130.468

6.3.2.2 Clinical Temporal Abstraction

Similarly to the NICU case study, abstraction of physiological data is performed to quantify if a state is normal or abnormal. This is completed by adopting the temporal abstraction criteria detailed in (Baker, et al., 2012) and shown in Figure 6.27 where a physiological feature such as SpO_2 is categorized into 3 groups, normal, abnormal and danger. Baker et al. (2015) recommended this abstraction for use on data captured from patients age 16 and over.

Figure 6. 27 : Physiological Data Abstraction Guideline



			Danger (Red)	Abnormal (Yellow)	Normal (Green)	Abnormal (Yellow)	Danger (Red)		Treatment If Danger Sign	Physician's modifications to protocol
A	Airway	Conscious Level (Glasgow Coma Scale = GCS)	3-8	9-14	15			A	PROTECT AIRWAY Lateral position Chin lift / jaw thrust Oro-pharyngeal airway Suction	•..... •.....
		Airway sounds			Normal airway sounds		Abnormal airway sounds eg. gurgling/ snoring / stridor			
B	Breathing	Respiratory Rate / minute	<8	8-11	12-18	19-30	>30	B	HYPOXIA? Sit patient up (if no shock) Increase Oxygen	•..... •.....
		Inspired Oxygen			Air	<80% or ≤10L/min	80-100% Or >10L/min			
		Oxygen Saturation (%)	<90	90-94	95-100					
C	Circulation	Heart Rate / minute	<40	40-59	60-100	101-130	>130	C	SHOCK? Tip bed head-down IV RL/NS 500ml in 30mins Recheck & repeat 500ml in 30min as long as Danger Sign persists If >2 litres given in 2hrs: Call doctor	•..... •..... •.....
		Systolic Blood Pressure (mmHg)	<90	90-99	100-180	>180				

Other treatments to consider

Bag & Mask Ventilation
Intubation
Modify Ventilator settings
Adrenaline
Atropine
Dextrose (IV 10% 5ml/kg)
Naloxone
Pain relief (Morphine)
Paracetamol
Salbutamol

ICU Muhimbili National Hospital 2014

After utilizing the categories in Figure 6.27, the results of the abstraction are comparable with the case study completed on NICU patients. However, unlike in NICU data, where 3 physiological data elements were abstracted, on the elderly patient data, only the blood oxygen saturation was used for classification of a state. As TPRMine algorithm allows flexibility depending on data sources and data domain and as such, using just 1 variable for classification was possible as the elderly patient data HR and RR datasets were not available.

6.3.2.3 Quantification of Temporal Clusters

After the abstraction, quantification of the states identified for each of the periods, before, during and after an ER Visit was performed. The results are shown in Table 6.12 where TPRMine identifies that there are 9 states that a patient can transition to during the period before an ER visit.

Using the temporal abstraction applied to the SpO_2 data, those states that are normal, abnormal, danger are easily identified. In the period 1 day before an ER Visit there are 3 normal states (green), 5 abnormal states (yellow) and 1 danger state (red).

Table 6. 13 : Quantification of Temporal Clusters with Temporal Abstraction

Cluster Means 1: 1 Day Before ER

	s1	s2	s3	s4	s5	s6	s7	s8	s9
ave_weight	87.4344	62.4564	70.637	103.917	93.8623	68.9085	61.7465	36.8663	91.2987
ave_pulse	85.51	92.7001	68.4551	70.2015	80.9251	70.174	85.9908	88.0365	69.0001
ave_spo2	92.4195	95.5889	93.2124	91.0428	93.6488	95.7322	92.6914	88.2651	94.9999
ave_bpmin	79.6296	76.3638	67.6896	75.6716	73.0444	66.139	72.0283	66.3006	138.997
ave_bpmax	130.558	128.985	126.873	139.664	135.754	107.75	135.111	108.419	70.0013

Cluster Means 2: Day of ER

	s1	s2	s3	s4	s5	s6	s7	s8	s9
ave_weight	61.6821	98.1506	41.3843	76.672	62.7554	93.8425	60.5143	81.5521	68.3202
ave_pulse	96.0065	67.8959	85.9093	76.2542	80.7403	80.4798	86.7387	86.8592	79.6403
ave_spo2	95.565	91.1241	89.2445	91.8775	94.7998	93.6293	92.8774	90.8169	92.5007
ave_bpmin	63.1167	81.5134	72.6984	77.7116	73.7128	72.997	67.6891	81.0877	73.3683
ave_bpmax	133.289	149.528	102.49	134.404	123.473	135.442	120.767	131.105	126.206

Cluster Means 3: 1 Day Post ER

	s1	s2	s3	s4	s5	s6	s7	s8
ave_weight	48.1715	64.8232	91.2758	93.6392	59.3537	99.5019	69.2012	65.2493
ave_pulse	91.0383	68.0116	69	78.3895	82.6387	78.562	92.4015	84.4968
ave_spo2	92.8198	90.9741	95.0013	93.502	91.8174	93.1963	94.0373	93.3892
ave_bpmin	64.8779	74.8744	138.946	75.4529	108.168	72.538	79.933	68.8962
ave_bpmax	120.138	133.325	70.0114	134.718	151.263	139.285	122.887	130.468

To demonstrate the results of TPRMine algorithm on the actual patients within the three periods, 3 patients are randomly selected.

Figure 6.28 shows the temporal transition of the three patients in the day before the ER visit, Figure 6.29 are the same patients during the day when ER Visit happened and the Figure 6.30 shows the same 3 patients the day after the occurrence of an adverse event.

Figure 6. 28 : State Transitions by 3 Patients on Day Before ER Visit.
The Y-Axis represents the states in the period and the X-Axis represents the unit time of each transition. The red line indicates the state this is characterized as a danger state.

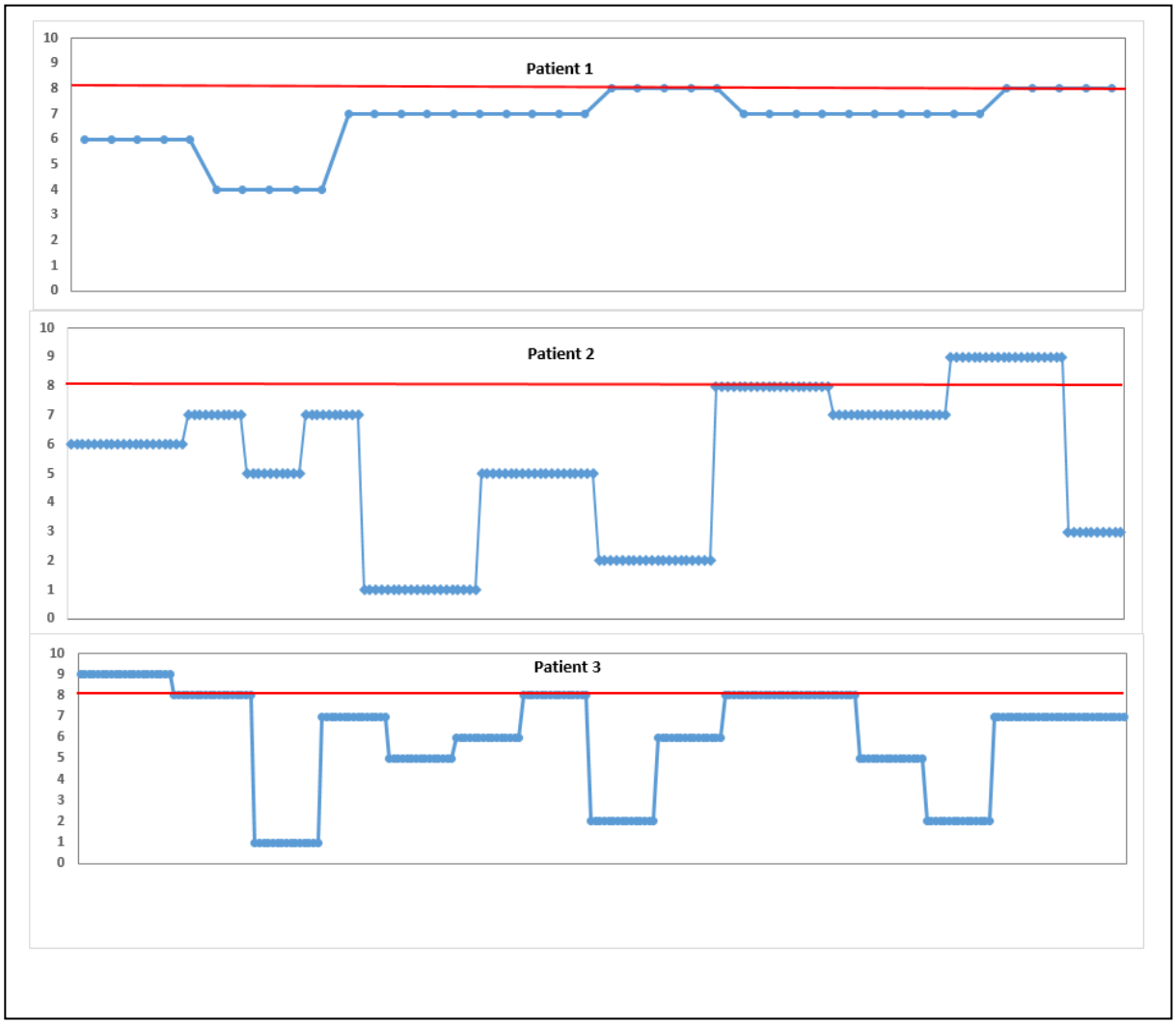


Figure 6. 29 : State Transitions by 3 Patients on Day of ER Visit.
The Y-Axis represents the states in the period and the X-Axis represents the unit time of each transition. The red line indicates the state this is characterized as a danger state.

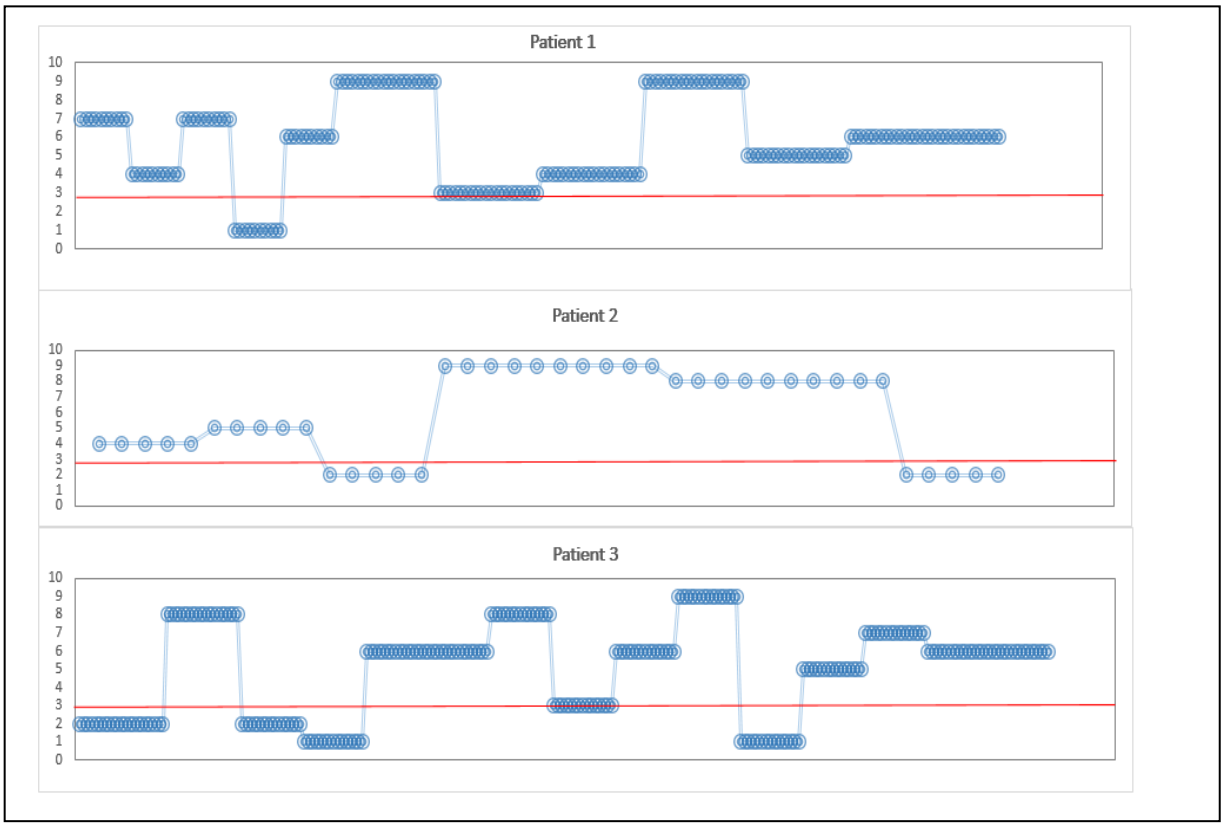
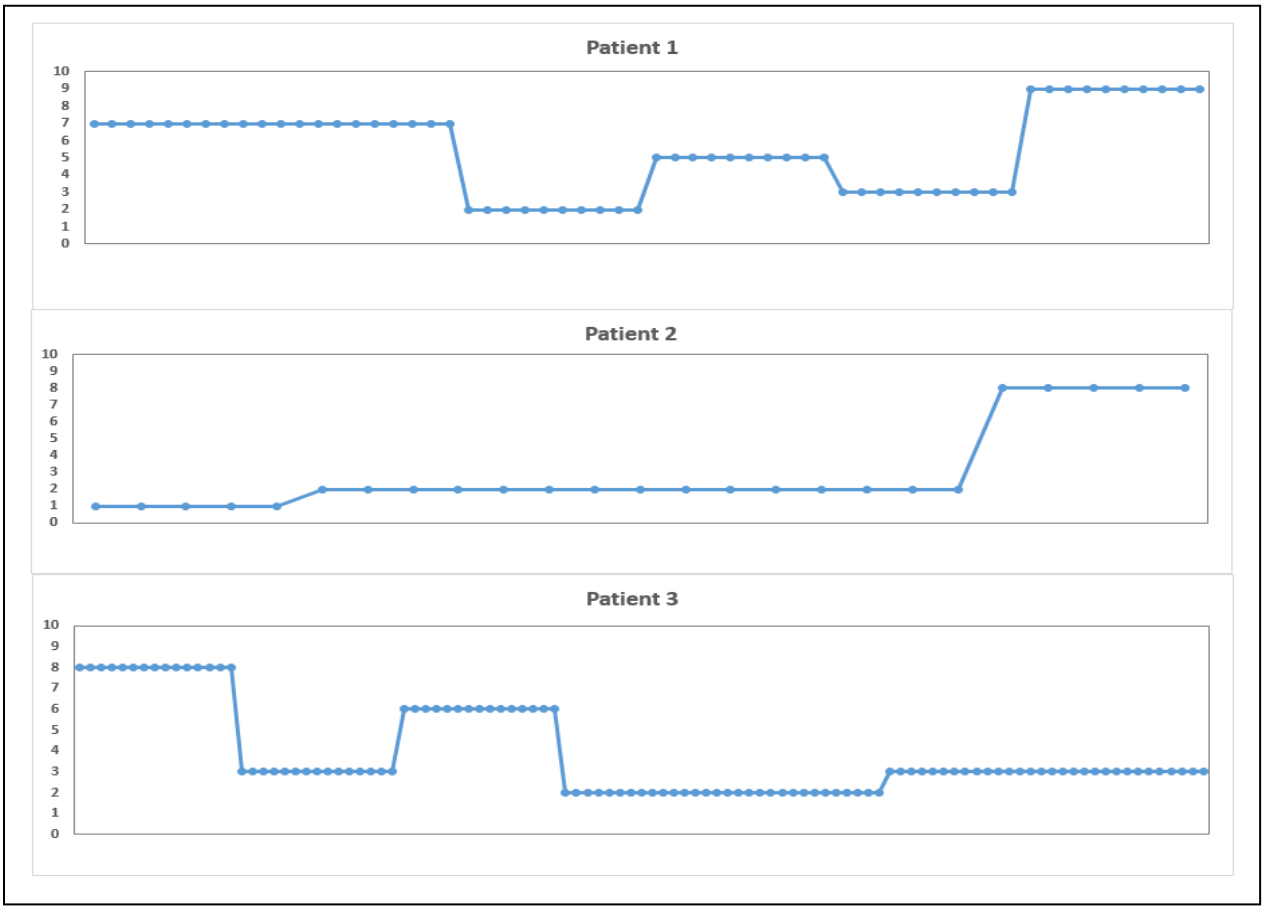


Figure 6. 30 : State Transitions by 3 Patients on Period Post ER Visit.
The Y-Axis represents the states in the period and the X-Axis represents the unit time of each transition. The patients have do not transition to a danger state in the Period Post ET Visit.



Frequent temporal patterns on each of the 3 time periods were generated as shown in Table 6.13. Note that it's possible to filter which frequent patterns to view based on confidence or support levels.

Table 6. 14 : Frequent Patterns Generated in 3 Different Periods

Period	LHS	RHS	Support	Confidence
PreER1Day	{ave_bpmax=[70,120)}	{ave_bpmin=[66,70.3)}	0.222222222	0.666666667
PreER1Day	{ave_bpmax=[70,120)}	{ave_spo2=[93.3,95]}	0.222222222	0.666666667
PreER1Day	{ave_bpmax=[70,120)}	{Cluster=[6.33,9]}	0.222222222	0.666666667
PreER1Day	{ave_bpmax=[132,140]}	{Cluster=[3.67,6.33)}	0.222222222	0.666666667
PreER1Day	{Cluster=[3.67,6.33)}	{ave_bpmin=[70.3,75.3)}	0.222222222	0.666666667
PreER1Day	{ave_bpmin=[70.3,75.3)}	{Cluster=[3.67,6.33)}	0.222222222	0.666666667
PreER1Day	{ave_spo2=[88,92),ave_bpmin=[66,70.3)}	{ave_bpmax=[70,120)}	0.111111111	1
PreER1Day	{ave_spo2=[88,92),ave_bpmax=[70,120)}	{ave_weight=[36,66)}	0.111111111	1
PreER1Day	{ave_weight=[36,66),ave_bpmax=[70,120)}	{ave_spo2=[88,92)}	0.111111111	1
PreER1Day	{ave_spo2=[88,92),ave_bpmax=[70,120)}	{Cluster=[6.33,9]}	0.111111111	1
ER Day	{ave_pulse=[79.7,85.3)}	{ave_bpmin=[72,74.3)}	0.333333333	1
ER Day	{ave_bpmin=[63,72),ave_bpmax=[102,125)}	{ave_weight=[41,61.7)}	0.111111111	1
ER Day	{ave_bpmin=[63,72),ave_bpmax=[125,133)}	{ave_weight=[41,61.7)}	0.111111111	1
ER Day	{ave_weight=[41,61.7),ave_bpmax=[125,133)}	{ave_bpmin=[63,72)}	0.111111111	1
ER Day	{ave_bpmin=[63,72),ave_bpmax=[125,134)}	{ave_weight=[41,61.7)}	0.111111111	1
ER Day	{ave_weight=[41,61.7),ave_bpmax=[125,134)}	{ave_bpmin=[63,72)}	0.111111111	1
ER Day	{ave_spo2=[92.3,95],ave_bpmin=[63,72)}	{ave_weight=[41,61.7)}	0.111111111	1
ER Day	{ave_weight=[41,61.7),ave_spo2=[92.3,95]}	{ave_bpmin=[63,72)}	0.111111111	1
ER Day	{ave_weight=[41,61.7),ave_bpmin=[63,72)}	{ave_pulse=[85.3,96]}	0.222222222	1
ER Day	{ave_weight=[41,61.7),ave_pulse=[85.3,96]}	{ave_bpmin=[63,72)}	0.222222222	1
Post ER 1Day	{ave_spo2=[92.7,93.3)}	{ave_bpmin=[72.7,73.3)}	0.222222222	0.666666667
Post ER 1Day	{ave_pulse=[85,90],ave_spo2=[93.3,94]}	{Cluster=[1,3.67)}	0.111111111	1
Post ER 1Day	{ave_spo2=[93.3,94],ave_bpmin=[62,72.7)}	{ave_bpmax=[70,125)}	0.111111111	1
Post ER 1Day	{ave_weight=[90,93],ave_bpmax=[70,125)}	{ave_spo2=[93.3,94]}	0.111111111	1
Post ER 1Day	{ave_bpmin=[62,72.7),Cluster=[1,3.67)}	{ave_bpmax=[70,124)}	0.111111111	1
Post ER 1Day	{ave_spo2=[89,92.7),ave_bpmax=[70,124)}	{Cluster=[1,3.67)}	0.111111111	1
Post ER 1Day	{ave_pulse=[85,90],ave_bpmax=[70,124)}	{Cluster=[1,3.67)}	0.111111111	1
Post ER 1Day	{ave_weight=[65.3,90),ave_spo2=[92.7,93.3),ave_bpmin=[72.7,73.3)}	{ave_bpmax=[132,148]}	0.111111111	1
Post ER 1Day	{ave_weight=[65.3,90),ave_spo2=[92.7,93.3),ave_bpmin=[72.7,73.3)}	{ave_pulse=[85,90]}	0.111111111	1
Post ER 1Day	{ave_weight=[65.3,90),ave_pulse=[85,90],ave_spo2=[92.7,93.3)}	{ave_bpmin=[72.7,73.3)}	0.111111111	1

.

6.4 Conclusions

This chapter has presented details on contributions to health informatics and medicine completed within this thesis. The health informatics contributions involved extensions to the multi-dimensional temporal data mining framework presented in (McGregor, 2011) and then integrating these enhancements to Artemis platform in (McGregor et al, 2013) to enable discovery of temporal relationships in abstractions and generating temporal patterns from time oriented physiological data. This process has been facilitated by utilizing the innovative TPR process and associated TPRMine algorithm developed within the computer science contribution of this thesis and presented in Chapter 5 for application to physiological data

streams. The results from this application demonstrate the potential for knowledge discovery through detection of temporal relationships and the subsequent temporal patterns among physiological data streams. This is achieved through detection of the many states patients may transition to at any given time. Integration of clinical context is demonstrated as effective at quantifying risky and non-risky states resulting to patient classification with vital scores.

In medicine, the proposed method has been applied to two clinical case studies in (a) NICU and in (b) RPM by using patient physiological data for classification of temporal events and episodes augmented with clinical context. Within NICU, the algorithm has been applied to detect the number of states a patient in an NICU bed can transition to in a given period of time. Additionally, a quantification of these states is completed leading to creation of a vital scoring. With this, it's possible to understand the percent of time a patient remains in a high or low vital score.

In the second clinical application, data utilized is captured from elderly patients that received care using RPM program. TPRMine algorithm is demonstrated as effective at identifying the states these patients transition to before, during and after an adverse event such as hospital admissions or ER visits. Quantification of the states that are normal, abnormal or in danger based on temporal abstraction guidelines is also facilitated by the TPRMine algorithm. Additionally, in both clinical case study applications, TPRMine enables discovery of those patterns that are deemed as frequent in each period of time thereby forming a temporal knowledge base. Experimental setup and evaluations of the proposed methods are discussed in chapter 7.

Chapter 7

7 Implementation of TPR Process and TPRMine Algorithm

7.1 Experimental Setup

This chapter presents the experimental design adopted to support the proposed methods. A pictorial view of this design is shown in Figure 7.1. This process is facilitated by a cloud computing architecture with 4 core processes. The computing of the algorithms is performed in R programming language which made it easier to adopt many existing packages such as the Mclust clustering for unsupervised learning, the Depmixs4 algorithm for HMM models and finally the Rules package for Association rules mining. DB2 database and Data Studio are also utilised for temporal data management.

Two prototypes are developed for processing real-time and static data.

(a) Real-Time streaming using IBM InfoSphere Streams, R, DB2, Data Studio and SQL

(b) Static data streaming using R, DB2, Data Studio and SQL

Each of these two prototypes includes modules that process each of the components shown in Figure 7.1. In order to handle real-time data streams, IBM InfoSphere streams was used to process incoming data streams within the Artemis Framework. Due to time constraint, results in this thesis uses the static prototype implemented with R following the process in Figure 7.2.

Figure 7. 1 : Experimental Process Flow

Modules for processing incoming data streams generating a temporal knowledge base

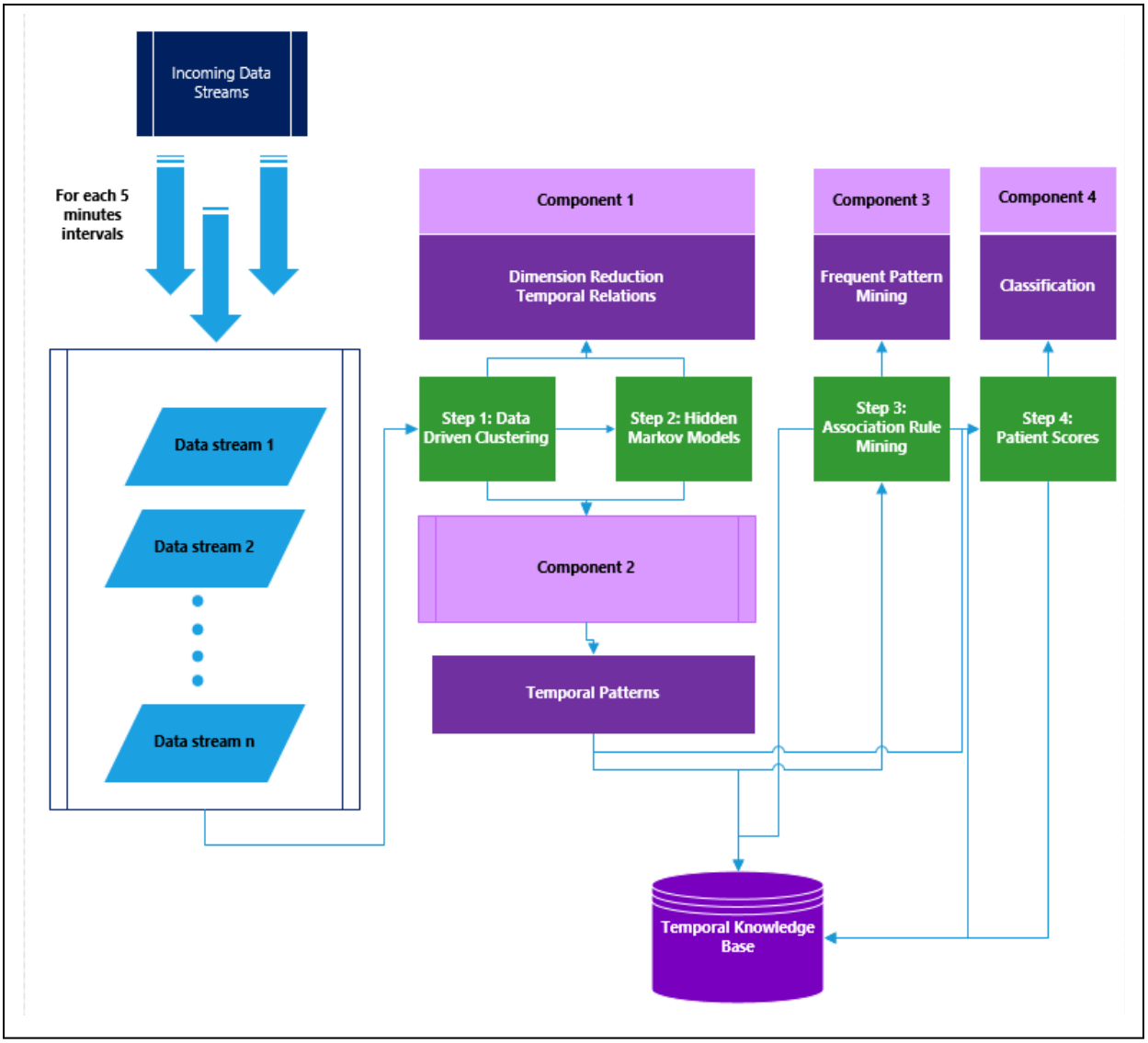
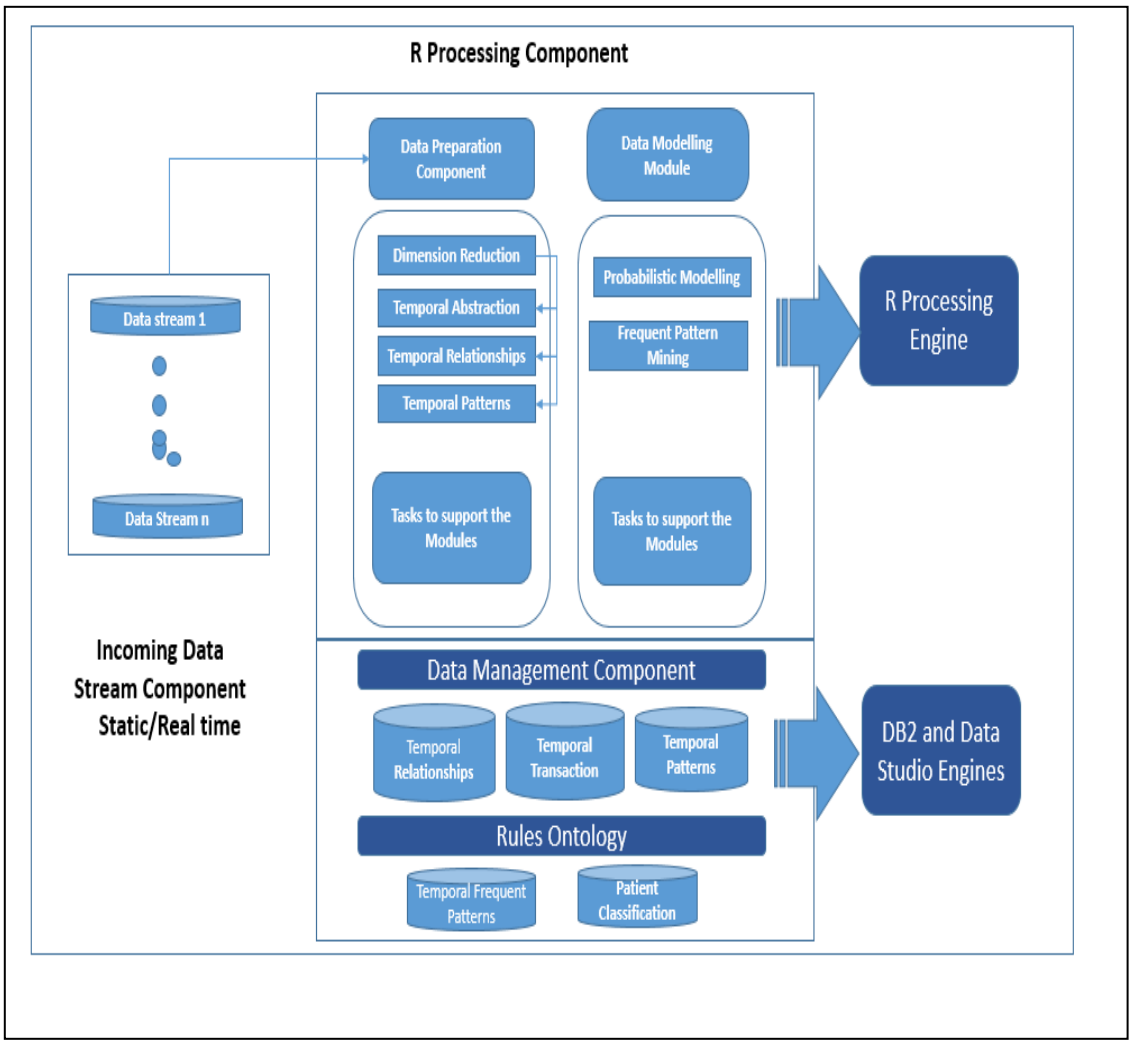


Figure 7. 2 : Temporal Pattern Recognition Process in R



7.2 Evaluation

There are three layers of testing that has been used to evaluate the overall performance of the proposed method. For feasibility, the evaluation has been structured based on research contribution to each domain. Clinical evaluation is out of scope of this thesis.

In the computer science contribution, several components are proposed that enhance the CRISP DM and CRISP-TDM frameworks with components for dimension reduction, generation

of temporal abstraction and discovery of frequent temporal patterns. As such, evaluating the effectiveness of the proposed algorithms to support these components is needed. To facilitate the testing, this thesis has used both theoretical and empirical evaluation approaches.

The theoretical approach involves assessing the capability of the proposed algorithms to generate effective knowledge about the underlying data and compare to existing CRISP-DM and CRISP-TDM frameworks. To facilitate this process, this thesis has utilized the qualitative system evaluation theory techniques for complex systems as detailed (Latham, 2014).

7.2.1 Qualitative Evaluation

7.2.1.1 Comparison of CRISP-TDM with Enhancement in TPRMine

First a comparison of the CRISP-TDM compared to the new enhancement proposed with TPRMine. Figure 5.1 in chapter 5 shows the extensions proposed to incorporate the TPRMine process. Four components developed do not exist in the current CRISP-TDM model as such the functions included in the new enhancement are not in there in the old model.

In the existing data preparation process, there is no process to handle the large volume of data generated from many clinical settings such as the NICU where every second multiple physiological time series data are generated for each patient in an NICU bed. Component 1 includes functions to scale the data into temporal abstractions which are then used as high level representation of the underlying dataset. The temporal abstractions generated are then quantified with mathematical modelling adopting HMM leading to generation of temporal sequences and temporal patterns in a specified period of time. In addition to understanding

the underlying relationships in time series data, the temporal abstraction approach utilized in this thesis is a dimension reduction strategy that facilitates the high level representation of time series data which can then be supplied to the modelling process.

Next a component is added in the Modelling phase which provides an ability to understand which are the frequent temporal patterns by incorporating principles in association rule mining shown in (Agrawal & Srikant, 1995) and event sequence mining in (Liu, Wu., & Zhang, 2014). The identification of frequent temporal patterns based on some underlying data domain facilitates building a classification system for which to generate new hypothesis on a given data stream. This is demonstrated using the vital scoring of 2 different data domains (NICU and Elderly Care) as a means for classifying the states a patient can transition to in a given period of time.

A summarised comparison of enhancement to the existing CRIPS-TDM by the TPRMine approach is shown in Table 7.1. This comparison includes 5 key attributes and associated research questions following the qualitative system evaluation method detailed in (Latham, 2014).

Table 7. 1: Functional Comparison of CRISP-TDM and the TPRMine Frameworks

Function	Questions about the Framework	CRISP-TDM	TPRMine
Scalability	What is the ability to handle multiple data streams	Designed to process one data stream at a time	The framework developed allows scaling of period of time on what data to be processed, allowing processing of multiple high frequency data streams
Temporal Abstraction	Can the framework generate temporal abstractions that maintain the temporal nature of the original data	Only simple abstractions are possible at houly basis and has no way of maintaing temporal nature of	The framework developed incorporates methods for generating temporal abstractions that are highlevel representation of the underlying data. This approach also allows understanding relationships on abstractions within a given time interval thereby generating temporal patterns
Knowledge Discovery	Can the framework support understanding temporal relationships in data?	No components for understanding relationships among temporal abstractions or deriving temporal	The framework incorporate processes to detect, represent, understand relationships and classify temporal patterns from real-time data streams with capabilities to filter and prune irreverent patterns.
Classification System	Can the framework support building a classifcation system using temporal patterns	No such component as no process for deriving temporal patterns	The framework includes components to integrate domain knowledge thereby able to classfy events or episodes from the temporal patterns discovered
Data Management	Can the framework support storage of temporal data processed	Minimal data stored limiting further forensic analysis	The framework includes components to store temporal data generated which allows possiblity of future forensic evaluation of the entire process flow

7.2.1.2 Comparison of $STDM_0^n$ to Enhancement in TPRMine

In Contribution to Health Informatics this thesis proposes research to extend the STDM_0^n framework by (McGregor, 2013). These enhancements are characterized in purple in Chapter 6 (Figure 6.2). In particular, this research proposes extensions to; data preparation and modelling phases, rules ontology and data management layers.

With respect to data preparation a new component is added to support dimension reduction, generation of temporal abstractions using temporal sequences, understanding relationships among sequences and generation of temporal patterns. To support this

component, an extra task for dimension reduction and similarity measures is added in $STDM_0^n$ Agent Tasks. This also involves adding more functions to the temporal and relative agents which will need to perform the tasks in the added component.

With respect to modelling, a component for frequent pattern mining and building a classification model is added and an enhancement to include more tasks in the $STDM_0^n$ Agent has been added. This allows the functional agent to facilitate real-time temporal frequent pattern mining including the ability to support null hypothesis testing on frequent patterns.

Regarding the $STDM_0^n$ Rules Ontology, the discovered temporal frequent patterns forms an extra set of rules that are dynamically updated as new temporal patterns are generated.

The data management layer is also enhanced to include an extra set of data comprised of unique temporal frequent patterns and their quantification metrics, this is vital to allows classification of study cohorts based similarity measures. To facilitate this process, this thesis incorporates application of similar principles attempted in temporal databases for holding and updating frequent temporal patterns and their associated timelines (Combi et al, 2010).

Table 7.2 shows a summarized comparison of the components in $\mathrm{STDM}_0^{\mathrm{n}}$ to TPR Mine extension.

Table 7. 2: Comparison of $STDM_0^n$ vs TPRMine Process

Component	Questions about the Framework	$STDM_0^n$	TPRMine
Data Preparation	Does the framework include components for reduction of data processed	No Component for Dimension Reduction	Framework includes a component to support dimension reduction, generation of temporal abstractions using temporal sequences, understanding relationships among sequences and generation of temporal patterns.
Modeling	Does the framework support identification of frequent patterns	Framework doesn't not include components for mining frequent patterns	Includes a component for frequent pattern mining and building a classification model. This allows the functional agent to facilitate real-time temporal frequent pattern mining including the ability to support null hypothesis testing on frequent patterns.
Data Management	Does the framework incorporate component for additonal data generated in the process flow	No structures to house temporal pattern generated in the modeling phase	The data management layer is also enhanced to include an extra set of data comprised of unique temporal frequent patterns and their quantification metrics, this is vital to allows classification of study cohorts based similarity measures.

7.2.1.3 Enhancement to the Artemis Framework on Knowledge Extraction

Instantiation of the existing CRISP-DM and $STDM_0^n$ is accomplished in the current ARTEMIS architecture. In particular in the application to physiological data streams captured in NICU, two algorithms that adopted the existing frameworks were developed to understand the variability of HR and RR from patients in NICU.

In continuation of creating a robust decision support system for clinical application, this thesis has enhanced the knowledge extraction component in Artemis (McGregor, 2009). As such we have demonstrated the ability to incorporate modules for frequent pattern mining algorithms that generate and quantify temporal patterns to support development of a classification system to classify normal and abnormal patterns from time series data.

Two existing algorithms have been developed that are part of the existing knowledge extraction process within Artemis. The two algorithms process only 1 data set at a time. The

HR variability algorithm and the RR variability. Even though these two algorithms have been effectively applied in clinical care in detection of patients at risk of LONs, they do not provide an understanding of the underlying relationships within the underlying datasets and only process different patients in parallel. As such it's not possible to join analytics from multiple patients to understand characteristics of different patient cohort.

With the TPR Mine algorithm, this thesis is able to process multiple datasets at the same time, generating the temporal patterns a patient may transition to at a given period of time, deriving what are the frequent patterns among different cohort of patients and then building a classification system.

7.2.2 Empirical Evaluation

The empirical approach involve assessing the functionality of the proposed framework as both static and real-time environments. To facilitate this process, this thesis reviewed architectural and performance metrics.

On assessment of the architecture some of the principles assessed include; load balancing, data storage and sensitivity adjustments when handling varying time series data streams in discovery of frequent temporal patterns.

On the performance metrics, this thesis emulates the profiling principles from the Harstone Benmark that defines tests for measuring execution times in real-time systems details in (Donohoe, Shapiro, & Weiderman, 1990).

Performance profiling determines how much time a system spends at each component and how varying workloads affects each component. The metrics measured includes;

- Latency: How much time it takes to process data at each component
- Response time: End to end system execution time
- Throughput: How much data processed successfully over a given period of time

7.2.2.1 Latency Test

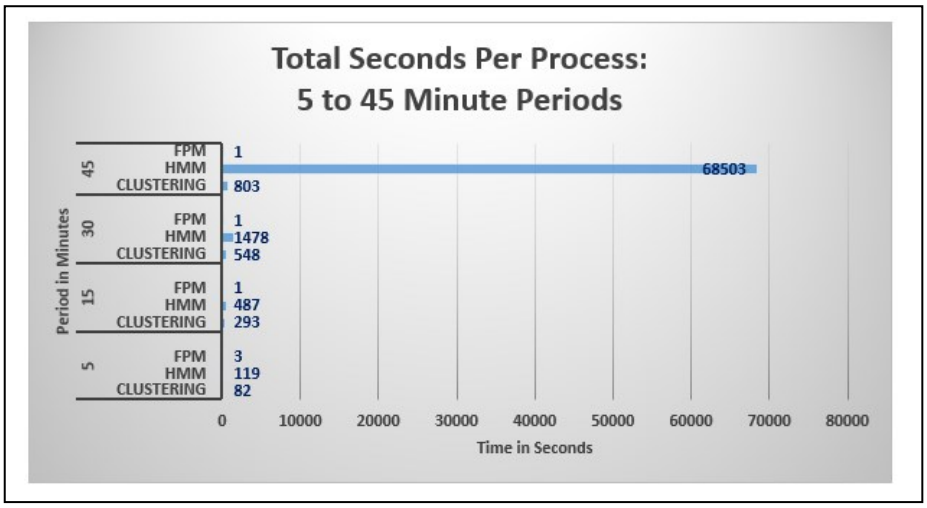
A comparison of how much time it takes for each component to process data in a particular time window is completed as shown in Figure 7.3. The use of time windows is a significant part of the thesis contribution with respect to effective scaling in order to handle real-time data streams processing. As such, it is important to understand how changes to time windows and an increase in data streams impact the components in TPRMine.

The TPRMine has adopted a widowing strategy where data for a specific time interval is processed. As this thesis utilizes high frequency data streams, then there has to be an efficient process that effectively manages input into multiple components. These components have functions that may take some time to generate an output. An evaluation of TPRMine is completed using system evaluation functions in (R-Benchmark, 2017) with respect to time and space required in each of the components for clustering, HMM and Frequent patterns as follows.

Selecting data generated at a random hour, different time periods (windows) are formed as follows, period 5 includes 5 minutes worth of data, period 15 contains 15 minutes of data

and so on till period 45 that contains 45 minutes worth of data. Note that windows less than an hour are utilized in order to assess how efficient the processes is at gathering patterns at 3/4 of an hour. Figure 7.3 shows that it takes 82 secs to complete clustering, 119 secs to compete HMM and 3 secs to complete FTP. As the time window increases, so does the number of seconds it takes to process each of the components. At 45 period, it takes 68503 secs to process HMM component.

Figure 7. 3 : Total Latency

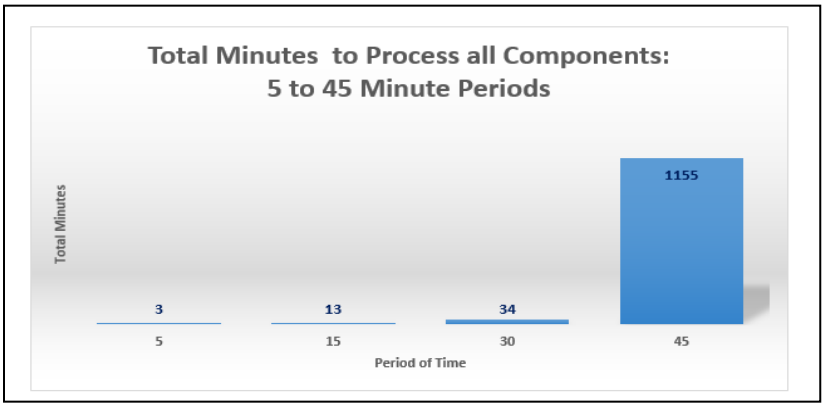


7.2.2.2 Response Time Test

In response time an evaluation of the amount of time it takes to complete the TPRMine process is completed by combining the times for each of the 3 main components, clustering, HMM and FTP. Figure 7.4 shows the amount of time in minutes it takes to complete the three processes. It takes 3 minutes to process data captured in 5 minutes, 13 minutes for data

captured in 15 minutes, 34 minutes for data captures in 30 minutes and 1155 minutes (over 19 hrs) for data captured in 45 minutes.

Figure 7. 4 : Total Response Time for 3 Major Components in TPRMine Process



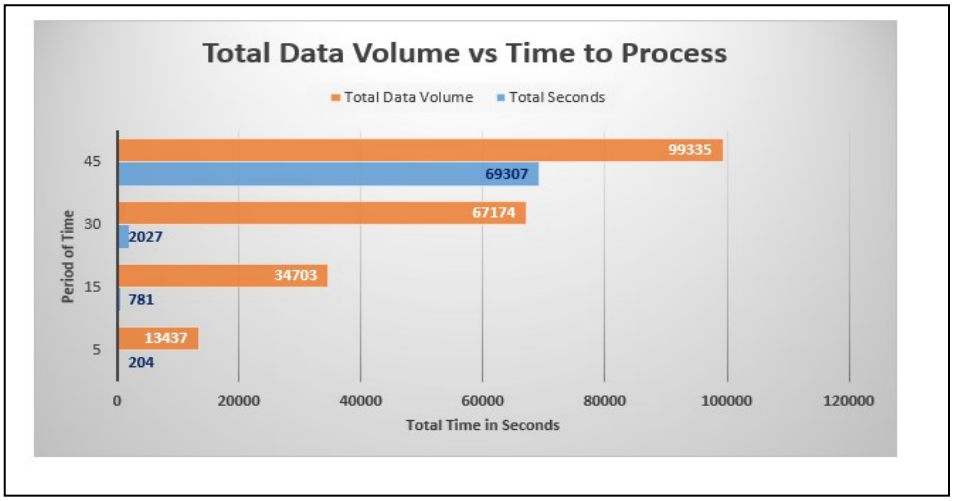
7.2.2.3 Throughput Time Test

A combination of latency and throughput is completed in order to understand how much data is processed for a specific time window, the results are included in Figure 7.5. The volume of data is in the total number of records within a specific time interval i.e. at period 5, it takes 204 seconds to process 13,437 records while at period 45, it takes 69307 secs to process 99,335 records. These results shows that TPRMine is efficient at processing 5 minutes data windows which then allows effective processing of continuous data streams as the delay between 5 minutes to the next 5 minutes is less than 3 minutes.

Another aspect is that, as the data volume changes between 5, 15, 30 and 45 time intervals, there was no change detected in the number of unique data driven clusters generated when processing the different sizes of data. To this respect, in addition to the minimal time it takes for TPRMine to process 5 minute data windows, selection of 5 minute data windows has no impact on the overall unique clusters generated. From a clinical

perspective, this demonstrates that heart rate variability changes were present over a longer period of time, for other clinical context where changes appear less frequently such as detection of bradycardia isolated events, the results are expected to be different and this highlights why the use of 5 minute intervals are important as a standard.

Figure 7. 5 : Combination or Latency and Throughput



Reliability and Validity Testing: The evaluation of health informatics and medicine contributions are unified into one process where reliability and validity of the framework is assessed. This process allows answering the following questions (a) Can identification of frequently occurring temporal patterns be used in generating data driven clinical rules that can form a classification system and (b) how effective the developed classification system is able to classify normal and abnormal patterns from time series data.

To facilitate this process, testing involved instantiation of two different clinical case studies as discussed in section 6.3.

In the first case study: Using physiological data streamed from monitors at the bedside in NICU; how effective is the framework at enabling the generation of temporal patterns, identification of frequent patterns and building a classification. Additionally, how effective is the framework at identifying normal and abnormal behavior in data from differing neonatal patients who may be at risk of LONS. McGregor et al has developed algorithms for classifying patients at risk of contact LONS. This is a threshold based approach as shown in Chapter 6, Figure 6.7.

The TPRMine patient scoring algorithm is tested against the HRV/RRV thresholds on how well it is able to quantify patients at risk of LONS.

In particular, after running the HRV/RRV algorithm on patients 6 hours worth of data, results shows various threshold score for different patients. Two patients are randomly selected for evaluation. One patient has no variability threshold falling in score D, while another patient is at score D for the entire 6 hours. Now applying TPRMine algorithm to the same patients data identifies that for the patient in score D, the number of state transitions is much less compare to patients in thresholds A or B.

These results would indicate that there are different temporal patterns on a patient detected at risk of LONS compared with other patients who fall in other HRV/RRV thresholds.

Second Case Study: Using data captured from RPM, the evaluation goal is to test how well the framework is able to develop a classification system that detects an adverse event such as a potential hospital admission or emergency room visit.

Using 3 periods of time before during and after an adverse event such as ER Visit, TPRMine has detected different patterns on a random cohort of patients. In particular, before and during the ER Visits, there were a higher number of states identified as a danger compared to the period after the event.

Reliability Testing: The TPRMine method is able to handle multiple patients data including data from two different domains and effectively scale to handle high frequency data from NICU as well as low frequency data from RPM. In NICU, approximately every second a record is generated, in 1 hr this results to approx. 3518 tuples generated within the heart rate, respiratory rate, SpO_2 and pulse pleth streams resulting in a total of 14,400 records for a patient. With 36 patients in NICU results to 518,400 records generated in 1 hour. The TPRMine method has been applied on 11 hours of 36 patient data effectively processing over 5.7 million records. The two case study also demonstrate the ability to scale input data streams, in NICU 4 different data streams are utilized and in RPM 5 data streams are used.

Hypothesis-Testing Validation: With respect to clinical validation the TPRMine process allows hypothesis testing thus facilitating quantification of temporal patterns obtain from different patient cohort.

In Neonatal Case study: Incorporation of null hypothesis testing allows quantification of different patterns against an alternate hypothesis with respect to patients at risk of contacting LONS.

A null Hypothesis test is performed to assess if a newborn baby born premature and at risk of contacting LONs has physiological features that exhibit differing temporal patterns in given period of time. To facilitate this process, the HRV/RRV threshold results are combined to form data two types of patients. By using 6 hours of patients data, the patient when HRV/RRV threshold was at Score D has a lower percent of state transitions compared to a patient in threshold scores A and B as demonstrated in chapter 6, section 6.3.

With respect the elderly patients RPM data, the number of states that are classified as abnormal or in danger are more in the period before or during an adverse event.

Consistency Testing: The final layer of evaluation assess the consistency of the framework by measuring repeatability and reproducibility. These two measures allow assessing the ability to reproduce similar results from varying patient cohort and in differing testing times. To support the evaluation process, two cohorts of data that is randomized to avoid any experimental bias is utilized. Results of running both NICU and RPM data shows consistent output across all the processes. For ease of evaluation of the consistency, this thesis adopt principles in (Ivanov, Ali-Löytty, & Piché, 2014) and utilize the transition matrices from the HMM component has generated the same probability distribution and a slight change in the underlying data is well represented in the resulting matrices.

Chapter 8

8 Conclusion and Future Research

8.1 Summary and Contribution

Currently, a significant amount of data is generated every second in many healthcare settings. Particularly, in critical care, bedside monitors generate thousands of numeric and wave data per second. For example, numeric data such as heart rate, respiratory rate, and blood oxygen saturations generate 1 reading per sec (1Hz) resulting to about 3600 readings per hour. Additionally, a single electrocardiogram wave generates about 7680 readings each second resulting in approximately 460, 800 readings per hour. As noted in chapter 2, trying to analyse such data to detect any temporal patterns that may exist is a central challenge and current methods to process this data becomes like a black box to end users resulting to manual annotations recorded from monitors every 30-60 minutes (McGregor et al., 2012). This is an ineffective process for the early detection of life threatening conditions such as sepsis where early discovery can lead to prompt treatment.

This thesis aimed to address the current challenges described in chapter 1 by extending existing knowledge discovery frameworks thereby proposing a method to detect and represent relationships that may exist in temporal abstractions (TA) and temporal patterns (TP) derived from time oriented data. This process involved incorporating techniques for down sampling data streams and in so doing introducing information about the behaviour of the stream. In clinical care, down sampled streams captured from patients can represent higher abstractions of behaviours in the physiology of patients. Providing such abstractions

as input to mining algorithms can lead to discovery of unknown relationships among physiological time oriented data. Consequently, the discovered relationships could lead to detection of onset of conditions and aid in classifying abnormal or normal behaviours in patients. This process could be vital to clinicians and health care providers as they make critical decisions on best care for patients.

This thesis achieves this goal by developing a method that extends the existing knowledge discovery frameworks to include components for detecting and representing temporal relationships in time series data. Instantiation of the developed method is completed within a big data analytics platform and applied to the analysis of streaming patient data. This process leads to research contributions in three research domains computer science, health informatics and medicine.

In computer science contribution, an extension to CRISP-TDM^n model is completed in this thesis by adding 4 components to the existing system with functions for understanding relationships among abstractions, generating temporal patterns, mining frequent occurring patterns and building a classification system.

To facilitate this process, this thesis introduced TPRMine process and an associated TPRMine algorithm which adopts a stepwise approach to temporal pattern discovery by first applying a scaled mathematical formulation of the time series data. This is achieved by modelling the problem space as a finite state machine representation where for a given timeframe, a time series data segment transitions from one state to another based on some

weights. This is accomplished by using markov chains that associates probability distributions to likelihood of transitioning from one state to the next.

To be able to understand how many states a process can transition to in a given time period, data driven clustering have been adopted. This approach allows higher representation of underlying data with cluster centroids at the same time understanding the different states data may transition to for a given time period. Representation of high frequency data with their cluster centroid is a dimension reduction strategy that allows scaled data streams to be passed on as input for further processing including mining frequent patterns.

Contribution to health informatics is completed by enhancement of the temporal multi-dimensional knowledge discovery framework in (McGregor, 2013) by incorporating the new components developed within the computer science contribution of this thesis. Instantiation of this process is completed by extending the knowledge extraction component in Artemis (McGregor, 2009) consequently continuing to create a robust decision support system for clinical application. These enhancements are demonstrated in chapter 6 whereby the proposed TPR process and TPRMine algorithm is applied to patient physiological data thus generating and quantifying temporal patterns. These patterns are then used for development of a classification system to classify normal and abnormal patterns as well as patient vital scoring.

Contribution to medicine is demonstrated in two clinical case studies. First application is in the NICU context where supplying patient physiological data streams as input to TPRMine

algorithm enhances the threshold scoring approach in (McGregor, 2012). This is completed by integrating a more granular data representation to demonstrate that application of TPRMine to multiple physiological data streams from patients in NICU could enable discovery of temporal patterns within the hour of high or low threshold.

This is accomplished by generation of data driven clusters including artefacts such as cluster means and cluster variances which can be utilized for retrospective analysis. Additionally, the data driven clusters forms the states that a patient in NICU may transition to in a given period. Quantification of the total number of states a patient may transition to in a given period of time is completed using a scoring mechanism. A demonstration of formulation of hypothesis tests using the quantified states is completed. This process can facilitate clinical case studies that could lead to discovery of pathophysiological patterns in physiological data streams captured from patients at risk of conditions such as LONS.

The second clinical application is in elderly patients who received care using a RPM program. Earlier research indicated there are specific factors that leads to multiple visits to emergency rooms. However, even though we had good predictors of an adverse event in earlier publications related to this research as detailed in (Inibhunu, Schauer, Redwood, Clifford & McGregor, 2017), we still did not understand any difference in patterns leading to and after an adverse event.

Using data such as average pulse rate, SpO_2 and patient's weight, TPRMine allowed formulating 3 different temporal periods; before, during and after an ER Visit thereby

facilitating generation of temporal relationships and patterns on each of the associated temporal periods.

The overall method demonstrated in this thesis extends the STDM_0^n by (McGregor, 2011) as such several data structures are added to the original framework allowing more knowledge base to be captured for each period of time. These new data structures allow further analysis of the derived temporal patterns thereby facilitating formulation of new hypothesis about the underlying data domain. In the NICU, this allows formulation of hypothesis with respect to detection of changes in patients suspected of diagnosis of conditions such as LONS. In elderly care this process allows the formulation of hypotheses on the difference between patients at risk of an adverse event and those who have no risk.

Evaluation of the functional performance of each of the components in TPRMine shows that selecting a shorter period of time for which to process data leads to faster processing of continuous data streams. As data volume increases, the time it takes to process 45, 30 and 15 minutes worth of data is 340, 10, and 4 times respectively as compared to processing 5 minutes worth of data. Additionally, within the case study contexts of this thesis, the number of unique clusters generated is not impacted by reduction of data size. As such, this thesis demonstrates that the approach adopted is efficient at processing high frequency data streams.

8.2 Limitations and Challenges

This research falls under the big data paradigm of high volume, velocity, and variety and each of these have their own unique challenges. In NICU, the amount of data generated each second from each patient in a bedside is quite large and it's not an easy task to sift through this massive amount of data when trying to get meaningful understanding of any underlying patterns that may exist within the data.

In order to handle the huge volume of data streams, this research has first use a static approach on generating temporal patterns and then apply this process to a scaled set of real-time data streams captured over a 12 hour period. Scaling of data streams can be done by size or time, this research has investigated a combination of these two approached to reduce the computation needed to generate temporal patterns. More work needs to be completed to assess the system stability with more data volumes, example is processing several months of NICU data.

Another challenge is the sheer variety of data generated from monitoring and sensory devices in critical care. This makes it even more challenging when trying to correlate patterns from different data points that all have unique contextual meaning with respect to the state of a patient. This research realizes, it's not feasible to correlate all the different data sets, as such, a limited number of physiological features were processed in both clinical case studies. Additional datasets can be added when the proposed system is stable and research in Artemis progresses. This includes adding components for standardizing the variety of data streams

With respect to data management, this thesis has utilized relational databases, as such enforces all output to be in some structured manner. However, when one needs to make a trade off between time, space and volume, future works can explore other storage structures such as the NOSQL databases in (Ke, Gong, Li, & et al., 2014) for storing semantically indexed temporal data.

8.3 Future Research

This thesis introduces a foundational framework for further knowledge discovery especially in domains where high frequency data about patients in NICU is generated every second. This thesis demonstrates a prototype that uses data from two different clinical domains and results shows an ability to quantify the different states a patient may transition to which could be associated with an underlying clinical condition. These results will be included in upcoming research publications.

However, many more data streams are captured from patients every second such as blood pressure, ECG waves etc. TPRMine algorithm shows it's scalable to allow inclusion of additional data streams, further more, there is potential to include more details about clinical diagnosis such as Apnea, IVH etc. As such, the process introduced in TPRMine serves as a foundation for further knowledge discovery in clinical settings where data about patients continues to be generated every second. The ability to quantify all this data using TPRMine could be instrumental to decision making about patient case in critical conditions.

With respect to scalability, two types of scalability are demonstrated in the TPR process proposed in this thesis, (a) allowing inclusion of multiple and diverse data streams, (b)

application to different data domain, NICU and Elderly patients. Another aspect of scalability is to facilitate separate knowledge discovery process based on different context about a data domain. There may exist compress facts and information about any data domain under investigation, as such this thesis recognizes that integration of such context is required before the application of TPRMine algorithm.

In particular, in NICU, there are fundamental differences among patients admitted for care. There are neonates who are preterm and those that are full term infants, preterm have very low birth weight and are 3 to 10 times more likely to be predisposed to neonatal sepsis as compare to full term infants with normal birthweight (Hoogen, 2009). To this respect the temporal behaviors on data generated from such cohort of patients can be very different. As such, TPR Process can therefore be applied separately to these different groups of patients resulting to differing knowledge base. This is enabled in components for data preparation proposed in this thesis allowing the ability to scale the selection of temporal windows using different patient properties. In this case two sets of patients *(setA, setB)* where properties in *setA* are different from properties in *setB*, then *setA* and *setB* are two problem spaces and each with unique characteristics and thus requires application of TPRMine exclusively. All patients in setA have similar properties and can be termed as a homogenous population. For example, if data domain is in NICU and cohorts of patients are identified as diverse i.e. preterm and full term infants, then each of these patients have unique characteristics and therefore would require application of TPR Process separately. In particular, those preterm can be termed as a homogenous population that is quite different from the full term population. Demonstration of application of TPR Process and TPRMine algorithm into such

different cohort of NICU patients can be completed in the future thereby understanding temporal behaviours among different groups of NICU patients which could highlight susceptibility to certain conditions such as neonatal sepsis.

With respect to the TPRMine algorithm, the current application assumes data is already standardized in order to formulate the temporal data windows. However, future enhancements can be completed by integrating a data standardization component especially when augmenting multiple data streams of varying data types and from vast data domains.

Another aspect is storage. As noted in the literature, storage of time-oriented data requires temporal databases that are capable of holding events and transaction times. TPRMine has utilized relational data structures for data management. Future enhancements can be explored to utilize non-relational (NOSQL) databases for storing the continuous temporal knowledge discovered from processing high frequency data streams with TPRMine.

8.4 Concluding Statements

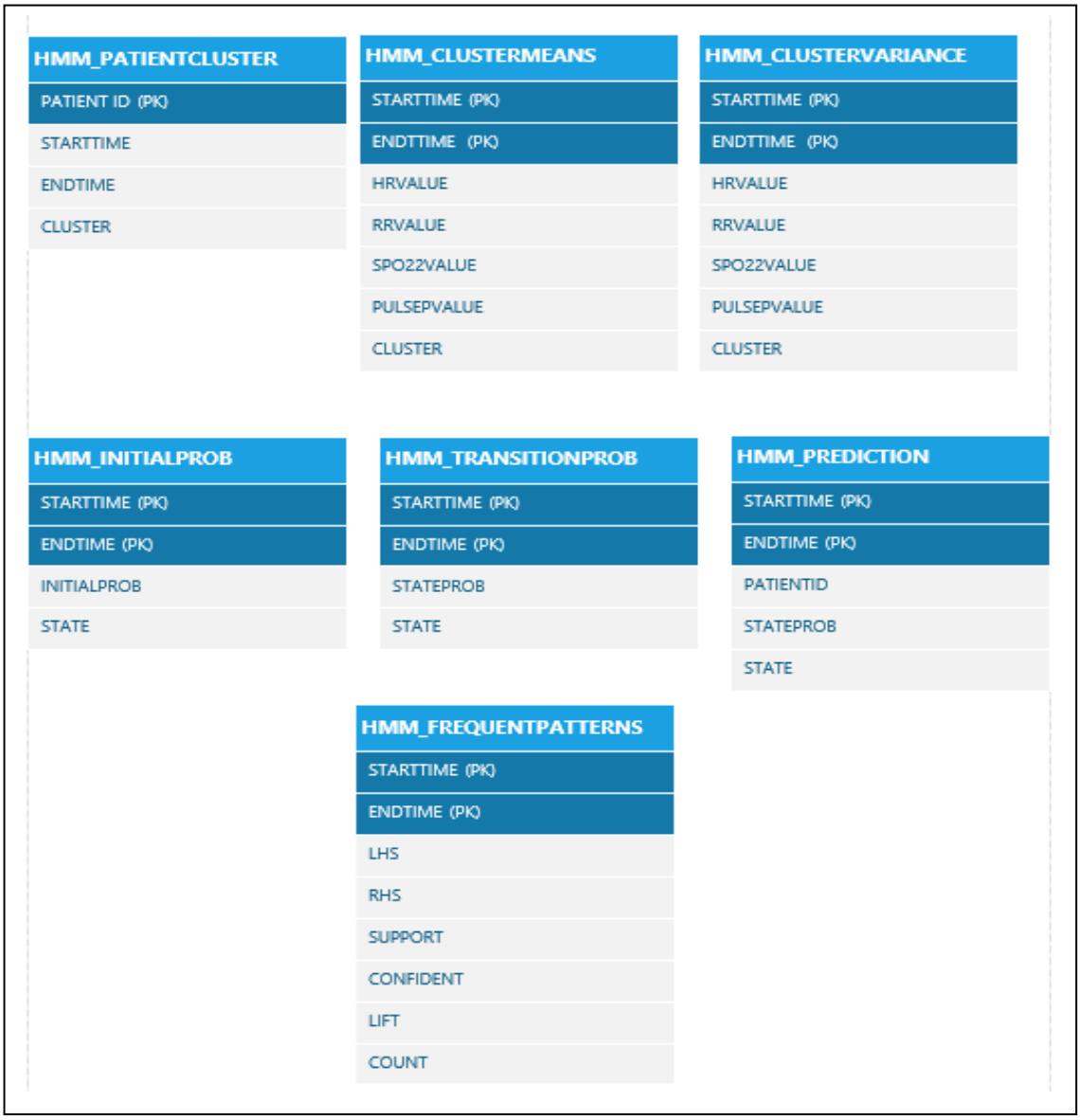
As vast amount of data continue to be generated from monitors and sensors in many complex systems and data domains, so is the need for utilizing such data when making critical decisions. Particularly, in clinical settings, sensors continuously monitor patients physiology resulting to significant amounts of data generated every second. If temporal patterns can be derived from such data and available to clinicians promptly, they could utilize it in critical decisions on care of patients. Existing systems attempting to derive hidden knowledge from such data are often overwhelmed as they are not built to handle the volume, variety and speed of continuous data streams.

This thesis has addressed these challenges by proposing a method that extends the existing knowledge discovery frameworks to include components for detecting and representing temporal relationships and patterns in time series data. Furthermore, instantiation of the developed method within a cloud based knowledge discovery platform has demonstrated the potential for processing high frequency physiological data streams leading to detection of temporal behaviours in vast patient cohort.

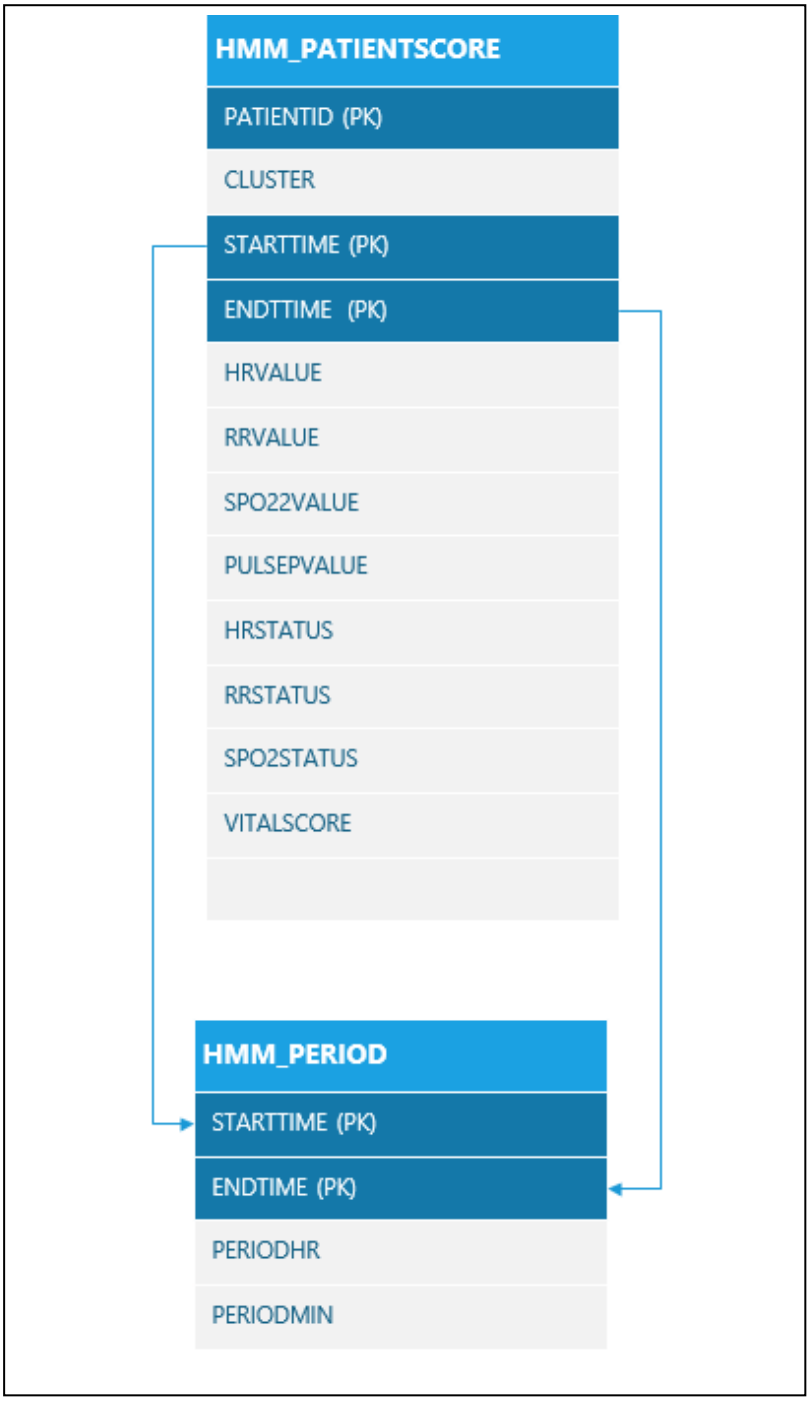
There is much still to be learned about the temporal behaviour of many things and by introducing the proposed method, this thesis not only provides a framework to explore further such temporal behaviours, but also offers a process for advancing temporal knowledge discovery.

Appendices

Appendix 1a Temporal Data Structures



Appendix 1b Temporal Data Structures



References

Abdur, F., & Ibrahim, K. (2016). A Probabilistic model for early prediction ofabnormal clinical events using vital sign correlations in home-based monitoring. 2016 IEEE International Conference on Pervasive Computing and Communications.

Agrawal, R., & Srikant, R. (1995). Mining Sequential Patterns. *Proceedings of the International Conference on Data Engineering (ICDE).* Washington DC.

Ali, R., Elhelw, M., Atallah, L., & et al. (2008). Pattern Mining for Reouting Behaviour Discovery in Pervasive Healthcare Environments. . *IEEE Int Conference in Information Technology and Applications in Biomedicine.*

Altiparmak, F., Ferhatosmanoglu, H., Erdal, S., & Trost, D. C. (2006). Information Mining Over Heterogenerous and High-dimensional Time-Series Data in Clinical Trials Databases. *IEEE Transactions in Information Technology in Biomedicine, 10*(2).

Alvarez, M. R., Felix, P., & Carinena, P. (2013). Discovering Metric Temporal Constraint Networks on Temporal Databases. *Artificial Intelligence in Medicine, 58*, 139-154.

Amer, A. Y., Julie, V., Femke, W., Dieter, M., Pieter, V., Valerie, S., . . . Jean-Marie, A. (2019). Feature Engineering for ICU Mortality Prediction Based on Hourly to Bi-Hourly Measurements. *Appl. Sci. https://doi.org/10.3390/app9173525*, 3525.

American Academy of Pediatrics. (2019). Retrieved from https://www.healthychildren.org/English/family-life/health-management/pediatric-specialists/Pages/What-is-a-Neonatologist.aspx

American Heart Association Guidelines. (2005). *Management of Symptomatic Bradycardia.* Retrieved from DOI: 10.1161/CIRCULATIONAHA.105.166558

Askie, e. a. (2003). Oxygen-Saturation Targets and Outcomes in Extremely Preterm Infants. *N Engl J Med;*, 349:959-967 DOI: 10.1056/NEJMoa023080.

Baker, T., Schell, C. O., Lugazia, E., Blixt, J., Mulungu, M., Castegren, M., . . . Konrad, D. (2012, 10). Vital Signs Directed Therapy: Improving Care in an Intensive Care Unit in a Low-Income Country. *PLoS ONE 10(12)*.

Batal, I., Sacchi, L., Bellazzi, R., & Hauskrecht. (2009). A Temporal Pattern Mining Approach in Classifying Electronic Health Records Data. *AMIA Symposium Proceedngs.*

Batal, I., Valizadegan, H., & Cooper, G. H. (2012). A Temporal Pattern Mining Approach for Calssifyin Electronic Health Record Data. DOI: http://dx.doi.org/10.1145/2508037.2508044.

Batal, I., Valizadegan, H., Cooper, G. F., & al., e. (2013). A Temporal Pattern Mining Approach for Classifying Electronic Health Record Data. *ACM Transactions on Intelligent Systems and Technology, 4*(4), Article 63.

Baton, B., Li, L., Newman, N., Das, S., & al., e. (2014). Evolving blood pressure dynamics for extremely preterm infants. *Journal of Perinatology 34*, 301–305.

Beck, M., Jensen, A., Nielsen, A., Perner, A., Moseley, P. L., & Brunak, S. (2016). Diagnosis trajectories of prior multi-morbidity predict sepsis mortality. *Sci Rep 6, 36624*.

Bellazi , R., Larizza, C., & Magni, P. (2005). Artificial Intelligence in Medicine. *Temporal Data Mining for the Quality Assesment of Hemodialysis Service, 34*, 25-39.

Berney, S. C., Rose, J. W., Denehy, L., Granger, C. L., Ntoumenopoulos, G., & Elise Crothers. (2017). Commencing Out-of-Bed Rehabilitation in Critical Care—What Influences Clinical Decision-Making? *Critical Care Conference.* Honolulu, Hawaii.: Society for Critical Care Medicine. https://doi.org/10.1016/j.apmr.2018.07.438.

Blueprint Genetics. (2017, Mar). *https://blueprintgenetics.com.* Retrieved from https://blueprintgenetics.com/wp-content/uploads/2017/03/Validation_of_clinical_testing_ANAA41-02.pdf

Boaz. A., S. Y. (2005). A Framework for Distributed Mediation of Temporal-abstraction Queries to Clinical Databases. *Artificial Intelligence in Medicine, 34*, 3-24.

Borchania, H., Bielzaa, C., Torob, C., & et al. (2013). Predicting Human Immunodeficiency Virus Inhibitors using Multi Dimensional Bayesian Netwok Classifiers. *Artificial Intelligence in Medicine, 57*, 219-29.

Borrie, A., Jonsson, G., & Magnusson, M. (2002). Temporal pattern analysis and its application in sport: An explanation and exemplar data. *Journal of Sports Sciences*, 845-852.

Bouarfa, L., & Daneklman, J. (2012). Workflow Mining and Outlier Detection from Clinical Activity Logs. *Journal of Biomedical Informatics, 45*, 1185-1190.

Bressan, N., James, A., & C., M. (2012). Trends and Opportunities for Integrated Real Time Neonatal Clinical Decision Support. *Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatic*, (pp. 2-7). Hong Kong and Shenzhen, China.

C, T. (2015). Respiratory rate: overloofed but vital. Vol 23 (3).

Campbell, E. A., Bass, E. J., & Masino, A. J. (2020). Temporal condition pattern mining in large, sparse electronic health record data: A case study in. *Journal of the American Medical Informatics Association*, 558–566.

Campos, C., Juarez, J. M., Palma, J., Marin, R., & Palacios, F. (2011). Avian Influenza: Temporal Modelling of A Human to Human Transmission Case. *Expert Systems with Applications, 38*, 8865-8885.

Catley, C., Smith, K., McGregor, C., & Tracy, M. (2009). Extending CRISP-DM to Incorporate Temporal Data Mining of Multi-dimensional Medical Data Streams: A Neonatal Intensive Care Unit Case Study. *22nd IEEE International Symposium on Computer-Based Medical Systems.*

Chandrasekharan , P., & et, al. (2017). Apnea, Bradycardia and Desaturation Spells in Premature Infants – Impact of a Protocol for Duration of “Spell-Free” Observation on Inter-Provider Variability and Readmission Rates. *J Perinatol. doi: 10.1038/jp.2017.174*.

Chaovalitwongse w., F. Y. (2007). On the Tine Series K-Nearest Neighbor Classification of Abnormal Brain Activity. *IEEE Transactions on Systems, Man and Cybernetics - Part A Systems and Humans, 37*(6).

Chen. (1990). A Comparison on Neural Network models for Pattern Recognition. *IEEE CH2898-5/90/0000/0045*, 45-46.

Chen, C. H., Hong, T. P., & Tseng, V. S. (2012). Fuzy Data Mining for Time Series Data . *Applied Soft Computing, 12*, 536-542.

Chittaro, L. (2001). Information Visualization and Application to Medicine. *Artificial Intelligence in Medicine, 22*, 81-88.

Choi, Y., Bressan, N., James, A., Pugh, E., & McGregor, C. (2013). Design of temporal analysis of neonatal vagal spells at different gestational ages using the Artemis' framework. . , Pages e4–e5.

Ciaos, K., & Moore, G. (2002). Uniqueness of medical data mining. . *Artif Intell Med* , 1-24.

Ciddio, M., Mari, L., Gatto, M., Rinaldo, A., & Casagrandi, R. (2015). The temporal pattern of disease severity and prevalence in schistomiasis. *Chaos An International Journal of Nonlinear Science 25*, doi: 10.1063/1.4908202. http://dx.doi.org/10.1063/.

CIHI. (2015). Retrieved from https://secure.cihi.ca/free_products/NACRS_ED_QuickStats_Infosheet__2014-15_ENweb.pdf.

Combi, C., & Oliboni, B. (2012). Visually Defining and Querying Consistent Multi-Granular Clinical Temporal Abstractions. *Artificial Intelligence in Medicine, 54*, 75-101.

Combi, C., Keravnou-Papailiou, E., & Shahar, Y. (2010). *Temporal Information Systems in Medicine.* Springer Science+Business Media.

Combi, C., Pozzi, G., & Rossato, R. (2012). Querying Temporal Clinical Databases on Granular Trends. *Journal of Biomedical Inforamtics, 45*, 273-291.

Cukier, F., Andre, M., Monod, N., & Dreyfus-Brisac, C. (1972). Apport de I'EEG au diagnostic des hemorragies intraventriculaires du premature. *Revue d'Electroencephalographie et de Neurophysiologie Clinique*, 318-22.

Dagum, E. B., & Cholette, P. (2006). *Benchmarking, Temporal Distribution, and Reconciliation Methods in Time Series.* ISBN-10: 0-387-31102-5 .ISBN-13: 978-0387-31102-9: Lecture Notes in Statistics.

Dash, M., & Liu, H. (1997). Feature Selection for Classification. *Intelligent Data Analysis* , 1 131–156.

Denai, M., King, O. K., Ross, J. J., & Mahfouf, M. (2007). Online Qualitative Abstraction of Cardiovascular Hemodynamics for Post Cardiac Surgery Decision support. *IEEE Life Sciences Systems and Applications Workshop.*

Donohoe, p., Shapiro, R., & Weiderman, N. (1990). *Hartstone Benchmark User's Guide, Version 1.0.* Retrieved from https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=11273

Dougherty, G. (2013). *Pattern Recognition and Classification.* Springer. 2013 DOI 10.1007/978-1-4614-5323-9.

Egho, E., Jay, N., Raissi, C., & et, a. (2014). A contribution to the discovery o multidimensional patterns in healthcare trajectories. *J Intell Inf Sys DOI 10.1007/s10844-014-0309-4*, 42:283–305.

Fairchild, K. D., Sinkin, R. A., Davalian, F., Blackman, A. E., & et, a. (2014). Abnormal heart rate characteristics are associated with abnormal neuroimaging and outcomes in extremely low birth weight infants. *Journal of Perinatology*, 34, 375–379.

Fanaroff, J., & Fanaroff, A. (2006). Blood pressure disorders in the neonate: Hypotension and hypertensionSeminars in Fetal & Neonatal Medicine 11. 174 - 181.

Fernando, K., McGregor, C., & James, A. (2016). Correlation of Retinopathy of Prematurity and Blood Oxygen Saturation in Neonates using Temporal Data Mining: A Pilot Study: A Platform presentation. *IEEE EMBC.* Orlando, FL.

Fontana, N., & et al. (2014). *Changes in cardiorespiratory physiological data streams before, during and after intraventicular haemorrhage in neonates less than 32 weeks gestational age.* Oshawa, Canada: unpublished report at UOIT.

Fraley , C., & Raftery , A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 611–631.

Fraley, C., Adrian, R. E., Murphy, T. B., & Scrucca, L. (2012). *Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification and Density Estimation.* Technical Report No. 597, Department of Statistics, University of Washington.

Frencken et al. (2018). Myocardial Injury in Patients With Sepsis and Its Association With Long-Term Outcome. *Circ Cardiovasc Qual Outcomes. DOI: 10.1161/CIRCOUTCOMES.117.004040*.

Fu, T. (2011). Review on Time Series Data Mining. *Engineering ApplicationsofArtificialIntelligence*, 164–181.

Fuchs, E., Gruber, T., Pree, H., & Sick, B. (2010). Temporal Data Mining Using Shape Space Representation of Time Series. In *Neurocomputing* (pp. 379-393).

Gagniuc, P. A. (2017). *Markov Chains: From Theory to Implementation and Experimentation.* . NJ: John Wiley & Sons.

Goddard-Finegold, J., & Mizrahi, E. (1987). Understanding and Preventing Prenatal, Intracerebral, Peri- and Intraventricular Hemorrhage. *Journal of Child Neurology*.

Goncalves, H., Bernardes, J., Rocha, A. P., & Ayres-de-Campos, D. (2007). Linear and nonlinear analysis of heart rate patterns associated with fetal behavioral states in the antepartum period. *Early Human Development*, 83, 585–591.

Gordan, R., Gwathmey, J. K., & Xie, L.-H. (2015). Autonomic and endocrine control of cardiovascular function. *World Journal of Cardiology* , 204–214.

Gorst , S. L., Armitage, C. J., Brownsell, S., & Hawley, M. S. (2014). Home Telehealth Uptake and Continued Use Among Heart Failure and Chronic Obstructive Pulmonary Disease Patients: A Systemic Review. *Annals of Behaviour Medicine, 43*(3), 323-336.

Gotz, D., Wang, F., & Perer, W. (2014). A Methodology for Interactive Mining and Visual Analysis of Clinical Event Patterns Using Electronic Health Record Data. *Journal of Biomedical Informatics*.

Griffin , M., & Moorman , J. (2001). Toward the early diagnosis of neonatal sepsis and sepsis-like illness using novel heart rate analysis. *Pediatrics*, 97-104.

Griffin, P., O'Shea , M., Bisonette , A., Harrell , F., Lake , G., & Moorman , J. (2003). Abnormal Heart Rate Characteristics PrecedingNeonatal Sepsis and Sepsis-Like Illness. *Pediatric Research* , 920–926 .

Gronlund, J. U., Korvenranta, H., Kero, P., Jalonen, J., & Valimaki, I. A. (1994). Elevated arterial blood pressure is associated with peri-intraventricular haemorrhage. *European journal of pediatrics*, 153(11), 836–841.

Gul, M., & Carbas, F. (2009). Statistical pattern recognition for structural health monitoring using time series modeling: Theory and experimental verifications. *Mechanical Systems and Signal Processing 23(7)*, 2192-2204.

Gustafson, A. K. (2004). The Neuropsychological outcome of preschool and early school age children with Mild Neonatal Germinal Matrix-Intraventricular Hemorrhage. *Masters Thesis.*

Guyet, T., & Quiniou, R. (2008). Mining temporal patterns with quantitative intervals. *IEEE International Conference on Data Mining Workshops*, (p. DOI 10.1109/ICDM.).

H., C. (2014). Top-k Frequent Patterns Over Data Streams Sliding Windows. *Journal of Intelligent Information Systems, 42*, 111-131.

Hadad, A., Evin, D., Drozdowicz, B., & Chiotti, O. (2007). Temporal Abstraction for the Analysis of Intensive Care Information. *16th Argentine Bioengineering Congress and the 5th Conference of Clinical Engineering* . doi:10.1088/1742-6596/90/1/012074.

Haghighi, P. D., Gilick, B., Krishnaswamy, S., & et al. (2010). A Situation-Aware Adaptive Visualization for Sensory Data Stream Mining . *Sensor-KDD*, 43-58.

Hale, T. (2017, Jul 26). *How Much Data Does The World Generate Every Minute?* Retrieved from The IFLScience Newsletter: https://www.iflscience.com/

Han, J., & Aggrawal, C. C. (2014). *Frequent Pattern Mining.* New York: Springer.

Han, J., & Kamber, M. (2001). *Data Mining Concepts and Techniques.* Academic Press.

Hanna, B. D., White-Traut, R. C., Silvestri, J. M., Vasan, U., Rey, P. M., Patel, M. K., & et, a. (2000). Heart rate variability in preterm brain-injured and very-low-birth-weight infants. Biology of the neonate. *Biology of the Neonate*, 77(3), 147–155.

Heidjen , M. V., & Lucas, P. J. (2013). Describing Disease Processes using a Probabilistic Logic of Qualitative Time. *Artificial Intelligence in Medicine, 59*, 143-155.

Ho, T. B., Kawasaki, S., Le, S. O., & et al. (2003). Combining Temporal Abstraction and Data Mining to Study Hepatitis. *Prceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Hoogen, A. (2009). *Infections in neonatal intensive care, Prevalence, Prevention and Antibiotic use.* ISBN: 978-90-393-5119-2.

Hoppner, F., & Klawonn, F. (2002). Finding Informative Rules in Interval Sequences . *Intelligent Data Analysis, 6*(3), 237-256.

Huang, Z., Juarez, J. M., Duan, H., & Li, H. (2014). Reprint of ‘‘Length of stay prediction for clinical treatment process using. *Expert Systems with Applications, 41*, 274-283.

Huang, Z., Lu, X., & Duan, H. (2012). On Mining Clinical Pathways Patters from Medical Behaviors. *Artificial Intelligence in Medicine, 56*, 35-50.

Hui, L., Chen, Y., Weng, J. T., & Lee, S. (2012). Incremental mining of temporal patterns in interval-based database. *Knowl Inf Syst*.

Hunter, J., Freer, Y., & Gatt, A. (2012). Automatic Generation of Natural Language Nursing Shift Summaries in Neonatal Intensive Care: BT-Nurse. *Artificial Intelligence in Medicine, 56*, 157-172.

Huvanandana, J., Thamrin, C., Tracy, M. B., Hinder, M., Nguyen, C. D., & McEwan, A. L. (2017). Advanced analyses of physiological signals in the neonatal intensive care unit . *Physiological Measurement*.

Ibaraki, T. (1976). Theoretical Comparisons of Search Strategies in Branch-and-Bound Algorithms. *International Journal of Computer and Information Sciences*, 1-30.

Inibhunu, C., & McGregor, C. (2016). Dimension Reduction and Similarity Measures for Temporal Pattern Recognition in Critical Care. *IEEE EMBS.* Orlando, FL.

Inibhunu, C., & McGregor, C. (2016). Machine learning model for temporal pattern recognition. *IEEE EMBS International Student Conference (ISC)*, (pp. 1-4). Ottawa.

Inibhunu, C., & McGregor, C. (2017). Towards Temporal Pattern Discovery aided by Remote Patient Monitoring Services: A case study on ER Visit Factors . *WIML Workshop.* Long Beach, CA.

Inibhunu, C., & McGregor, C. (2018). Fusing Dimension Reduction and Classification for Mining Interesting Frequent Patterns in Patients Data. *Machine Learning and Data Mining in Pattern Recognition. MLDM* (pp. 1-15). New York: Lecture Notes in Computer Science, vol 10935. Springer, Cham.

Inibhunu, C., & McGregor, C. (2018). State Based Hidden Markov Models for Temporal Pattern Discovery in Critical Care. *Life Sciences Conference (LSC)* (pp. 77-80). Montreal, Canada: IEEE.

Inibhunu, C., Jalali, R., Doyle, I., Gates, A., Madill, J., & McGregor, C. (2019). Adaptive API for Real-Time Streaming Analytics as a Service. *41st EMB.* Berlin: IEEE.

Inibhunu, C., McGregor, C., Schauer, A., Redwood, O., & Clifford, P. (2016). The Impact of Gender, Medical History and Vial Status on Emergecy Visits and Hosptal Admissions: ARomte Pateint Monitoring Case Study. Sidney, Australia: IEEE Life Sciences.

Inibhunu, C., Schauer, A., Redwood, O., Clifford, P., & McGregor, C. (2017). Predicting Hospital Admissions and Emergency Room Visits using Remote Home Monitoring Data. *1st IEEE conference in Life Sciences.* Sidney: IEEE Press.

Ivanov, P., Ali-Löytty, S., & Piché, R. (2014). Evaluating the consistency of estimation. *Proceedings of 2014 International Conference on Localization and GNSS* (pp. 1-5). 10.1109/ICL-GNSS.2014.6934171. .

Jensen, R., & Shen, Q. (2008). Dimension ReductionCopyright. In *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches* (pp. 61-84). Institute of Electrical and Electronics Engineers.

Joshi, R., Deedee, K., Guo, C., Bikker, J., Feijs, L., van Pul, C., & Andriessen, P. (2019). statistical Modeling of Heart Rate Variability to Unravel the Factors Affecting Autonomic Regulation in Preterm Infants. *Scientific Reports*, 1-9.

Kamaleswaran, R., Collins., C., James, E., & McGregor, C. (2016). CoRAD: Visual Analytics for Cohort Analysis. Proceedings in. *IEEE International Conference on Healthcare Informatics*, (pp. 516 - 526). Chicago.

Kamsu-Foguem, A., Tchuente-Foguem, G., Allart, L., & et al. (2012). User-Centered Visual Analysis using a Hybrid Reasoning Architecture for Intensive Care Units. *Decision Support Systems, 54*, 496-509.

Kasanen, E., Lukka, K., & Siitonen, A. (1993). The constructive approach in management accounting. *Journal of Management Accounting Research*, 243-264.

Ke, S., Gong, J., Li, S., & et al. (2014). A Hybrid Spatio Temporal Data Indexing Methos for Trajectory Databases. *Sensors, 14*(7), 12990-13005.

Khazaei, H., McGregor, C., Eklund, J., El-Khatib, K., & Thommandram, A. (2014). Toward a Big Data Healthcare Analytics System: a Mathematical Modeling Perspective. *IEEE 10th World Congress on Services*, (pp. 208-15). Alaska.

Klimov, D., Shahar, Y., & Taieb-Maimon, M. (2010). Intelligent Visualization and Exploration of time-Oriented Data od Multiple Patients. *Atificial Intelligence in Medicine, 49*, 11-31.

Koksal, N., Baytan, B., Bayram, Y., & Nacarkii, E. (2002). Risk Factors for Intraventricular Haemorrhage in Very Low Birth Weight Infants. . *Indian J Pediatr*, 69 (7) : 561-564.

Krejcie, M. (1970). Determining Sample Size for Research Activities. *Educational and Psychological Measurement #30*, 607-610.

Krueger, C. A., Gyland, E. A., & Theriaque, D. W. (2008). Neonatal heart rate variability and intraventricular hemorrhage: a case study. . *Pediatric Nursing*, 34(5), 401–404. .

Kruse. (2010). Temporal Pattern Mining. *Proceedings of the Intrenational Conference on Signals and Electronic Systems* (pp. 3-8). Poland: IEEE Press.

Kruse, R., Steinbrecher, M., & Moewes, C. (2010). Data Mining Applications in th Automotive Industry. *Automotive Industry, in: Proceedings of the 4th International Workshop on Reliable Engineering Computing (REC 2010),* (pp. 23-40). Singapore: Professional Activities Centre.

Lammarsch, T., Aigner, W., Bertone, A., & et al. (2014). Mind the Teime: Unleashing Temporal Aspects in Pattern Discovery. *Computers & Graphics, 38*, 38-50.

Lara, J., Moreno, G., Perez , P., & et al. (2008). Comparing Posturographic Time Series through Events Detection. *21st IEEE Internation Symposium on Computer-Based Medical Systems.*

Latham, N. (2014). Retrieved from http://www.pointk.org/resources/files/Latham_Human_Services_Systems.pdf

Laxman, S., & Sastry, P. S. (2006). A Survey of Temporal Data Mining". *S¯adhan¯a Vol. 31, Part 2*, 173–198.

Lee, V. E., Jin, R., & Agrawal, G. (2014). Frequent Pattern Mining in Data Streams. In C. C. Aggrawal, & J. Han, *Frequent Pattern Mining* (pp. 199-224). New York: Springer.

liao, S., Chu, P., & Hsiao, P. (2012). Data Mining techniques and applications - A decade review from 2000 t0 2011. *Expert Systems with Applications 39* , 11303–11311.

Lin, J., Williamon, S., Borne, K., & DeBarr, D. (2012). Lin J., Williamson S., Borne K., DeBarr D., "Pattern Recognition in Time Series". Advances in Machine Learning and Data Mining for Astronomy,. In E. b. Srivastava, *Advances in Machine Learning and Data Mining for Astronomy, Chapter 1.* https://doi.org/10.1201/b11822-36.

Lin, K., Xie, J., Hu, Y., & Kong, G. (2018, Apr 18). Application of support vector machine in predicting in-hospital mortality risk of patients with acute kidney injury in ICU]. *50(2)*, pp. 239-244. Retrieved from PMID: https://www.ncbi.nlm.nih.gov/pubmed/29643521

Liu, H., Wu., X., & Zhang, s. (2014). A new supervised feature selection method for Pattern Classification. *Computational Intelligence, Volume 30, Number 2,*, 342-361.

Magnusson, M. (2000). Discovering hidden time patterns in behaviour: T-Patterns and their detection. *Behavior Research Methods, Instruments, & Computers*, 32 (I), 93-110.

Maharaj, E., & D'Urso, P. (2011). Fuzzy clustering of time series in the frequency domain. *Inf. Sci. 181(7)*, 1187-1211.

Mai, S., Tim, T., & Rob, S. (2012). Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. *International Journal of Information and Education Technology*, Vol. 2, No. 3, June.

Marchione, E., & Johnson, S. (2013). Spatial, Temporal and Spatio-Temporal Patterns of Martime Piracy. *Journal of Research in Crime and Delinquency 50(4)*, 504-524.

Marret, S., Parain, D., Jeannot, E., Eurin, D., & Fessard, C. (1992). Positive rolandic sharp waves in the EEG of the premature newborn: a five year prospective study. *Archives of Disease in Childhood*, 948-951.

Martins, S. B., Shahar, Y., Goren-bar, D., & et al. (2008). Evaluation of an Architecture for Intelligent Query and Exploration of Time-Oriented Clinical Data. *Artificial Intelligence in Medicine, 43*, 17-34.

McGregor , C., & Stacey, M. (2007). High Frequency Distributed Data Stream event Correlation to improve Neonatal Clinical Management. *Proceedings of the 2007 Inagural International Conference in Distributed Event-Based systems.* Toronto.

McGregor, C. (2011, Jan 27). *Canada Patent No. WO2011009211 A1.*

McGregor, C. (2018). Using Constructive Research to Structure the Path to Transdisciplinary Innovation and Its Application for Precision Public Health with Big Data Analytics. *Technology Innovation Management*, pp. 7–15.

McGregor, C., & Kneale, B. (2007). Simulated Neonatal Intensive Care Units to Support Neonatologist International Mobility. *IASTED Int Conference Telehealth.*

McGregor, C., & Smith, K. P. (2009). A Survey of Phisiological Monitoring Data Model to Support The Ervices of Critical Care. *IEEE Internal Computer Software and Application Conference.*

McGregor, C., Catley, C., & James, A. (2011). A Process Mining Driven Framework for Clinical Guideline Improvement in Critical Care", Learning from Medical Data Streams. *13th Conference on Artificial Intelligence in Medicine (LEMEDS)*, (pp. 35-46).

McGregor, C., Catley, C., & James, A. (2012). Variability analysis with analytics applied to physiological data streams from the neonatal intensive care unit. *25th IEEE International Symposium on Computer-Based Medical Systems (CBMS).* Rome, Italy: IEEE.

McGregor, C., Catley, C., James, A., & Padbury, J. (2011). Next Generation Neonatal Health Informatics with Artemis. *Medical Informatics Europe*, (pp. 115-119). Oslo, Norway.

McGregor, C., James, A., Eklund, M., Sow, D., Ebling, M., & Blount, M. (2013). Real-time Multidimensional Temporal Analysis of Complex High Volume Physiological Data Streams in the Neonatal Intensive Care Unit. *IMIA* (pp. 362 - 366). MEDINFO 2013, doi:10.3233/978-1-61499-289-9-362.

McGregor, C., James, A., Eklund, M., & et al. (2013). Real-time MultiDimentinoal Temporal analysis of Complex High Volume Physiological Data Steams in Neonatal Intensive Care Unit. DOT:10.3233/978-1-61499-289-9-362.

McGregor, C., Kneale, B., & Tracy, M. (2007). On-Demand Virtual Neonatal Intensive-Care Units suporting Rural, Remote and urban Healthcare with Bush Babies Broadband. *Journal of Network computing Application, 30*, 1309-1313.

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (1999). *Introduction to Probability and Statistics.* Duxbury Press.

Minne, L., Ludikhuize, J., Jonge, E., & al, e. (2011). Prognostic Models for Predicting Mortality in Eldery ICU Patients: A Systemic review. *Intesinve Care in Medicine, 37*, 1258-1268.

Mohan, P., Shekhar, S., Shine, J., & Rogers, J. (2012). Cascading Spatio-Temporal Pattern Discovery. *IEEE Transactions in Knowledge and Data Engineering*, (pp. 1977-1992).

Mozar, M. (1989). A Focussed Backpropagation Algorithm for Temporal Pattern Recognition. *Complex Systems 3* , 349-381.

Murty, N., & Devi, V. (2011). *Introduction to Pattern Recognition and Machine Learning.* ISBN 978-9814335454.

Naik, T., Bressan, N., James, A., & McGregor, C. (2013). Design of temporal analysis for a novel premature infant pain profile using artemis. *Journal of Critical Care*, Issue 1 e4.

Orphanoua, K., Athena, S. A., & Keravnough, E. (2013). Temporal Abstraction and Temporal Bayesian Networks in Clinical Domain: A Survey. https://dx.doi.org/10.1016/j.artmed.2013.12.007.

Oshea, T., Allred, E. N., Kuban, K. C., Hirtz, D., & et, a. (2012). Intraventricular Hemorrhage and Developmental Outcomes at 24 Months of Age in Extremely Preterm Infants. *Journal of Child Neurology*, 22-29.

Pekalska, E., & Duin, R. (2001). Automatic pattern recognition by similarity representations. *Electronics Letters. Vol 37. No. 3*.

Peng, Y., Wang, G., Kou, G., & Shi, Y. (2011). An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing*, 2906–2915.

Portela , F., Santos, M. F., Silva, A., & et al. (2013). Pervasive Ensemble Data Mining Models to Predict Organ Failure and Patient Outcome in Intensive Medicine. *IC3K*, 410-425.

Povinelli, R. (1999). *Time Series Data Mining: Identifying temporal patterns for characterization and prediction of time series events.* PHD Dissertation.

Pugh, E., Thommandram, A., Ng, E., & McGregor, C. (2013). Classifying neonatal spells using real-time temporal analysis of physiological data. *Journal of Critical Care, Volume 28, Issue 1*.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, vol. 77*, 257-286.

Rajaraman, A., Ullman, J., & Leskovee, J. (2010). *Mining of Massive Datasets.*

R-Benchmark. (2017). Retrieved from https://www.r-bloggers.com/5-ways-to-measure-running-time-of-r-code/

Reza, S., Tanvi, B., & William, R. (2019). *Early hospital mortality prediction using vital signals.* Retrieved from https://arxiv.org/pdf/1803.06589.pdf

Robertson, J. D. (2006). Prevention of intraventricular haemorrhage: A role for recombinant activated factor VII. *Journal of Paediatrics and Child Health*, 325–331. doi:10.1111/j.1440-1754.2006.00867.x.

Robison, S. (2012). The Storage and Transfer Challenges of Bog Data. *MIT Sloan Management Review; Cambridge Vol. 53, Iss. 4*, 1-4.

Sacci, L., Larrizza, C., Combi, C., & Bellazzi, R. (2007). Data Mining with Temporal Abstractions: Learning rules from Time Series. *Data mining , 15*, 217-247.

Sadleir, R., & Tang, T. (2009). Electrode Configurations for Detection of Intraventricular Hemorrhage in the Premature Neonate. Physiological Measurement. *Physiological Measurement*, 30: 63-79.

Saeed, M., Lieu, C., Raber, G., & Mark, R. (2002). MIMIC II: a massive temporal ICU patient database to support. *Computers in Cardiology. IEEE*, 641–644.

Sarah, A., Pauwels, E., Tavenard, R., & Gevers, T. (2010). T-Patterns Revisited: Mining for Temporal Patterns in Sensor Data. *Sensors*, 7496-7513; doi:10.3390/s100807496.

Schneider, R. (2011). *Survey of Peaks/Valley indentificatio in Time Series.* Retrieved from http://www.ifi.uzh.ch/dbtg/teaching/thesesarch/ReportRSchneider.pdf

Sedgewick, R., & Flajolet, P. (1996). *An introduction to the Analysis of Algorithms.* Addison-Wesley.

Shabtai, A., Shahar, Y., & Elovici, Y. (2012). A Distributd Architecture for Efficient Parallelization and Comuting of Knowledge Based Temporal Abstractions. *J. Inteel Inf Syst. DOI 10.1007/s10844-011-0190-3*, 249-286.

Shahar, Y. (1997). Framework for Knowledge-Based Temporal Abstraction. *Arificial Intelligence in Medicine*, 79.

Shearer, C. (2000). The CRISP-DM Model: the new blueprint for data mining. *J Data Warehousing*, 13-22.

Shmulevich, & Povel, D. (2000). Measures of Temporal Pattern Complexity". *Journal of New Music Research, 29*, 61-69. 0929-8215/00/2901-061.

Silven, A. S., Dojat, M., & Garbay, C. (2005). Multi-level temporal abstraction for medical scenraio construction. *International Journal on Adaptive Control, 19*, 377-94.

Singh, A., Tamminedi, T., Yosiphon, G., Ganguli, A., & Yadegar, J. (2010). Hidden Markov Models for modeling blood pressure data to predict acute hypotension. *IEEE International Conference on Acoustics, Speech and Signal Processing.*

Singh, H. R., Garekar, S., Epstein, M. L., & L'Ecuyer, T. (2005, July). *Supraventricular tachycardia (SVT) in neonates.* Retrieved from Victoria Agency for Health Information: https://pdfs.semanticscholar.org/019f/0a3581984ed4ba268d1dc5a6ebaacaea9a8c.pdf

Spokoiny, A., & Shahar, Y. (2007). An active database architecture for knowledge-based incremental abstraction of complex concepts from continuously arriving time-oriented raw data. *Journal of Intelligent Information Systems, 31*, 1-33.

Spokoiny, A., & Shahar, Y. (2008). Incremental Application of Knowledge to Continously Arrtiving Time-Oriented Data. *Journal of Intelligent Information Systems, 31*, 1-33.

Stacey, M., & McGregor, C. (2007). Temporal Abstration in Intelligent Clinical Data Analysis: A Survey. *Artificial Intelligence in Medicine, 39*, 1-34.

Stein, B. D., Charbeneau, J. T., Lee, T. A., Schumock, G. T., & et al. (2010). Hospitalization for Acute Exacerbations of Chronic Obstructive Pulmonary Disease: How You Count Matters. *COPD: Journal of Chronic Obstructive Pulmonary Disease. , 7*(DOI: 10.3109/15412555.2010.481696), 164-171.

Sternickel, K. (2002). Automatic pattern recognition in ECG Time Series. *Computer Methods and Programs in Biomedicine 68* , 109–115. PII: S0169-2607(01)00168-7.

Sucar, L. (2015). *Probabilistic graphical models: principles and applications .* ISBN 1447166981.

Sung, C.-H., & Priebe. (1988). *IEEE International Conference on Neural Networks.* San Diego.

Sung, C.-H., & Priebe, C. (1988). Temporal Pattern Recognition. *Conf on Neural Networks,* (pp. 689-696). San Diego, CA; July 24-27: Proc. IEEE Int'l.

Tang, T. (2010). *Detection of Intraventricular Hemorrhage in neonates using Electrical Impedance Tomography.*

Tanguiane, A. (1994). A Principle of Correlativity of Perception and Its Application to Music Recognition. *Music Perception: An Interdisciplinary Journal*, Vol 11, Issue 4.

Taylor, C. (2015). Respiratory rate: overloofed but vital. Volume 23, No. 3.

Thommandram, A., Eklund, J., McGregor, C., Pugh, E., & James, A. (2014). A Rule-based Temporal Analysis Method for Online Health Analytics and It's Application for Real-Time Detection of Neonatal Spells. *IEEE International Congress on Big Data.*

Thommandram, A., Pugh, J. E., Eklund, J. M., McGregor, C., & James, A. G. (2013). Classifying Neonatal Spells Using eal-Tie Temporal Analysis of Physiological Data Stream. *Algorithm Development*.

Tin, W., Milligan, D., Pennefather, P., & Hey, E. (2001). Pulse oximetry, severe retinopathy, and outcome at one year in babies less than 28 weeks gestation. . *Arch Dis Child Fetal Neonatal Ed* , 84:F106-F110 doi:10.1136/fn.84.2.F106.

Titarenko, S., Titarenko, V., Aivaliotis, G., & Palczewski, J. (2019). Fast implementation of pattern mining algorithms with time stamp uncertainties and temporal constraints. *J Big Data 6,*, 6-37.

Tory, K., Suveges, Z., Horvath, E., Bokor, E., Sallay, P., Berta, K., & et, a. (2003). Autonomic dysfunction in uremia assessed y heart rate variability. *Pediatr Nephrol*, 18:1167–1171.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory. .* New York: Springer-Verlag Inc.

Verduijn, M., Sacchi, L., Peek, N., & et al. (2007). Temporal Abstraction for Feature Extraction: A comparative Case Sudy in Prediction from Intesive Care Monitoring Data. *Artificial Intelligence in Medicine, 41*, 11-12.

Vettier, B., Amate, L., & Garbay, C. (2012). A Multi-Hypothesis Monitoring Architecture: Application to Ambulatory Physiology. *Starting AI Researchers Symposium. DOI;10.3233/978-1-61499-096-3-348.* Montpellier, France.

Vranas, K. C., Jopling, J. K., Sweeney, T. E., Ramsey, M. C., Milstein, A. S., Slatore, C. G., . . . Liu, V. X. (2017). Identifying Distinct Subgroups of Intensive Care Unit Patients: a Machine Learning Approach. *Crit Care Med*, 45(10): 1607–1615.

Wang, J., Liu, P., Shw, M., Nahavandi, S., & Kouzani , A. (2013). A Biomedical time Series Clustering based on non-Negative Sparse Coding and Probabilistic Topic Model. *Computer Methodologies and Programs in Biomedicine, 111*, 629-641.

Ward, B. W., Schiller, J., & Goodman, R. (2014). Multiple Chronic Conditions among US Adults: A 2012 Update. *Preventing Chronic Disease. DOI: http://dx.doi.org/10.5888/pcd11.130389, no. 11*.

Warrick, P., Hamilton, E., & Macieszczak, M. (2005). Neural Network Based Detection of Fetal Heart Rate Patterns. *Proceedings of International Joint Conference on Neural Networks.* Montreal, Canada.

Weissman, A., Zimmer, E. Z., Aranovitch, M., & Blazer, S. (2012). Heart rate dynamics during acute pain in newborns. *Pflugers Arch - Eur J Physiol*, 464:593–599. DOI 10.1007/s00424-012-1168-x.

White-Traut, R. C., Nelson, M. N., Silvestri, J. M., Patel, M., Berbaum, M., & Gu G-G., &. M. (2004). Developmental Patterns of Physiological Response to a Multisensory Intervention in Extremely Premature and High Risk Infants. *JOGNN Clinical Issues*, 266-275.

Y., S. (1996). Knowledge-based temporal abstraction. *Artificial Intelligene in Medicine, 8*, 267-98.

Yet, B., Perkins, Z., Fenton, N., & et al. (2014). Not Just Data: A Method for Improving Prediction with Knowledge. *Jornal of Biomedical Informatics, 48*, 28-37.

Yuan, G., Drost, N., & Mclvor, R. (2013). Respiratory Rate and Breathing Pattern. *Clinical Review Volume 10 No. 1*, 23-25.