University of California

Los Angeles

SuperAlarm: System and Methods to Predict In-Hospital Patient Deterioration and Alleviate Alarm Fatigue

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biomedical Engineering

by

Yong Bai

2016

Abstract of the Dissertation

SuperAlarm: System and Methods to Predict In-Hospital Patient Deterioration and Alleviate Alarm Fatigue

by

Yong Bai
Doctor of Philosophy in Biomedical Engineering
University of California, Los Angeles, 2016
Professor Xiao Hu, Co-Chair
Professor Kalyanam Shivkumar, Co-Chair

A diverse array of continuous, multi-parameter and alarm-equipped physiologic monitoring devices have been deployed in modern intensive care units (ICUs) and other critical care settings to detect changes in a patient's status. Alarm signals activated by the monitors are intended to alert caregivers to either abnormalities in a patient's state or device malfunctions in order to prevent adverse events, and hence improve quality of care and patient safety. The majority of patients who eventually experience adverse events such as in-hospital cardiac arrest (IHCA) frequently exhibit signs of clinical deterioration that are evidenced in symptoms and abnormalities in the physiological vital signs and laboratory test results preceding the events. Unfortunately, the signs of deterioration are often unrecognized and missed by caregivers due to the widespread and well-documented alarm fatigue problem, which is attributable to the excessive number of false and nuisance alarms generated by the physiologic monitors.

The overarching goal of the present dissertation is to predict patient deterioration, particularly code blue events and offer a potential solution for alarm fatigue problem by leveraging monitor alarms available from physiologic monitors and laboratory test

results available in the electronic heath record (EHR) system. Several studies are performed in this dissertation to achieve this goal.

First, clinicians are routinely challenged by an overwhelming number of heterogeneous raw data to make diagnosis and treat patients. In an intensive care unit setting, we hypothesize that more predictive patterns of patient deterioration exist not in streams of individual data modality but instead in multivariate data streams. To overcome the issue of data overload, we developed a data fusion framework to identify multivariate combinations of monitor alarms and laboratory test results that co-occur frequently in a time window preceding code blue events but rarely among control patients. We proposed two approaches to integrate laboratory test results with monitor alarms. We exploited the maximal frequent itemset algorithm to mine the multivariate combinations in a time window preceding code blue events. The resultant combinations were further filtered out if they also occurred sufficiently often among all control patients. Those combinations that meet the above two criteria are termed "SuperAlarm patterns".

Moreover, deploying SuperAlarm patterns to monitor patients and detecting the emerging ones can alert caregivers to the changes in the patient's status. The emerging SuperAlarm patterns are termed "SuperAlarm triggers". The consecutive SuperAlarm triggers over time form "SuperAlarm sequences". We further hypothesize that temporal patterns may exist in these SuperAlarm sequences. Therefore, we developed a sequence classifier to recognize temporal patterns in SuperAlarm sequences. The sequence classifier essentially functions as a filter of SuperAlarm triggers. In addition, we tested the hypothesis that SuperAlarm sequences may contain more predictive temporal patterns than monitor alarms sequences. We proposed a novel method to sample subsequences, and utilized the term frequency inverse document frequency (TFIDF) to represent the subsequences. We used the information gain (IG) to select the most relevant SuperAlarm patterns to the code blue events, and the weighted support vector machine (SVM) to perform classification. The results have demonstrated that sequence

classifier based on SuperAlarm sequences outperforms that based on monitor alarm sequences.

Furthermore, a large-scale, comprehensive patient dataset is required for the development and evaluation of advanced SuperAlarm algorithms. To fulfill this need, we created a new SuperAlarm study database by consolidating and aggregating a large volume of temporal physiologic and clinical data. The new SuperAlarm study database included patient demographics, admission-discharge-transfer (ADT) information, monitor alarms, laboratory test results, physiologic waveforms and vital signs that were collected from a cohort of a large amount of identified adult coded patients and control patients admitted to UCLA and UCSF Medical Centers. We designed codebooks to map and unify alarms and laboratory tests extracted from the two institutions. We also developed a software application to extract physiologic waveforms and vital signs, and save them into binary files for further analysis.

Finally, we proposed a novel representation method to convert SuperAlarm sequences into fixed-dimensional vectors, called time weighted supervised sequence representation (TWSSR). Unlike TFIDF representation method, the TWSSR is not only a supervised weighting scheme that takes into account the distribution of SuperAlarm triggers in the SuperAlarm sequences between coded patients and control patients, it also incorporates the timing information on the weight of a SuperAlarm trigger in a SuperAlarm sequence. We used the monitor alarms and laboratory test results in the established SuperAlarm study database to mine SuperAlarm patterns and further generated SuperAlarm sequences. The support vector machine based recursive feature elimination (SVM-RFE) algorithm was applied to perform classification in conjunction with the feature selection process. The results have suggested that the performance of the sequence classifier based on the TWSSR representation method is higher than that based on TFIDF method.

In summary, we have proposed the SuperAlarm framework to integrate heterogeneous EHR and patient monitoring data to develop predictive models. This framework recognizes patterns not only across different data modalities but also across the temporal dimension through sequence classification. The SuperAlarm framework by means of data fusion could provide a potential paradigm that transforms patient monitoring into a more integrated and precise system for recognizing adverse events and ensuring prompt interventions and treatments, and subsequently improve patient outcome.

The dissertation of Yong Bai is approved.

Noel G. Boyle

William Hsu

Nader Pouratian

Kalyanam Shivkumar, Committee Co-Chair

Xiao Hu, Committee Co-Chair

University of California, Los Angeles
2016

To my wife, Yunwei Xiang, my son, Haomiao Bai,
and my parents, Wenjun Bai and Qianying Xiang
for all of your countless love and support.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express my sincerest gratitude to my adviser Dr. Xiao Hu for his professional guidance and endless support on both my research and life throughout my graduate study at UCLA. I deeply appreciate his comprehensive and thorough knowledge of research topics and his collaborative initiative for high-quality works. I have tremendously benefited from his creative spirit, thought-provoking scholarship, valuable suggestion and constant encouragement, which have helped develop my research skill and motivate me to reach where I am today. Without his persistent assistance and great patience, the completion of this dissertation could not have been possible.

I would also like to extend my great indebtedness to my committee members, Dr. Kalyanam Shivkumar, Dr. Nader Pouratian, Dr. William Hsu and Dr. Noel G. Boyle for their time, efforts, helpful feedback and valuable suggestions on evaluating my work, which broaden my horizons and shape my dissertation from various perspectives. Special thanks go to Dr. Boyle for his insightful comments and constructive role as a coauthor over the years. I would also like to thank Dr. Hsu for providing me with the opportunities to enroll in his classes during my study, which have opened my eyes to the biomedical image process.

I am truly grateful to Dr. Barbara Drew, Dr. Duc Do, Dr. Patricia Harris, Dr. Richard Fidler and Dr. Michele Pelter for their clinical concerns and support on my work. It is also my honor to take Dr. Drew's ECG course at UCSF, which has provided me with the great education experience of the human heart's electrical activity and effectively helped my transit from engineering to physiology.

I am eternally thankful to my fellow colleagues, past and present, of the research teams at UCLA and UCSF. Special appreciation is offered to Dr. Robert Hamilton for always being happy to help me on various aspects of studying and living. I still remember the unforgettable moments in my first Thanksgiving Day in the U.S. because of his invitation. I would like to thank Dr. Shaozhi Wu for introducing me to this excellent group which has had a profound impact on my life and interests. I would also like to thank Dr. Quan Ding for his willingness to discuss about my works, share me with his study experiences and help me in many ways for our established friendship. I also want to thank all other labmates from whom I have learned much and received various assistance: Dr. Fabien Scalzo, Dr. Shadnaz Asgari, Dr. Rebeca Salas-Boni, Andrea Villaroman, Dr. Jorge Arroyo Palacios and Dr. Yalda Shahriari. I would also like to thank all my friends for the wonderful time we have experienced together, even though I may not be able to thank each one individually.

Last but not least, I would like to express my deep appreciation to my family for your love, kindness, inspiration and contribution. I am sincerely indebted to my wife, Yunwei Xiang, for her incredible patience, support and tolerance during my Ph.D. study. Words fail to sufficiently convey my immense gratitude to her.

Grant Agent Acknowledgments

The works in this dissertation are partially supported by R01NHLBI128679, UCSF Middle-Career scientist award, and UCSF Institute of Computational Health Sciences.

Co-author Acknowledgments

Chapter 2 is adapted from the article: **Bai Y**, Do DH, Harris PR, Schindler D, Boyle NG, Drew BJ, Hu X. "Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction". *Journal of Biomedical Informatics*, 53: 81–92, 2015.

Chapter 3 is adapted from the article: **Bai Y**, Do DH, Ding Q, Arroyo-Palacios J, Shahriari Y, Pelter M, Boyle N, Fidler R, Hu X. "Is the sequence of SuperAlarm triggers more predictive than sequence of the currently utilized patient monitor alarms?". *IEEE Transactions on Biomedical Engineering*, submitted, 2016.

Chapter 5 is adapted from the article: **Bai Y**, Do DH, Villaroman A, Fidler R, Boyle

NG, Hu X. “Prediction of patient deterioration using SuperAlarm sequence: a time weighted supervised sequence representation method”. *Physiological Measurement*, In preparation for submission, 2016.

Vita

2006	B.E. in Information Management and Information System Beijing Information Technology Institute, Beijing, China
2006–2009	Graduate Student Researcher in Computer Applied Technology University of Electronic Science and Technology of China, Chengdu
2009–2011	Visiting Researcher, Department of Neurosurgery University of California, Los Angeles
2011–2016	Graduate Student Researcher in Biomedical Engineering University of California, Los Angeles
2013	M.S. in Biomedical Engineering University of California, Los Angeles
2013–2016	Research Intern, Department of Physiological Nursing University of California, San Francisco

Publications

Bai Y, Do DH, Harris PR, Schindler D, Boyle NG, Drew BJ, Hu X. "Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction". *Journal of Biomedical Informatics*, 53: 81–92, 2015.

Bai Y, Do DH, Ding Q, Arroyo-Palacios J, Shahriari Y, Pelter M, Boyle N, Fidler R, Hu X. "Is the sequence of SuperAlarm triggers more predictive than sequence of the currently utilized patient monitor alarms?". *IEEE TBME*, submitted, 2016.

Bai Y, Do DH, Villaroman A, Fidler R, Boyle NG, Hu X. "Prediction of patient deterioration using SuperAlarm sequence: a time weighted supervised sequence representation method". *Physiological Measurement*, In preparation for submission, 2016.

Ding Q, **Bai Y**, Tinoco A, Mortara D, Do D, Boyle NG, Pelter MM, Hu X. "Developing new predictive alarms based on ECG metrics for bradyasystolic cardiac arrest". *Physiological measurement*, 36(12):2405–22, 2015.

Do DH, Hayase J, Tiecher RD, **Bai Y**, Hu X, Boyle NG. "ECG changes on continuous telemetry preceding in-hospital cardiac arrests". *Journal of electrocardiology*, 48(6):1062–8, 2015.

Drew BJ, Harris P, Zègre-Hemsey JK, Mammone T, Schindler D, Salas-Boni R, **Bai Y**, Tinoco A, Ding Q, Hu X. "Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients". *PloS one*, 9(10):e110274, 2014.

Salas-Boni R, **Bai Y**, Harris PR, Drew BJ, Hu X. "False ventricular tachycardia alarm suppression in the ICU based on the discrete wavelet transform in the ECG signal". *Journal of electrocardiology*, 47(6):775–80, 2014.

Hu X, Do D, **Bai Y**, Boyle NG. "A case-control study of non-monitored ECG metrics preceding in-hospital bradyasystolic cardiac arrest: Implication for predictive monitor alarms". *Journal of electrocardiology*, 46(6):608–15, 2013

Bai Y, Sow D, Vespa P, Hu X. "Real-time processing of continuous physiological signals in a neurocritical care unit on a stream data analytics platform". *In:Beng-Ti Ang C,(ed) Intercranial Pressure and Brain Monitoring XV, 1st Edition, Acta Neurochirurgica Suppl*, Springer International Publishing, 122(1), 2016, DOI 10.1007/978-3-319-22533-3.

Salas-Boni R, **Bai Y**, Hu X. "Cumulative time series representation for code blue prediction in the intensive care unit". *AMIA Summits on Translational Science Proceedings*, 2015:162, 2015.

Arroyo-Palacios J, Rudz M, Fidler R, Smith W, Ko N, Park S, **Bai Y**, Hu X. "Characterization of shape differences among ICP pulses predicts outcome of external ventricular drainage weaning trial". *Neurocritical Care*, 1–10, 2016.

CHAPTER 1

Introduction

Code blue designates the most life-threatening patient events in a hospital demanding immediate assessment of the patient and initiation of necessary resuscitation effort. Code blue events are rigorously documented by a hospital's quality department. There are typically three types of codes: cardiac arrest that requires defibrillation and/or chest compression to re-establish spontaneous circulation, respiratory compromise that requires immediate intubation and ventilation to supply oxygen, and medical emergencies that are less specific regarding the cause of the call and interventions. The objective of this dissertation is to develop a data fusion system to predict patient deterioration, particularly code blue events and offer a potential solution for alarm fatigue problem based on data-driven pattern recognition and machine learning algorithms.

Two fundamental studies are performed to pursue this goal: (1) identification of multivariate patterns hidden in data streams of clinical alarms available from physiologic monitors and laboratory test results available in electronic health record (EHR) systems. Such a multivariate pattern is termed "SuperAlarm pattern" in this dissertation; and (2) recognition of temporal patterns in sequences of the consecutively emerging SuperAlarm patterns (termed "SuperAlarm triggers") when deploying them to monitor patients over time. In addition, as development and validation of the SuperAlarm approachs necessitate a large-scale and comprehensive database, substantial effort has been devoted to such endeavors, leading to an additional Chapter in this dissertation that describes the consolidation and aggregation of physiological and clinical data extracted from Medical Centers at University of California, Los Angeles (UCLA) and University of California, San Francisco (UCSF). Although the focus of this thesis is on prediction of code blue events, the general framework and approaches proposed here are also applicable to prediction of other adverse events such as sepsis.

This Chapter first introduces the background and motivation of the studies in this dissertation, including the descriptions of patient monitoring, alarm fatigue and patient deterioration. The survey of related works on solutions for alarm fatigue problem and early detection of patient deterioration is then reported in Section 1.2. Finally, the organization of the dissertation are presented in Section 1.3.

1.1 Background and Motivation

Critically ill patients admitted to intensive care units (ICUs) and other critical care settings today often have complex health problems and become higher acuity than ever before. These high acuity patients frequently exhibit hemodynamic instabilities during admission, requiring intensive monitoring and appropriate, prompt therapeutic interventions. With the aid of significant advances in life-support technologies in the recent decades, considerable efforts have yielded the availability of an impressive array of continuous, multi-parameter bedside physiologic monitoring devices for the ICU patients in attempts to improve health care efficiency, quality of care and patient safety (Figure 1.1).

The bedside physiologic monitoring devices, ranging from electrocardiogram (ECG) machines, pulse oximetry devices, ventilators to monitors of blood pressure and other variables, play a pivotal role in the detection of changes in a patient's condition in real time. Serving as electronic sentinels of safety, these monitors are routinely used for displaying and collecting physiologic waveforms, and performing high frequency measurements of a myriad of patient vital sign parameters such as heart rate (HR), respiration rate (RR), systolic, diastolic and mean values for blood pressure (SysBP, DiaBP and MeanBP), and peripheral oxygen saturation (SpO_2), just to name a few.

An alarm is often activated when any of individual vital sign parameters falls outside a predefined "low" or "high" alarm limit for a few seconds, or a deviation from normal sinus cardiac rhythm is detected [1]. An alarm signal can be either audible sound or visual text message, depending on its prioritization that stratifies risk to an appropriate level of vigilance. That is, the alarm indicating immediate life-threatening event overrides less urgent one. The alarms, typically triggered independently, are intended to alert caregivers to abnormalities in a patient's status so that timely clinical decision can be made to avert complications and adverse outcomes. It has been reported that continuous monitoring as a key activity in patient surveillance process is indispensable and may compensate for suboptimal staffing levels [2].

There are growing evidences that the majority of high-acuity patients who eventually experience adverse events such as in-hospital cardiac arrest (IHCA) exhibit signs of deterioration as evidenced in symptoms and abnormalities in both vital signs and laboratory test results in hours preceding the events [2–9]. The recognition of these signs and symptoms is paramount to avoiding adverse events and improving patient safety. Continuously detecting instabilities in physiologic variables for the critically ill patients in the ICU by means of alarm-equipped monitors, in theory, provides caregivers with opportunities to immediately identify patient deterioration so as to faciliate early intervention and subsequently reduce the incidence of failure-to-rescue (FTR) [10].

However, due to the lack of guidelines in use of physiologic monitors for the patient care, caregivers tend to rely too heavily on these devices to call their attention to changes in a patient's status. Ironically, current bedside physiologic monitors would compromise patient safety as a result of their limitations such as the simplicity of threshold-based alarm generation algorithms [11]. Two significant drawbacks rooted in the currently existing bedside physiologic monitors are listed as follows, which may lead them to be inefficient to identify patient deterioration as automatic warning systems.

1. **Alarm fatigue problem.** In order to capture the greatest percentage of clinically relevant events, most if not all of traditional threshold-based monitor algorithms

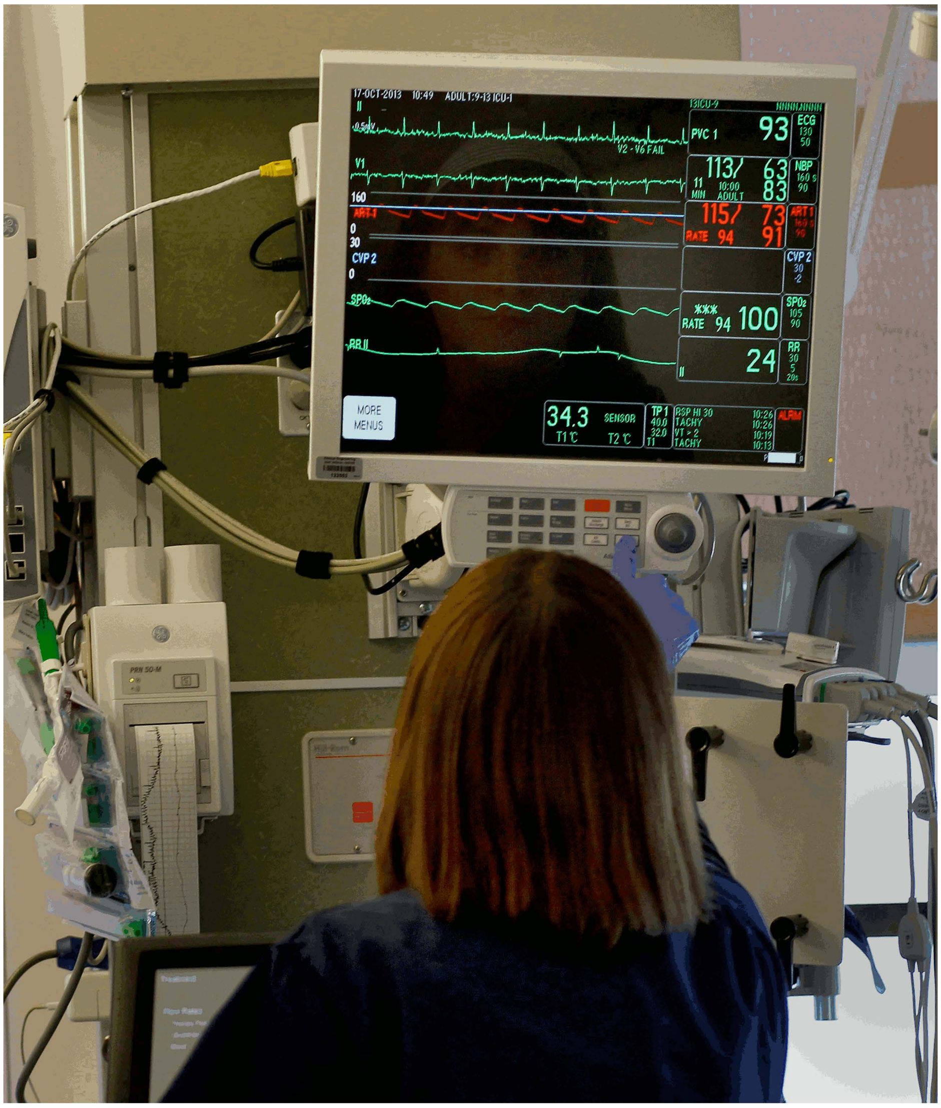


Figure 1.1: Physiologic monitoring devices in the ICU [1].

are intentionally set to be tight so as to have high sensitivity but low specificity due to the lack of a standard for alarm limit setting [11, 12]. This simple nature of alarm generation principle leads to a large volume of alarms plaguing the ICU (approximately 700 alarms [13] and 187 audible alarms [1] per patient per day). Previous studies have demonstrated that up to 99% of them are false alarms and nuisance (or false positive) alarms without clinical relevance and no need of clinical intervention (i.e., non-actionable) because of artifacts, patient's movements, cough, and so forth [14–18]. Alarm fatigue may therefore develop when caregivers are exposed to a large number of false and nuisance alarms [13, 19, 20]. The sheer number of false and nuisance alarms leads nurses to cognitive overload and becoming desensitized, distracting and interfacing with their ability to recognize changes in a patient's critical condition, and may cause them to disable, silence or ignore the critical alarms. As a consequence, prompt interventions to real critical events may be inadequate, delayed or even missed, leading to fatal outcomes [21, 22]. Although close vigilance is the reliable practice for patient monitoring in conjunction with correct operation of monitors, this method has been dramatically limited by the low nurse to patient (NTP) ratios and staffing levels. It has been reported that only about 47% of all alarms are responded by nurses [23] and the excessive false and nuisance alarms have caused unexpected alarm-related deaths in hospitals [24]. Indeed, the establishment of the "crying wolf" phenomenon [25] and the cacophony environment for both patients and caregivers are primarily attributed to excessive number of false and nuisance alarms, which have posed serious risks to patient safety. Unfortunately, this issue to date remains unsolved yet [26–28]. The Emergency Care Research Institute (ECRI) has ranked the alarm hazard as the "top 1" technology hazard from 2012 to 2015 [29–32] and "top 2" for 2016 [33].

2. **Lack of provision of relationships and patterns in the context of a patient's available data to identify patient deterioration.** Bedside physiologic monitors were originally designed to offer caregivers assistance in detecting instabilities in status of a small group of high-acuity patients by measuring physiologic parameters individually [34]. Yet this benefit has been challenged by both the increasing complexity of a patient's condition and the increasing number of more sophisticated monitoring devices in use. It has been reported that more than 36 critical physiologic patient variables have been monitored [35], and a multitude of separate, uncorrelated and stand-alone physiologic devices were deployed in acute care areas [20]. A diverse array of alarms are generated independently due to the

proliferation of these devices. However, due to the deficiencies of currently available bedside physiologic monitors in evaluating multiple parameters at a time, it is difficult and time-consuming for caregivers in the ICU to understand the relationships and underlying mechanisms of changes in a patient's condition [35]. A study has reported that caregivers have difficulties in learning more than 6 different alarm signals at a time [36]. Nurses may be overwhelmed and require longer response time when they encounter a diverse array of physiologic monitors that are attached to patients [37].

It is apparent that current patient monitoring paradigm in the ICU is inefficient in early detection of patient deterioration. Even though there are muitlfactors, the shortcomings emerging from the currently existing physiologic monitors are detrimental to recognition of signs and symptoms preceding adverse events. Studies have reported that adverse events occurred in 4% to 17% of admissions and up to 70% of such events were preventable, however, the risk of death from such an event as unexpected cardiac arrest remained at 50% to 80% due to the unrecognized patient deterioration [38]. Thereby, the development of more advanced alarm generation algorithms are required to reduce false and nuisance alarms, and alleviate alarm fatigue problem. Furthermore, with the progressively data-intensive environment in the ICU, the development of an integrated and intelligent system that allows to incorporate different types of patients data into a single or fewer indicators of physiologic instability may have potentials to reduce caregivers' reaction time and enhance quality of patient care by facilitating identification of deteriorating patients who require prompt interventions and treatments [39,40].

1.2 Related Works

From the technology perspective, numerous studies have attempted to eliminate noise and provide caregivers with the critical and essential information about a patient's physiological status so that immediate interventions and treatments are able to be offered to prevent complications and even death when patients are deteriorating. In general, this has been done separately by developing (1) advanced algorithms to reduce false alarms generated by bedside physiologic monitors; and (2) score-based systems to warn caregivers about patient deterioration.

1.2.1 False alarm reduction

Many methods have been proposed to address the alarm fatigue problem by reducing false alarms in three essential aspects [11,41]: (1) improving physiological signal extraction to filter artifacts; (2) developing advanced algorithms for alarm generation; and (3) enhancing alarm validation. An overview of some representative methods proposed recently for false alarm reduction are reported in the following paragraphs.

- **Trend-based approaches.** The trend-based approaches for improving alarm algorithms attempt to identify patterns of changes in signals for the specific physiologic variables. This can be done by exploiting qualitative representation of trends in the signals [42], which is also called qualitative trend analysis (QTA) [43] or qualitative shape analysis (QSA) [44]. Using QTA, a trend is constituted by a sequence of consecutive semi-quantitative episodes with the corresponding time interval length. The episodes are represented as basic shapes to express variation of the measured signal such as increasing, decreasing, steady, and so forth, which are called primitives. Charbonnier *et al.* [45] applied QTA with the linear segmentation technique to the SysBP, SpO_2 and maximal pressure in the airways, respectively. Consequently, 33% of total false alarms were reduced. Later, the authors further reported to filter up to 80% of false SpO_2 alarms without missing any true ones based on an adaptive method for extracting and aggregating the trends of 10 different physiologic variables: SpO_2, SysBP, DiaBP, MeanBP, HR, RR, expired volume (VE), minute ventilation (MV), maximal flow in the airways and maximal pressure in the airways [46].

- **Signal quality analysis.** As physiological signals such as ECG and ABP in the ICU are often contaminated by noise and artifacts, signal quality analysis is of paramount importance when performing signal processing and applying advanced methods such as the fuzzy logic analysis and machine learning algorithms to reduce false alarms. Zong *et al.* [47] employed the fuzzy logic approach to reduce false ABP alarms based on the signal quality indices (SQIs) derived from the ABP waveforms and then fused information from simultaneous ECG and ABP signals. The results suggested that 98.2% false alarm reduction rate was achieved, with 0.2% true alarm reduction rate. One other study has proposed to estimate HR using beat detection algorithm with a fusion method based on a Kalman filter tracking algorithm and calibrated signal quality metrics (based on statistical, temporal, spectral and cross-spectral characteristics) derived from ECG and ABP signals. The false alarms of bradycardia and tachycardia were then identified based on the estimated HR [48]. In another study false alarms were suppressed by applying support vecotor machine (SVM) to the SQIs derived from single-lead 10-second ECG segments associated with annotated labels of "good" or "bad" and types of arrhythmia alarms such as asystole, bradycardia, tachycardia, ventricular tachycardia (VT) or ventricular fibrillation/tachycardia(VFib). The authors reported that for VT, the false alarm reduction rate was 13% with 0.4% true alarm suppression [49]. Salas-Boni *et al.* [50] reported to reduce false VT alarms by extracting features from the ECG signal based on the discrete wavelet transform. The results showed that a false VT suppression of 21% with zero true alarm suppression can be achieved using MIMIC II dataset [51] while a 36% false VT suppression with zero true alarm suppression can be yielded using the UCSF alarm study dataset [1].

- **Data fusion approaches.** A data fusion approach for reducing false alarms is multivariate, which incorporates information derived from physiologic variables other than the one under observation. For example, Aboukhalil *et al.* [52] ultilized morphological and timing information derived from the ABP waveforms to reduce false arrhythmia-related alarms, including asystole, extreme bradycardia, extreme tachycardia, VT and VFib, respectively. With the alarm dataset from the MIMIC II database [51], the overall false alarm reduction rate was 59.7%, with the highest rate of 93.5% for asystole while the lowest rate of 33.0% for VT. True alarm suppression rates were zero except for VT (9.4%). Li *et al.* [53] proposed to suppress false arrhythmia-related alarms by applying the relevance vector machine (RVM) classifier on the features extracted from ECG, ABP and photoplethysmograph (PPG) signals (including HR, SpO_2, SQIs and rates of changes in parameters). The results have shown that false alarm suppression rate with no true alarm suppression were 86.4% for asystole, 100% for extreme bradycardia, 27.8% for extreme tachycardia and 19.7% for the ventricular tachycardia alarms. Borges *et al.* [54] proposed to reduce false HR alarms based on the fused information about the heart rate variability, the heart rate difference between sensors and the spectral analysis of low and high noise of each sensor extracted from from ECG, ABP and PPG. The results demonstrated that 92.5% false alarm reduction rate can be reached by applying neural network algorithm on the fused data.

1.2.2 Patient deterioration detection

Over the past decade, a variety of score-based systems, also known as “track and trigger” systems (TTSs) have been developed to facilitate early detection of patient deterioration and prediction of adverse events so as to ensure immediate and appropriate interventions. The score of a TTS is an aggregate measure evaluated in a weighted manner, typically based on several routine physiologic variables such as HR, RR, SysBP and so on. The weight of each physiologic variable is determined based on the knowledge and opinion of domain experts. A TTS warns caregivers to prompt interventions and treatments when the score exceeds the predefined criteria. Such TTSs have been exhaustively surveyed in [55] and [56]. Here, a brief review of some TTSs that have been applied in hospitals is given as follows.

- **EWS and MEWS.** The early warning scoring (EWS) system [57] is a simple tool commonly implemented at general ward level based on five physiologic variables: HR, RR, SysBP, temperature and a measure of level of a patient's consciousness (i.e., alert-verbal-painful-unresponsive (AVPU) score). The total score of EWS is a sum of all individual scores of the five parameters, each of which is assigned a 0–3 point according to its measured value (the greater the deviation from the normal range is, the larger the individual score is assigned). In 1999, Stenhouse *et al.* [58] presented a modified version of the EWS system, known as modified early warning scoring (MEWS) that was first introduced in surgical patients. In addition to the physiologic variables as used in EWS, MEWS includes urine output. Although MEWS has benefits to identify patient deterioration [59, 60], studies have reported its limitations such as inadequacy of the involved vital signs [61] and lack of standard guideline for response to the abnormal scores [62].

- **ViEWS.** The VitalPAC early warning score (ViEWS) [63] is another aggregate-weighted scoring system for prediction of mortality. The score of ViEWS is calculated by fusing multiple physiologic variables of HR, RR, SysBP, temperature, $\mathrm{SpO_2}$, fractional inspired oxygen concentration ($\mathrm{F}_i\mathrm{O}_2$) and the AVPU score. The weights for abnormalities of the measured physiologic variables used in the ViEWS are adjusted based on the clinical expert's experience and knowledge. A study has demonstrated that ViEWS outperforms other score-based systems in terms of predicting outcomes of death with 24-hour of the measured physiologic parameter set [63].

- **BioSign.** BioSign [64,65] is an algorithm for identification of patient deterioration by generating a score, called patient status index (PSI) based on fusion of five vital signs: HR, RR, BP, temperature and $\mathrm{SpO_2}$. It uses a multivariate Gaussian probabilistic model for the distribution of these vital signs for patients without crisis events. A patient crisis event is detected when these vital signs have a small probability according to this distribution estimated from a training data set.

- **Rothman Index.** Rothman *et al.* [66] developed a system to evaluate the risk of patient deterioration by derive a composite patient acuity metric, called the Rothman Index (RI) using 27 physiologic parameters including vital signs, laboratory test results, indicators of cardiac rhythms and nursing assessments. This approach is based on empirical accumulation of relative risks of its component variables in determining patient mortality after one year discharge from the hospital.

Although the approaches and systems reviewed above to address alarm fatigue problem and detect patient deterioration in isolation may perform satisfactorily in the local environments for which they were developed, they have significant limitations. For false alarm reduction, most approaches attempt to achieve the goal through secondary analysis of physiologic signals that are related to a few individual types of the interesting alarms. In addition, because true alarms not suppressed by these approaches were inherently designed to detect abnormalities after they occurred, rather than predicting patient deterioration, they are at best able to support a reactive patient care practice rather than proactive intervention. For patient deterioration detection, most score-based systems inevitably introduce additional alarms or alerts without providing direct relief of the existing alarm fatigue problem. Moreover, the scheme of score assignment is designed empirically [55]. Furthermore, most systems such as MEWS based on vital signs alone are inadequate to detect patient deterioration appropriately [61]. Currently, the ability of a tool to identify at-risk patients have not convincingly been clarified [2,10].

Bliss *et al.* [67] has suggested that clinical patterns exist in multiple physiologic variables. For example, a tension pneumothorax may have concomitant signs of tachycardia, high heart rate, rapid breathing, low SpO_2, low blood pressure and high pressure on the ventilator. Furthermore, Chopra *et al.* [34] has pointed out that future patient monitoring systems should shift focus from individual alarms to recognizing clinical patterns by integrating all patient-linked devices. Encouraged by these observations and concepts, in this dissertation we seek to predict code blue events and offer a potential solution for alarm fatigue problem by developing a data fusion system that incorporates clinical alarms available from physiologic monitors and laboratory test results available in the electronic heath record (EHR) system.

The system in the present dissertation is proposed to identify combinations of monitor alarms and laboratory test results that co-occur high frequently in a time window preceding code blue events but rarely among control patients. These combinations are composed of such super sets of frequent, multivariate clinical events, and hence we call the combinations "SuperAlarm patterns". The key feature of the SuperAlarm system is to use the event map to represent the patient data streams, as illustrated in Figure 1.2 where an example of heterogeneous physiologic data represented as clinical event map in 12-hour window preceding a code blue event is displayed. Two major categories of physiologic data are employed in this dissertation: monitor alarms (including arrhythmia alarms and vital sign parameter alarms) and laboratory test results. Moreover, we deploy the SuperAlarm patterns to monitor patients and detect the emerging SuperAlarm patterns which we term "SuperAlarm triggers". The consecutive SuperAlarm triggers over the monitoring time construct "SuperAlarm sequences". We further develop SuperAlarm a sequence classifier for temporal patterns recognition by exploring sequence representation methods that convert SuperAlarm sequences into fixed-dimensional numeric vectors. In addition, we establish a large scale, comprehensive database to facilitate the development and evaluation of such SuperAlarm algorithms. As a potential transformative paradigm of critical care monitoring in an integrated and precise manner, the SuperAlarm system is capable of recognizing patient deterioration (e.g., code blue events) without causing alarm fatigue so as to ensure early interventions and hence improve patient monitoring by leveraging heterogeneous data streams available from in-hospital patients.

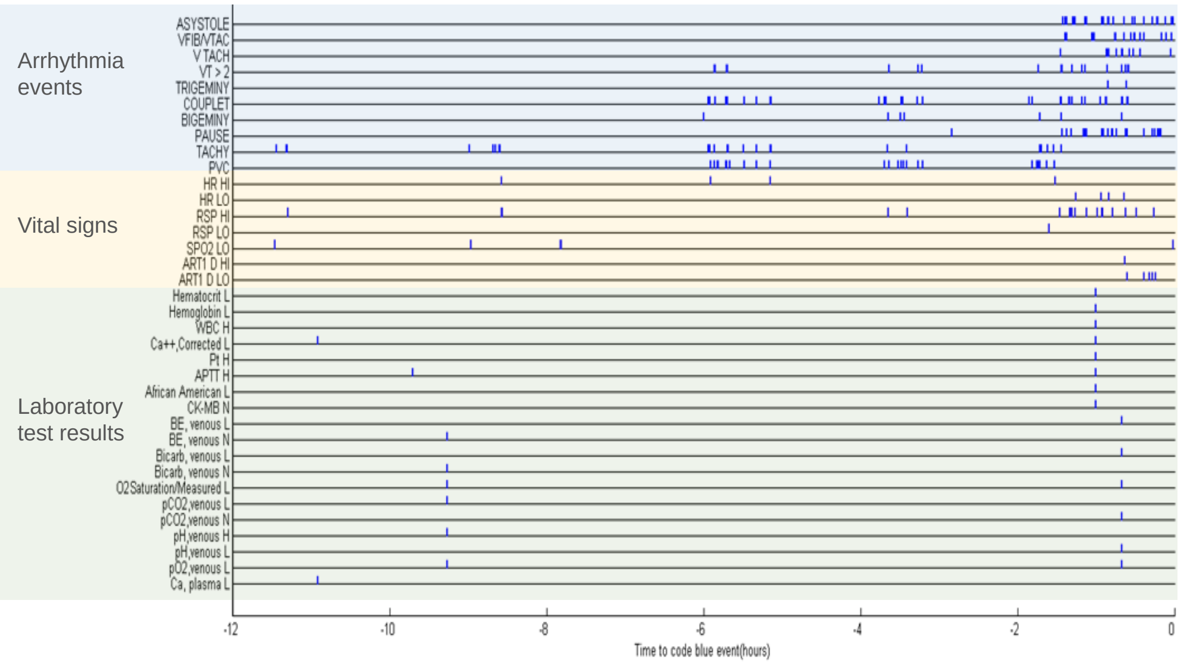


Figure 1.2: An example of heterogeneous physiologic data represented as clinical events.

1.3 Organization of the Dissertation

The present dissertation is organized as follows.

In Chapter 2, we develop a data fusion framework for identification of SuperAlarm patterns and further use the SuperAlarm patterns to predict the target endpoint — code blue events. We proposed two approaches to integrate data streams of monitor alarms and laboratory test results. We also describe the algorithm to mine SuperAlarm patterns based on the integrated dataset. We demonstrate the advances of the SuperAlarm patterns in predicting code blue events and reduce alarm frequency.

In Chapter 3, we further deploy the SuperAlarm patterns to monitor patterns and generate SuperAlarm sequences. We develop a sequence classifier for recognition of temporal patterns in SuperAlarm sequences. We also test the hypothesis that SuperAlarm sequences may contain more predictive temporal patterns than monitor alarms sequences. We demonstrate that SuperAlarm sequences can achieve higher performance in prediction of code blue events and reduction of alarm burden than that using monitor alarm sequences.

In Chapter 4, we describe a large-scale, comprehensive SuperAlarm study database that includes physiological and clinical data collected from adult coded patients and control patients admitted to the ICUs in the Medical Centers at UCLA and UCSF. We develop two naming schemes for monitor alarm and laboratory tests in order for automatic consolidation and aggregation of patient data with diverse terminologies and nomenclatures that are originated from the two institutions, respectively. We present the patient characteristics, statistical summary of monitor alarms and laboratory test results in the database. We also show examples of time series of physiological waveforms and vital signs available in the database even though physiological signal process is beyond the scope of the studies in the present dissertation.

In Chapter 5, we propose a novel representation method for conversion of SuperAlarm sequences into fixed-dimensional vectors in order to predict code blue events by recognizing temporal patterns in the sequences that are generated based on the established large-scale database. This representation method is not only a supervised weighting scheme that takes into account the distribution of sequences between coded patients and control patient, it also considers the impact of time on the weight of a SuperAlarm trigger that occurs in a SuperAlarm sequence. Classification is performed based on the proposed representation method. We demonstrate that the proposed approach can potentially assist caregivers in early predicting code blue events and reduce alarm burden, and subsequently provide a complementary tool to support clinical decision-making and enhance patient monitoring.

In Chapter 6, we draw the conclusion of the studies and summarize the research contributions in the present dissertation. We also discuss several research ideas for the future directions towards improving the patient monitoring and enhancing patient safety and quality of care based on the SuperAlarm framework proposed in this dissertation.

CHAPTER 2

A data fusion framework for identification of SuperAlarm patterns

In this Chapter, we develop a data fusion framework to identify SuperAlarm patterns in order to predict code blue events and reduce alarm frequency. The SuperAlarm patterns refer to multivariate combinations of monitor alarms and laboratory test results that co-occur high frequently in a time window preceding code blue events but rarely among control patients. In particular, we first propose two approaches to integrate patient data streams of monitor alarms and laboratory test results. Furthermore, two steps are then used to identify SuperAlarm patterns: (1) we exploit the maximal frequent itemset algorithm (MAFIA) to mine the the multivariate combinations in a T_w-long window preceding code blue events under a user-specified minimum support value min_sup; and (2) the resultant combinations are further filtered out if they also occur more than a FPR_{max} percentage of all T_w-long windows that are consecutively selected from all control patients. The remainder of combinations, termed SuperAlarm patterns, are applied on an independent test dataset to evaluate the performance in prediction of code blue events and reduction of alarm frequency.

2.1 Introduction

With technologic advances in medical devices over the past few decades, life-saving patient monitoring systems have become ubiquitous in modern hospitals [11]. Alarms annunciated by the monitoring systems are expected to alert caregivers to either changes in monitored physiological parameters of a patient or device malfunction, and to enhance quality of care and patient safety by detection of any abnormality [15].

In traditional monitor algorithms, an alarm is triggered immediately when the value of the monitored parameter exceeds or falls below the preset threshold [68]. Due to the lack of a standard for default threshold setting [69], this threshold-based algorithm is intentionally set to have high sensitivity in order to capture the greatest percentage of clinically significant events [17, 18]. As a consequence, there is low specificity and numerous alarms occur (about 700 alarms per patient per day [13]) and up to 99% of them are false alarms and nuisance (or false positive) alarms with no clinical relevance [13–17, 25]. Caregivers exposed to a large number of false and nuisance alarms become desensitized, leading to alarm fatigue problems [13,14,25]. Excessive false and nuisance alarms may compromise the quality of patient care and cause unexpected alarm-related deaths in hospitals [19]. The alarm hazard has been ranked as the "TOP 1" technology hazard for 2014 by the Emergency Care Research Institute (ECRI) [31].

Many studies have focused on addressing the alarm fatigue problem. Descriptions of many such algorithms were provided in reviews [11, 41]. For instance, Zong *et al.* [47] proposed an algorithm for reducing false arterial blood pressure (ABP) alarms by evaluating signal quality of ABP and the relationship between electrocardiogram (ECG) and ABP using fuzzy logic approach. Similarly, Aboukhalil *et al.* [52] reduced false critical ECG arrhythmia alarms using morphological and timing information derived from the ABP waveforms. Lastly, Li *et al.* [53] used a machine learning technique and data fusion method to reduce false arrhythmia alarms by combining signal quality and physiological metrics derived from the waveforms of ECG, photoplethysmograph, and optionally, ABP. Scalzo *et al.* [70,71] applied pattern recognition methods to reduce false intracranial pressure (ICP) alarms using the morphological waveform features extracted from the ICP signal. These approaches were developed to manage individual alarm types and further validation is needed to ensure that no true alarm is suppressed before their implementations by monitor vendors. Additionally, true alarms not suppressed by these approaches were designed to detect abnormalities after they occur, not to detect patient deterioration. Therefore, they are at best able to support a reactive patient care practice rather than a predictive one.

To detect patient deterioration, especially outside intensive care units, several score-based systems have been developed based on multiple parameters. The modified early warning score (MEWS) [59], for instance, was a simple tool to produce a fusion score based on the summation of an individual score assigned to each of five physiological parameters: systolic blood pressure (SysBP), respiratory rate (RR), pulse rate, temperature and patient consciousness. For each parameter, the greater the degree of deviation from the normal range, the larger the individual score assigned. However, the schema for score assignment was designed empirically [55]. Biosign [64, 65] was another algorithm to generate a patient status index (PSI) by fusing five vital signs: heart rate (HR), respiratory rate (RR), blood pressure (BP), temperature and arterial oxygen saturation (SpO_2). It used a multivariate Gaussian probabilistic model for the distribution of these vital signs for patients without crisis events. A patient crisis event was detected when these vital signs had a small probability according to this distribution estimated from a training data set. Rothman *et al.* [66] developed a system to calculate a patient acuity metric, called the Rothman Index (RI), to evaluate the risk of patient deterioration using vital signs, laboratory test results, indicators of cardiac rhythms, and nursing assessments. This approach was based on empirical accumulation of relative risks of its component variables in determining patient mortality after one year discharge from the hospital. Machine learning-based methods have also been proposed to detect patient deterioration. For instance, Clifton *et al.* [72] compared Gaussian mixture model (GMM) and support vector machine (SVM) with HR, RR, SpO_2, and SysBP as input. Tarassenko *et al.* [73] developed a centile-based early warning score system based on statistical properties of the vital signs (HR, RR, SpO_2 and SysBP) to identify deteriorating patients. Scores were determined when the statistical value of vital sign fell into certain range of centile.

It can be argued that those algorithms presented above for detection of patient deterioration introduce additional alarms or alerts without providing direct relief of the existing alarm fatigue problem. A potentially more desirable approach would incorporate patient monitor alarms and physiological signals from patient monitors. The idea to include monitor alarms as predictors of patient deterioration detection models has been tested by our group. In our previous paper [74], we proposed a novel data-driven approach using raw streaming alarm data to: 1) identify patterns that were combined with different monitor alarms using in-hospital code blue events; 2) select those patterns that occurred sufficiently often preceding code blue events but rarely in control patients; 3) empirically define and determine the optimal length of time window for the selected patterns; 4) assess the temporal characteristics of these patterns such as the sensitivity with respect to prediction window; and then 5) based on these factors, evaluate the performance of these patterns, which we called SuperAlarm patterns, under varying acceptable false positive rates. Because a SuperAlarm trigger necessarily requires simultaneous triggering of different alarms, it therefore has the potential to reduce alarm frequency.

In the present study, we follow the general framework we have previously proposed [74] and describe how we extend the conceptual domain of a SuperAlarm to incorporate laboratory test results as an additional source to compose SuperAlarm patterns. To do so, we propose several new methods so as to tackle complicating factors that arise when one incorporate non-streaming data (e.g., patients with very sparse data). We also address the need to exclude "crisis" alarms that clinicians would consider to be "no brainers" such as asystole. Specifically, we first explore a Non-Homogenous Poisson Process (NHPP) to model the occurrence rate of monitor alarms and obtain an objective threshold to exclude code blue patients with unexpectedly small number of monitor alarms preceding code blue events. We then develop two approaches to integrate laboratory test results with monitor alarms. We apply a new algorithm to discover SuperAlarm candidate patterns occurring frequently before code blue events.

These candidate patterns are composed of combinations of maximal number of monitor alarms and laboratory test results with occurrence rate greater than a support threshold. The candidate patterns are further filtered out if their false positive rates are greater than an acceptable false positive rate FPR_{max}, resulting in the final SuperAlarm patterns. By construction, these patterns are less redundant compared to those determined by the techniques of mining frequent itemsets (FI) or closed frequent itemsets (CFI) used in our previous work.

2.2 Methods

Figure 2.1 illustrates the flowchart of the proposed algorithm to discover SuperAlarm patterns. Key steps of this process are described in the following sections.

2.2.1 Data pre-processing

We follow the same pre-processing steps as used in our previous work [74]. We first unify the name of monitor alarm related to the same physiological parameter by ignoring difference in terms of monitor ports to which the sensors were attached. In addition, “crisis” alarms signaling asystole, ventricular fibrillation, and no breath are excluded. Our ultimate goal of this study is to predict code blue events; therefore, exclusion of these “crisis” alarms, which usually occur near the onset of code blue events, may avoid artificially increasing the prediction sensitivity of the SuperAlarm set.

2.2.2 Exclusion of patients with abnormally small number of monitor alarms

We found that some code blue patients had extremely small number of alarms within a T_w-long time window preceding code blue events. Given the retrospective nature of this study, it is impossible to determine the exact reasons why this occurred. However, it is highly plausible that the monitor alarms may be missed for those patients because

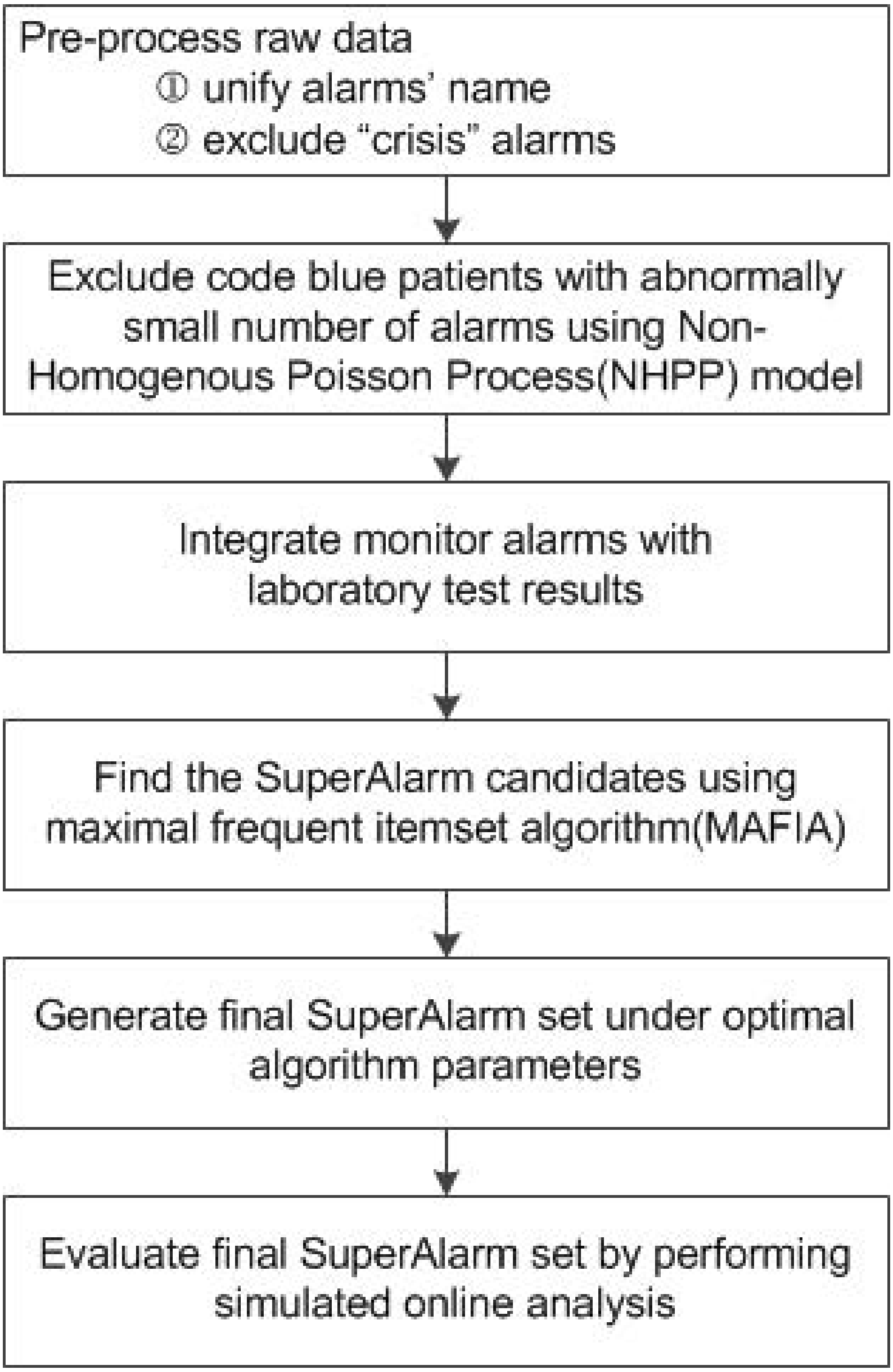

Pre-process raw data
① unify alarms' name
② exclude "crisis" alarms
Exclude code blue patients with abnormally small number of alarms using Non-Homogenous Poisson Process(NHPP) model
Integrate monitor alarms with laboratory test results
Find the SuperAlarm candidates using maximal frequent itemset algorithm(MAFIA)
Generate final SuperAlarm set under optimal algorithm parameters
Evaluate final SuperAlarm set by performing simulated online analysis

Figure 2.1: Flowchart of the proposed algorithm to discover SuperAlarm patterns.

of technical reasons including data loss from our data acquisition system or signals not registered properly by the monitors, etc. Including these patients to extract SuperAlarm patterns will provide incorrect results when determining the incidence of an alarm or alarm combinations among the code blue patients. Therefore, we exclude these patients from the study based on an objective criterion. We propose an approach to estimate the minimum number of alarms (called minimum-alarm-count-threshold) within a T_w-long time window preceding code blue events. Since monitor alarms become more frequent as time approaches the onset of code blue events [74], we assume that the arrival of monitor alarms follows a Non-Homogenous Poisson Process (NHPP) with a non-linear rate.

We denote μ_t as the rate of alarms occurring at t over time interval $(0, T]$ such that

$$\mu_t = e^{\alpha+\beta t}, 0 \leq t \leq T \tag{2.1}$$

The time interval $(0, T]$ is divided into N subintervals $\left(\frac{(k-1)T}{N}, \frac{kT}{N}\right], 1 \leq k \leq N$. Let y_k be the average number of alarms per patient over the subinterval k, we then utilize generalized linear model (GLM) to estimate the parameters α and β.

The estimated number of alarms over T_w is given by

$$\hat{n} = \int_0^{T_w} \mu_t dt = \int_0^{T_w} e^{\alpha+\beta t} dt \tag{2.2}$$

95% interval of $\hat{n}$ is (n_{lower}, n_{upper}). Thus, the minimum-alarm-count-threshold over the T_w is defined as

$$N_{minCount} = \lfloor n_{lower} \rfloor \tag{2.3}$$

where $\lfloor x \rfloor$ is the maximum integral number that is not greater than x.

We exclude those code blue patients whose number of alarms within a T_w-long time window preceding code blue events is less than the $N_{minCount}$ threshold. Regular monitor alarms from the rest of patients constitute the *Alarm* data set.

2.2.3 Integration of monitor alarms with laboratory test results

Two approaches are proposed to integrate monitor alarms with laboratory test results. Using the first approach as illustrated in panel A of Figure 2.2, we integrate the latest abnormal result of each type of laboratory tests with the array of monitor alarms within a T_w-long window. We select abnormal laboratory test results from our data set based on the associated flags reported by the electronic medical record (EMR) system. There are five flags for laboratory test results against the reference range: *HH* (extremely high), *H* (high), *L* (low), *LL* (extremely low) and *N* (normal). The abnormality flags for a given laboratory test result therefore include *HH*, *H*, *L* and *LL*. In this way, we ignore the numeric value of an abnormal laboratory test result and adopt the following representation: "[Test Name] [Abnormality]". For instance, if the laboratory test result "WBC" was flagged by *H*, then it would be represented as "WBC H". It can be seen from Figure 2.2(A) that *LA* and *LB* represent arrays of abnormal results from two different laboratory tests for a given patient. We will select LA_1 and LB_1 and integrate them with monitor alarms as they are the latest results of *LA* and *LB* with respect to T_0, respectively. Please note that we allow laboratory test results to fall outside the time window specified by T_w.

In the second approach, we use the difference between last two results of a laboratory test within a T_w-long window as a laboratory test trigger to be integrated with monitor alarms (panel B of Figure 2.2). As each laboratory test result can be indicated by one of the five flags *HH*, *H*, *L*, *LL* and *N*, there will be 25 possible triggers for a given laboratory test, which we called delta laboratory test results: $[HH \rightarrow HH$, $HH \rightarrow H$, $HH \rightarrow L$, $HH \rightarrow LL$, $HH \rightarrow N$, $\cdots$, $N \rightarrow HH$, $N \rightarrow H$, $N \rightarrow L$, $N \rightarrow LL$, $N \rightarrow N]$. For instance, if the last two results of laboratory test "Hemoglobin" were flagged by *N* and *L*, then the delta laboratory test result would be represented as "Hemoglobin $N \rightarrow L$". From Figure 2.2(B), we can see that *LA* represents an array of results from a laboratory test for a given patient. LA_1 and LA_2 will be selected and integrated with monitor alarms within T_w-long window since LA_1 and LA_2 are the two latest results for

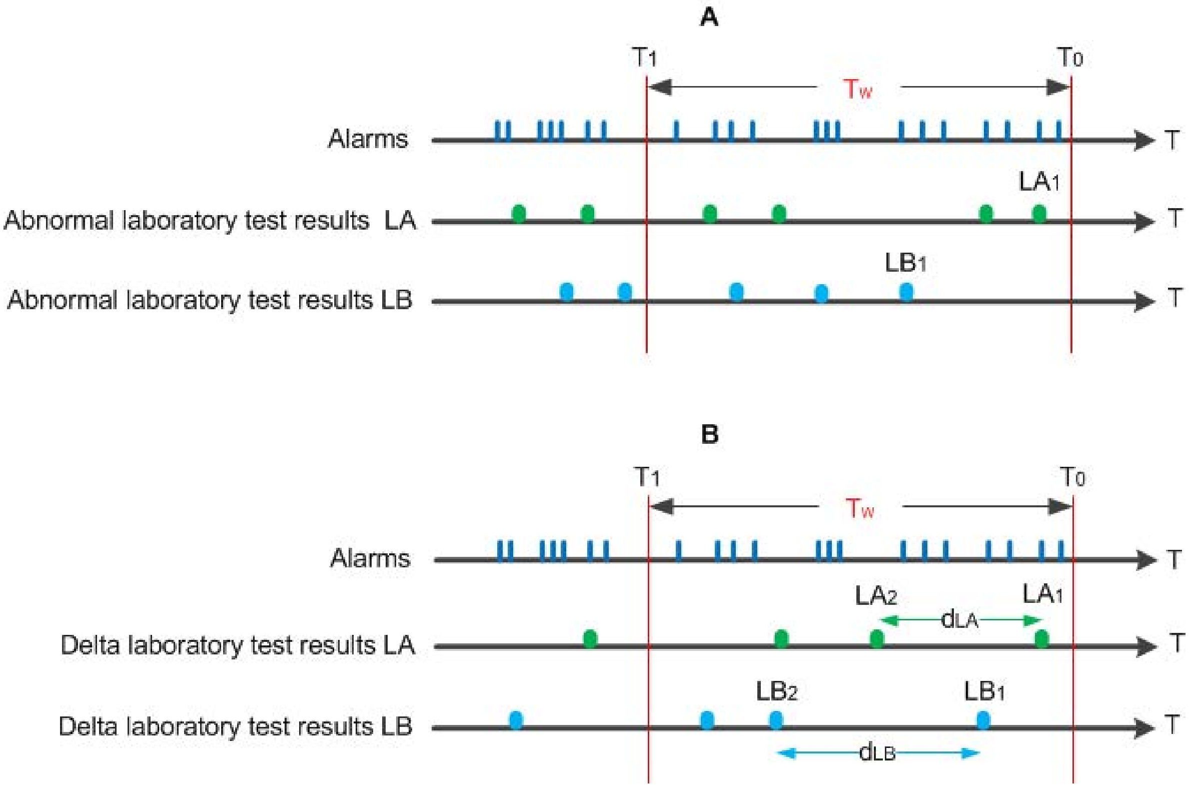

A
T1
T0
Tw
Alarms
T
LA1
Abnormal laboratory test results LA
T
LB1
Abnormal laboratory test results LB
T
B
T1
T0
Tw
Alarms
T
LA2
LA1
dLA
Delta laboratory test results LA
T
LB2
LB1
Delta laboratory test results LB
T
dLB

Figure 2.2: Two approaches to integrate monitor alarms with laboratory test results. The top horizontal axes in both (A) and (B) represent the alarms sequence while $T_w = T_0 - T_1$ is the time window. In the offline training phase, T_0 represents the onset of code blue events for code blue patients while it represents the end time point of a random T_w-long window in the consecutive 4-hour window for control patients. In the online test phase, T_0 represents the time of a new arriving monitor alarm or laboratory test result. (A) Integration of monitor alarms with the latest abnormal laboratory test results. (B) Integration of monitor alarms with the delta laboratory test results (i.e., last two laboratory test results).

laboratory test LA with respect to T_0.

Based on these two approaches, we create two extended data sets: the *Ab Lab + Alarm* data set, which is composed by the *Alarm* data set integrated with the abnormal laboratory test results, and the *Delta Lab + Alarm* data set, which consists of the *Alarm* data set integrated with the delta laboratory test results.

2.2.4 Discovery of SuperAlarm patterns

To facilitate discovery of SuperAlarm patterns, we first encode parametric monitor alarms by discretizing their numeric values using the Class-Attribute Contingency Coefficient (CACC) algorithm [75]. The CACC algorithm is a supervised discretization algorithm to generate intervals for given numeric attributes by finding the cutting points. It takes the contingency coefficient into account to measure the strength of dependence between individual attribute and classes. Therefore, the CACC algorithm allows us to utilize data from code blue patients and control patients to generate high-quality discretization schemes for parametric monitor alarms with the best correlation between these alarms and the type of patients (i.e., code blue patients and control patients). Laboratory test results do not need to be encoded since they are not represented with numeric values. The integrated data set of laboratory test results with encoded monitor alarms within T_w-long window preceding code blue events is then used to mine maximal frequent itemsets (MFI), i.e., SuperAlarm candidates.

- **Definition 1.** Support of an itemset: The support of an itemset is defined as the proportion of code blue patients in the data set who contain the itemset.

- **Definition 2.** Frequent Itemsets (FI) [76–79]: An itemset is frequent if its support is not less than a user-specified threshold of minimum support (i.e., *min_sup*).

- **Definition 3.** Maximal Frequent Itemsets (MFI) [80–83]: An itemset is maximally frequent if none of its superset is a frequent itemset. A superset of an itemset is an extension of the itemset.

It should be noted that the following relationship holds between MFI and FI: MFI $\subseteq$ FI. Classic Apriori-based methods mining FI employ a strategy of breadth-first traversal of the search space to find support information for all k-itemset ($k = 1, 2, 3, \cdots$). This method scans all $2^k - 2$ subsets of each k-itemset to determine whether or not the itemset is frequent based on the Apriori-principle, stating that the superset of any non-FI set is still a non-FI set [77]. Apriori-based method is computationally expensive when the dataset is huge or the frequent itemsets are very long [80, 81]. A different method called maximal frequent itemset algorithm (MAFIA) was proposed and it overcame this shortcoming [82].

MAFIA is a new algorithm for maximal frequent itemsets (MFI) mining using depth-first traversal on a lexicographic itemset lattice. Each node on the lattice includes *head* and *tail*. The *head* contains an itemset identifying the node while the *tail* contains frequent extensions of items lexicographically greater than any items of the head. In the process of depth-first traversal, each item in the nodes tail is determined and counted as a 1-extension. According to the Apriori-principle, the traversal process will stop if the support of $\{$*node's head*$\} \cup \{$*1-extension*$\}$ is less than a user-specified *min_sup* threshold. A candidate itemset will be added into MFI set if no superset of this candidate itemset exists in the MFI set. Three pruning strategies are applied to reduce the search space. These include: 1) parent equivalence pruning (PEP); 2) frequent head union tail pruning (FHUT); and 3) head union tail MFI (HUTMFI). MAFIA employs vertical bitmaps to represent data and uses an adaptive compression technique to enhance the performance. A vertical bitmap is a column layout to represent the patients for an itemset in the data set, and a bit in a bitmap is used to indicate whether or not the corresponding itemset appears in a given patient. For example, if patient i has itemset j, then bit i of the bitmap for itemset j is set to 1, otherwise, the bit is set to 0. Assume that bitmap(T) is a vertical bitmap for itemset T and bitmap(S) for itemset S, then the vertical bitmap for itemset $T \cup S$, bitmap($T \cup S$), is defined as bitwise-AND(bitmap(T), bitmap(S)).

In order to utilize MAFIA to mine MFI, we first build a matrix B to represent laboratory test results and encoded monitor alarms extracted within T_w-long window preceding code blue events. $B = \{x_{ij}\}$ is a $M \times N$ matrix, where M is the number of code blue patients and N is the number of encoded monitor alarms and laboratory test results ($1 \leq i \leq M, 1 \leq j \leq N$). $x_{ij} = 0$ if the i^{th} patient does not have the j^{th} alarm or laboratory test result, otherwise $x_{ij} = 1$. In other words, the j^{th} column of B represents a vertical bitmap for the j^{th} alarm or laboratory test result in the data set. The matrix B is then input into MAFIA under the user-specified *min_sup* threshold. As the process of searching goes down the lattice, the head of the node on the lattice grows longer. Due to the sparseness of bitmap especially at the lower support levels, MAFIA compresses the bitmap by removing the bit for patient P from itemset X if P does not contain X because MAFIA only needs information about the patients who contain the itemset X to count the support of the subtree rooted at node n. MAFIA employs an adaptive compression scheme to determine when to compress the bitmap. In the meanwhile, the three pruning strategies are applied to remove non-maximal sets and therefore reduce the search space. MAFIA adopts the *progressive focusing* technique to determine whether or not the extracted maximal frequent itemsets are complete. The details of MAFIA can be found in [82]. MAFIA outputs MFI which is a set of patterns consisting of maximal potential components of laboratory test results and monitor alarms.

2.2.5 Evaluation of SuperAlarm patterns

We evaluate the SuperAlarm patterns by performing both offline and simulated online analysis. Monitor alarms and laboratory test results from a randomly selected 20% of both code blue patients and control patients compose an independent test data set for the simulated online analysis. Those from the remaining 80% of both groups of patients constitute the training data set that is used to build a 10-fold cross-validation set (10-fold CV set) in the offline analysis phase. Optimal parameters of the proposed algorithm are determined based on the performance of the SuperAlarm candidates generated by MAFIA from the 10-fold CV set. The final SuperAlarm set is then generated from the whole training data set under the optimal parameters. This final SuperAlarm set is eventually employed to perform simulated online analysis.

2.2.5.1 Offline analysis to determine optimal algorithm parameters and generate the final SuperAlarm set

To find the final SuperAlarm patterns, we determine the optimal values of algorithm parameters of T_w-long time window and minimum support threshold *min_sup*. This is done by performing cross-validation analysis. According to the integration approaches mentioned in section 2.2.3, we extract monitor alarms and laboratory test results within T_w-long window preceding code blue events from the first nine folds of the 10-fold CV set. MAFIA is employed to generate SuperAlarm candidates from this extracted data set under a user-specified *min_sup* threshold. These SuperAlarm candidates are then applied to the first nine folds of the 10-fold CV set for control patients to calculate false positive rate (FPR) values for each of the SuperAlarm candidates. FPR of a SuperAlarm pattern is defined as the percentage of T_w-long windows that trigger this pattern in control patients. This is achieved by partitioning the training data set for control patients into consecutive 4-hour windows from the beginning of monitoring to the end. A T_w-long window is randomly picked within each of these 4-hour windows. Laboratory test results and monitor alarms within the T_w-long window are used to determine whether a SuperAlarm pattern is triggered, and thereby the FPR of the SuperAlarm pattern is obtained. A SuperAlarm candidate will be removed if it has FPR value greater than a given threshold.

After removing the disqualified SuperAlarm candidates, we apply the rest of SuperAlarm patterns to the remaining one fold of the 10-fold CV set to obtain a pair of values of true positive rate (TPR) and false positive rate (FPR). TPR is defined as the percentage of code blue patients who trigger at least one of SuperAlarm candidates within a T_w-long window. FPR here is calculated in terms of percentage of T_w-long windows that trigger any of the SuperAlarm patterns in control patients. Varying the threshold will lead to various pairs of TPR and FPR, and hence a receiver operation characteristic (ROC) curve can be generated. This process is repeated for each of the 10 folds, resulting in 10 ROC curves. The final ROC curve is obtained by averaging the 10 ROC curves under a given algorithm parameter combination of T_w-long window and *min_sup*.

Given an acceptable false positive rate FPR_{max}, the optimal values for the parameters of T_w and *min_sup* are determined by choosing the one with maximal TPR value across all algorithm parameter combinations while possessing FPR value less than FPR_{max}. Under the optimal algorithm parameter combination, MAFIA is applied again to the whole training data to discover the complete SuperAlarm candidates. The whole training data set is created by coalescing the 10-fold CV data set into one single set. These complete SuperAlarm candidates are further refined to generate final SuperAlarm patterns by filtering out those patterns whose FPR values are greater than FPR_{max}.

2.2.5.2 Simulated online analysis

After discovering the final SuperAlarm patterns, we employ the independent test data set to simulate the application of these SuperAlarm patterns in real-time and assess their performance at predicting code blue events. Based on the method used in [74], at the moment of receiving a new monitor alarm or a new laboratory test result, the algorithm will determine whether any of the final SuperAlarm patterns can be found among the integrated laboratory test results and monitor alarms within a T_w-long window preceding the time of this new measurement. It should be noted that T_w is the optimal length of the time window determined in the training process.

By running the simulation across the sequence of monitors alarms and laboratory test results for a given patient, we obtain a new sequence of SuperAlarm triggers. Four metrics are used to assess the performance of SuperAlarm patterns at predicting code blue events:

(1) Sen^P@T: sensitivity function with respect to prediction window. This metric is calculated in terms of percentage of code blue patients triggering any of the final SuperAlarm patterns within a prediction window preceding code blue events. This is

the same definition used in our previous work.

(2) Sen^L@T: sensitivity function with respect to lead time. This metric is computed in terms of percentage of code blue patients triggering any of the final SuperAlarm patterns within a time window that starts at 12-th hour and ends at a lead time preceding code blue event.

(3) False SuperAlarm ratio. This metric is obtained as a ratio of hourly number of the final SuperAlarm triggers for control patients to that of regular monitor alarms, or that of regular monitor alarms plus laboratory test results if the final SuperAlarm patterns contain laboratory test results.

(4) Work-up to detection ratio (WDR). We define the work-up to detection ratio as $\frac{a+b}{a}$, where a is the number of code blue patients triggering any of the final SuperAlarm patterns within a time window preceding code blue events; b is the number of control patients triggering any of the final SuperAlarm patterns within a window of the same length. The window is randomly selected over the whole monitoring time for each control patient and this process is repeated $M = 1000$ times. Let $T_{ij} = 1(1 \leq i \leq N, 1 \leq j \leq M)$ if any SuperAlarm patterns are triggered within the j^{th} selected window in the control patient i and $T_{ij} = 0$ otherwise, where N is the number of control patients in the independent test data set. We estimate the expected value of whether any SuperAlarm patterns are triggered in the control patient i as $\hat{\mu_i} = \frac{\sum_{j=1}^{M} T_{ij}}{M}$ and the standard deviation of that as $\hat{\sigma_i} = \sqrt{\frac{\sum_{j=1}^{M} (T_{ij} - \hat{\mu_i})^2}{M-1}}$. The estimated value of b and its standard deviation are then calculated as $\hat{\mu_b} = \sum_{i=1}^{N} \hat{\mu_i}$ and $\hat{\sigma_b} = \sqrt{\sum_{i=1}^{N} \hat{\sigma_i}^2}$, respectively. At this point, the expected value and the standard deviation of WDR are finally computed as $\hat{\mu}_{WDR} = 1 + \frac{\sum_{i=1}^{N} \hat{\mu_i}}{a}$ and $\hat{\sigma}_{WDR} = \frac{\sqrt{\sum_{i=1}^{N} \hat{\sigma_i}^2}}{a}$, respectively.

For a given FPR$_{max}$, we perform the McNemars test to determine whether the performances of the three SuperAlarm sets generated from *Alarm* data set, *Ab Lab* + *Alarm* data set and *Delta Lab* + *Alarm* data set are significantly different from each other using the independent test data set. To do so, we first partition the data of each control patient into consecutive 4-hour windows from the beginning of monitoring to the end. The McNemars test is then done as follows. First, we randomly select one of the 4-hour windows from each control patient. Next, the three SuperAlarm sets are compared in pairs by applying each of them to both the data of each control patient within the selected 4-hour window and the data of each code blue patient within the optimal T_w-long window preceding code blue events. Third, this process of the McNemars test is repeated 1000 times. The performances of any two SuperAlarm sets are considered to be significantly different if the number of significant individual McNemars tests is greater than 95% of the total number of tests, which is equivalent to a p-value of 0.05.

2.2.6 Patient data

The monitor alarms and laboratory test results in the present study were extracted from a central repository of comprehensive data elements archived for patients hospitalized at the UCLA Ronald Regan Medical Center, Los Angeles, California. Patients involved in this study were from ICUs (neurosurgical, cardiothoracic, coronary care, medical, transplant surgical) or other acute care areas (cardiac observation unit, hematology and stem cell transplant unit, medical-surgical specialty unit, neuroscience and stroke unit, and liver transplant unit). The Institutional Review Board waiver of consent was obtained for this secondary analysis of the data. Study subjects include all adult patients (age $>$ 18 years) admitted from March 2010 to June 2012 who experienced code blue events. Control patients were admitted within the same period without codes, death, or unplanned ICU transfer. We further refined the selection of control patients by the following criteria [74]:

- Same APR DRG(All Patient Refined Diagnosis Related Group) or Medicare DRG;
- Same age (5 year);
- Same gender;

- Admission to the same hospital unit within the same month.

254 (54% male) code blue patients with age 61.6 ± 18.2 (mean ± std) and 2213 (68% male) control patients with age 63.5 ± 14.6 were included in this study. Seventy-six percent and 19% of code blue calls were noted for cardiac arrest and respiratory arrest, respectively. Seventy-one percent of code blue patients were admitted in ICUs, 23% in non-ICUs and 6% in other facilities such as operating room and procedure room. On the other hand, 74% of the control patients were from ICUs, 24% from non-ICU units and 2% from other facilities.

2.3 Results

2.3.1 Monitor alarms for the code blue patients and control patients

Monitor alarms preceding code blue events were extracted. There were 37 case patients in our data set having more than one code blue calls and we only extracted alarms prior to the first code blue call for current analysis. 662576 raw monitor alarms for code blue patients and 5363019 for control patients were collected. The monitoring time was 250.3 ± 406.1 (mean ± std) hours and 279.9 ± 384.3 hours for the case patients and control patients, respectively. Hourly number of monitor alarms was 18.9 ± 27.9 per code blue patient and 9.5 ± 9.8 per control patient. Within a 5 minutes window preceding code blue event, the number of code blue patients having at least one "crisis" monitor alarm signaling asystole, ventricular fibrillation and no breath was 38(15.0%), 31(12.2%) and 3(1.2%), respectively.

2.3.2 Laboratory test results for case patients and control patients

We extracted laboratory test results from 19 laboratory test panels, resulting in a total of 62 different laboratory tests. Table 2.1 provides descriptive statistics of the 19 laboratory test panels.

Table 2.1: Descriptive statistics of the 19 laboratory test panels that are used in the present study.

Panel Name	Code blue patients			Control patients		
	% of results	# of patients	% of patients	% of results	# of patients	% of patients
ABG[1]	16.643	228	89.8	7.08	481	21.7
Amylase	0.022	26	10.2	0.038	109	4.9
BNP[2]	0.220	132	52.0	0.395	402	18.2
CBC[3]	22.473	254	100.0	25.127	987	44.6
Chem10	5.767	236	92.9	5.366	817	36.9
Chem7	3.909	138	54.3	9.189	714	32.3
COAG[4]	9.972	251	98.8	10.03	891	40.3
GFR EST[5]	6.732	225	88.6	6.878	623	28.2
ISL[6]	1.137	155	61.0	1.75	417	18.8
Lipase	0.017	17	6.7	0.035	102	4.6
Liver Func Test	3.095	202	79.5	6.004	658	29.7
MEDS[7]	0.056	20	7.9	0.082	61	2.8
TROPONIN	0.807	179	70.5	0.857	593	26.8
Urinalysis	1.964	149	58.7	2.634	660	29.8
CSF[8]	0.102	22	8.7	0.079	58	2.6
POC[9]	15.925	250	98.4	14.417	896	40.5
Phenobarbital	0.021	3	1.2	0.001	5	0.2
vBG[10]	5.292	177	69.7	4.852	175	7.9
Calcium	5.846	247	97.2	5.186	915	41.3

"% of results" represents percentage of the collected laboratory test results belonging to the given panel. "# of patients" represents the number of code blue patients or controls who have laboratory test results belonging to the given panel. "% of patients" represents the percentage of code blue patients or controls who have laboratory test results belonging to the given panel. 1, arterial blood gas. 2, B-type natriuretic peptide. 3, complete blood count. 4, coagulation. 5, glomerular filtration rate, estimated. 6, immunosuppressant drug level. 7, medications. 8, cerebrospinal fluid. 9, point of care. 10, venous blood gas.

There were 191483 and 362960 laboratory results for code blue and control patients, respectively. For code blue patients, 37.1% of laboratory test results were flagged as H while 34.7% as *L*, 24.9% as *N*, 2.5% as *LL* and 0.8% as *HH*. For control patients, 45.5% of laboratory test results were flagged as *H* while 41.2% as *L*, 10.7% as *N*, 1.8% as *LL* and 0.8% as *HH*. It should be noted that the majority of laboratory test results for both case patients and control patients were flagged as either H or L, indicating that abnormalities were common in the laboratory test results among these patients.

2.3.3 Results of estimating parameters in Non-Homogenous Poisson Process (NHPP) model

In order to estimate parameters α and β in the Equation 2.1, we first extract monitor alarms for all code blue patients within a 12-hour time window preceding code blue events. The 12-hour window is then divided equally into 24 consecutive subintervals, each thirty minutes long. After counting the number of alarms in each of the 24 subintervals and applying the GLM model, we obtain the estimated values of the parameters $\hat{\alpha} = 2.59 \pm 0.20$ (mean ± std, $p < 0.01$) and $\hat{\beta} = -0.08 \pm 0.03 (p < 0.01)$. To set up our experiment, four values of T_w are assessed: 30 minutes, 60 minutes, 90 minutes and 120 minutes. Using Equation 2.3 with the estimated $\hat{\alpha}$ and $\hat{\beta}$, the value of minimum-alarm-count-threshold for each T_w is determined as 8 (95% CI: 8.5 to 19.3), 15 (95% CI: 15.8 to 38.1), 22 (95% CI: 22.2 to 56.4) and 27 (95% CI: 27.7 to 74.4), respectively. Accordingly, the number of excluded code blue patients for each T_w is 34, 40, 53 and 62, respectively.

2.3.4 Offline analysis results

Three *min_sup* thresholds are specified in this study: 0.10, 0.15 and 0.20. This creates 12 algorithm parameter combinations with the four T_w values. Figure 2.3 illustrates the ROC curves for each type of the SuperAlarm sets generated by MAFIA under the various algorithm parameter combinations. We observe that for a given FPR, TPR of the SuperAlarm set generated from the integrated data set is greater than that of the SuperAlarm set from the regular monitor alarm data set.

We specify the acceptable false positive rates FPR_{max} as 0.02, 0.05, 0.10 and 0.15 in

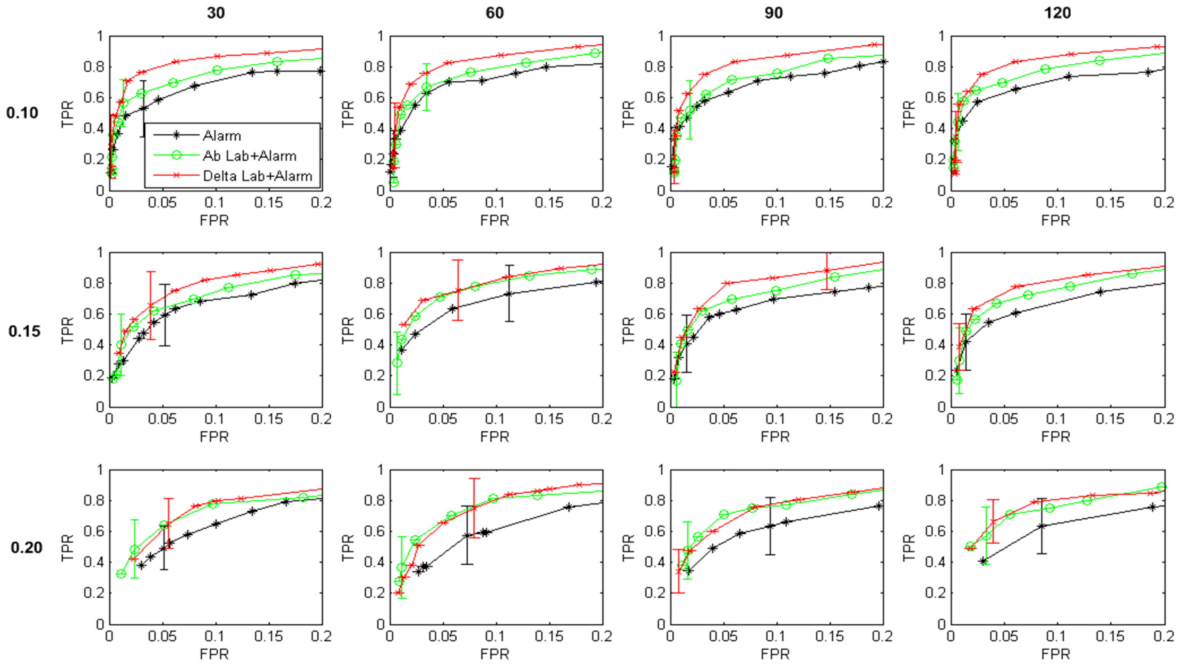


Figure 2.3: Receiver operator characteristic (ROC) curves of the three SuperAlarm sets generated under different combination of algorithm parameters in the offline training phase. The row represents *min_sup* thresholds while the column represents T_w-long time window (minutes). Since ROC curve with given *min_sup* and T_w is obtained by averaging the 10 ROC curves generated from the 10-fold CV set, we additionally mark maximum standard deviation for each of ROC curves using error bar.

the present study. Table 2.2 lists the optimal parameter combinations and the average sensitivity for each type of SuperAlarm sets based on each FPR_{max} threshold. It should be noted that this average sensitivity is calculated in the training phase based on the 10-fold CV set. We observe that for a given combination of optimal algorithm parameters, the sensitivity value of a given type of SuperAlarm set grows with increasing FPR_{max} threshold. For example, the sensitivity of SuperAlarm set generated from the *Delta Lab + Alarm* data set increases from 70.7% to 89.0% as FPR_{max} increases from 0.02 to 0.15. We also observe that the size of a SuperAlarm set, defined as the number of the SuperAlarm patterns generated from a given data set, becomes larger as FPR_{max} increases from 0.02 to 0.15 (11 to 51, 84 to 3345, and 59 to 428 for the *Alarm* data set, the *Ab Lab + Alarm* data set, and the *Delta Lab + Alarm* data set, respectively).

Table 2.2: Optimal algorithm parameters for each type of SuperAlarm sets based on varying FPR$_{max}$ thresholds.

FPR_{max}	SuperAlarm type	Optimal T_w (minutes)	Optimal *min_sup*	Sensitivity % (mean ± std)
0.02	Alarm	30	0.10	48.6 ± 15.4
	Ab lab + Alarm	120	0.10	58.0 ± 13.7
	Delta lab + Alarm	30	0.10	70.7 ± 18.1
0.05	Alarm	60	0.10	62.6 ± 14.1
	Ab lab + Alarm	60	0.15	70.6 ± 14.3
	Delta lab + alarm	30	0.10	76.2 ± 18.9
0.10	Alarm	60	0.10	70.9 ± 12.6
	Ab lab + Alarm	60	0.20	80.8 ± 15.9
	Delta lab + Alarm	90	0.15	83.3 ± 12.2
0.15	Alarm	60	0.10	79.6 ± 13.1
	Ab lab + Alarm	90	0.10	85.0 ± 9.7
	Delta lab + Alarm	30	0.10	89.0 ± 10.3

An example of the SuperAlarm pattern generated from the *Delta Lab + Alarm* data set is given as follows: SuperAlarm pattern “BRADY; APTT $H \rightarrow H$; WBC $H \rightarrow H$; Pt $H \rightarrow H$; Ca, plasma $L \rightarrow L$; Hematocrit $L \rightarrow L$; Hemoglobin $L \rightarrow L$”, which represents that if a patient was bradycardia, and the activated partial thromboplastin time(APTT), white blood cell count(WBC), prothrombin time (PT) remained high, but plasma calcium, hematocrit and hemoglobin remained low, then the patient may be at high risk.

2.3.5 Simulated online analysis results

Figure 2.4 shows the curves of Sen^P@T based on the four FPR$_{max}$ thresholds. We also plot the sensitivities of regular monitor alarms with and without “crisis” alarms, respectively. The sensitivity of regular monitor alarms with respect to prediction window is calculated in terms of percentage of code blue patients having any regular monitor alarms within the prediction window. As we expected, the sensitivity of regular monitor alarms with "crisis" alarms is greater than that without "crisis" alarms when time is near the onset of code blue events. We can observe that for a given FPR_{max}, the Sen^{P}@T value of a SuperAlarm set becomes higher as the length of prediction window is extended. We also observe that for a given length of prediction window, the Sen^{P}@T value of a SuperAlarm set increases with FPR_{max} threshold (specificity decreases).

Figure 2.5 displays the curves of Sen^{L}@T based on the four FPR_{max} thresholds. Sensitivities for regular monitor alarms with respect to lead time with and without "crisis" alarms are also shown. Here the sensitivity for regular monitor alarms is calculated as percentage of code blue patients having any of regular monitor alarms within 12-hour window prior to the lead time. We observe that no matter what the lead time is, the sensitivity for regular monitor alarms with respect to lead time is consistently 100%. It can be seen that for a given FPR_{max} threshold, the longer the lead time the lower the Sen^{L}@T value of the SuperAlarm set. We also observe that for the 2-hour lead time, the largest SuperAlarm sets obtained under the FPR_{max} threshold of 0.15 from the *Delta Lab + Alarm* data set, the *Ab Lab + Alarm* data set and the *Alarm* data set (as shown in figure 5(D)) achieve the SenL@T values of 90.0%, 83.3% and 80.0%, respectively.

In addition, Table 2.3 lists the false SuperAlarm ratio and the work-up to detection ratio for each type of SuperAlarm set based on a given FPR_{max} threshold. We also report the sensitivities with respect to different lengths of prediction window and lead time of half hour, 1 hour, 2 hours, 6 hours and 12 hours, respectively. From these results, we can see that for a given type of SuperAlarm set, a higher FPR_{max} threshold leads to a higher Sen^{P}@T and a higher Sen^{L}@T but also a larger work-up to detection ratio and a larger false SuperAlarm ratio. Taken as an example the SuperAlarm set generated from the *Delta Lab + Alarm* data set when the FPR_{max} threshold increases from 0.02 to 0.15, the Sen^{P}@T value for 1-hour prediction window and the Sen^{L}@T

Table 2.3: Performance metrics of the sensitivity with respect to prediction window (Sen^P@T), the sensitivity with respect to lead time (Sen^L@T), the false SuperAlarm ratio and the work-up to detection ratio. These metrics are calculated by applying the final SuperAlarm set to the independent test data set based on varying FPR$_{max}$ thresholds.

FPR$_{max}$	SuperAlarm type	Sensitivity (%)						False SuperAlarm ratio	Work-up to detection ratio (mean±std)	
		Metrics	Half hour	1 hour	2 hours	6 hours	12 hours	(%, mean±std)	12 hours	24 hours
0.02	Alarm	Sen^P@T	36.7	40.0	40.0	43.3	56.7	1.6±3.3	3.50.3	4.7±0.3
		Sen^L@T	40.0	36.7	30	30	33.3			
	Ab lab+Alarm	Sen^P@T	40.0	40.0	40.0	43.3	60.0	1.5±3.0	1.6±0.1	1.9±0.2
		Sen^L@T	40.0	36.7	36.7	36.7	33.3			
	Delta lab+Alarm	Sen^P@T	53.3	56.7	60.0	60.0	63.3	2.0±3.0	2.1±0.2	2.7±0.2
		Sen^L@T	46.7	43.3	36.7	40.0	40.0			
0.05	Alarm	Sen^P@T	43.3	43.3	43.3	50.0	66.7	4.2±6.5	4.4±0.3	5.7±0.3
		Sen^L@T	53.3	53.3	40.0	40.0	40.0			
	Ab lab+Alarm	Sen^P@T	70.0	70.0	70.0	73.3	80.0	3.2±4.7	2.4±0.2	3.0±0.2
		Sen^L@T	63.3	53.3	46.7	46.7	46.7			
	Delta lab+Alarm	Sen^P@T	66.7	66.7	70.0	70.0	83.3	3.6±8.1	2.8±0.2	3.7±0.2
		Sen^L@T	76.7	73.3	63.3	66.7	50.0			
0.10	Alarm	Sen^P@T	53.3	53.3	53.3	56.7	76.7	5.1±7.1	4.4±0.3	5.7±0.3
		Sen^L@T	60.0	56.7	46.7	46.7	40.0			
	Ab lab+Alarm	Sen^P@T	80.0	80.0	80.0	80.0	90.0	9.9±7.9	4.3±0.2	5.7±0.2
		Sen^L@T	76.7	73.3	70.0	70.0	60.0			
	Delta lab+Alarm	Sen^P@T	76.7	76.7	80.0	83.3	86.7	10.7±6.1	4.3±0.2	5.5±0.2
		Sen^L@T	83.3	80.0	76.7	66.7	53.3			
0.15	Alarm	Sen^P@T	66.7	70.0	70.0	76.7	90.0	14.7±10.5	7.8±0.3	10.1±0.3
		Sen^L@T	86.7	86.7	80.0	73.3	60.0			
	Ab lab+Alarm	Sen^P@T	83.3	83.3	83.3	83.3	93.3	13.0±9.4	3.9±0.2	4.8±0.2
		Sen^L@T	86.7	86.7	83.3	76.7	70.0			
	Delta lab+Alarm	Sen^P@T	93.3	93.3	93.3	93.3	93.3	14.8±9.8	6.5±0.2	8.0±0.2
		Sen^L@T	93.3	90.0	90.0	86.7	80.0			

value for 1-hour lead time increase from 56.7% to 93.3% and from 43.3% to 90.0%, respectively. However, the false SuperAlarm ratio and the work-up to detection ratio within 12-hour window also rise from 2.0% to 14.8% and from 2.1 to 6.5, respectively. We can also observe that when FPR$_{max}$ = 0.15, for instance, the Sen^L@T value of the SuperAlarm set generated from the *Delta Lab + Alarm* data set reduces from 93.3% to

80.0% with the extension of the length of lead time from half hour to 12 hours. It can be seen that for a given FPR$_{max}$ threshold and a given length of window, the Sen^{P}@T and the Sen^{L}@T of the SuperAlarm set generated from the *Delta Lab + Alarm* or the *Ab Lab + Alarm* data set are higher than that of the SuperAlarm set generated from the Alarm data set, whereas the work-up to detection ratio of SuperAlarm set generated from the *Delta Lab + Alarm* or the *Ab Lab + Alarm* data set is smaller than that of the SuperAlarm set generated from the *Alarm* data set.

Figure 2.6 shows SuperAlarm triggers within a 12-hour window preceding code blue events from the independent test data set consisting of 30 code blue patients. In each row of the plot, a white dot is placed at the time of a SuperAlarm trigger. These SuperAlarm triggers are from the largest SuperAlarm set obtained under the FPR$_{max}$ threshold of 0.15 from the *Alarm* data set (Figure 2.6(A)), the *Ab Lab + Alarm* data set (Figure 2.6(B)) and the *Delta Lab + Alarm* data (Figure 2.6(C)), respectively. It can be seen that the SuperAlarm triggers become more frequent as time approaches the onset of code blue events. We also observe that the SuperAlarm triggers generated from the *Delta Lab + Alarm* data set or the *Ab Lab + Alarm* data set are more frequent than that generated from the *Alarm* data set. These visual assessments match the quantitative results reported above.

For the FPR$_{max}$ threshold of 0.02, a large majority (954 and 983) of the 1000 repeated McNemars tests, conducted on randomly selected data, shows that the performances of SuperAlarm sets generated from the *Ab Lab + Alarm* data set and from the *Delta Lab + Alarm* data set are significantly different from that of the SuperAlarm set generated from the Alarm data set, respectively. Only 117 (11.7%) tests show that performance of SuperAlarm set generated from the *Ab Lab + Alarm* data set is significantly different from that of SuperAlarm set generated from the *Delta Lab + Alarm* data set. These McNemars tests demonstrate that the performances of the SuperAlarm sets generated from the *Ab Lab + Alarm* data set and from the *Delta Lab + Alarm* data set under the optimal algorithm parameters are significantly different from that of the

SuperAlarm set generated from the *Alarm* data set. However, the performance of the SuperAlarm set from the *Ab Lab + Alarm* data set is not significantly different from that of SuperAlarm set from the *Delta Lab + Alarm* data set. For the Alarm thresholds of 0.05, 0.10 and 0.15, we can draw the same conclusion because the numbers of the McNemars tests resulting in significantly different performances between these three types of SuperAlarm sets are [982(98.2%), 979(97.9%) and 248(24.8%)], [967(96.7%), 984(98.4%) and 304(30.4%)], and [962(96.2%), 974(97.4%) and 171(17.1%)], respectively.

2.4 Discussion

In this study, we have detailed the approaches and results from advancing a methodological framework of utilizing patient monitor alarms and laboratory test results to detect patient deterioration. Several new algorithmic elements have been introduced to the SuperAlarm framework that was created in our previous work [74]. Specifically, we excluded "crisis" alarms and code blue patients with unexpectedly small number of monitor alarms. We proposed two approaches to integrate monitor alarms with laboratory test results. The SuperAlarm patterns were discovered using MAFIA, which produced less redundant SuperAlarm patterns than those produced by Apriori-based methods used in our previous work. The results based on an independent test data set showed that SuperAlarm patterns discovered from the integrated data set of monitor alarms along with laboratory test results achieved higher sensitivity to predict code blue events and have fewer false triggers for control patients.

Patients studied here might have abnormally small number of monitor alarms for two broad reasons. It may be due to technical reasons such as data loss from the data acquisition system or due to pathophysiological reasons that sudden patient deterioration was not preceded by many alarms. If we included cases with small number of alarms due to technical reasons, the SuperAlarms sensitivity would thereby be incorrectly estimated.

Hence, we estimated the most likely minimum count of monitor alarms for code blue patients (i.e., minimum-alarm-count-threshold) and excluded those code blue patients whose count of alarms were less than the minimum-alarm-count-threshold. This practice may have excluded cases with small number of alarms due to pathophysiological reasons. Since this study was retrospective, we were not able to differentiate these two causes for a given case. Nevertheless, adopting the NHPP model to exclude patients may not impact the validity of our results for two reasons. First, it is known that ventricular fibrillation (VFib) cardiac arrest can occur suddenly. We therefore checked the number of excluded patients with VFib alarms. However, for each of the T_w-long time windows assessed in the present study, only 1 out of 34, 2 out of 40, 2 out 53 and 4 out of 62 excluded code blue patients had VFib alarms, respectively. Second, we did not exclude patients from the independent test data set and therefore the reported sensitivity may actually be an underestimate of its true value considering that patients with small number of alarms due to data loss may be included.

Compared to the SuperAlarm set consisting of only monitor alarms, the SuperAlarm set composed of monitor alarms and laboratory test results achieved higher sensitivity and lower work-up to detection ratio under an acceptable false positive rate (FPR_{max}). As we reported in section 2.3.4, both SuperAlarm sets generated from *Ab Lab + Alarm* data set and *Delta Lab + Alarm* data set yielded better performance than that generated from *Alarm* data set in terms of sensitivity to predict code blue events and the value of work-up to detection ratio. One likely explanation for this better performance might be that the laboratory test results provided more information about the patients condition. Another reason, according to Table 2.1, may be related to the fact that code blue patients on average had more laboratory tests performed, reflecting a higher clinical demand of those laboratory tests to manage patients whose clinical status were declining.

In the present study, both abnormal laboratory test results and delta laboratory test results were represented based on whether they were out of the standard reference range.

With this representation we were able to simplify the process of integrating monitor alarms because both data modalities can now be treated as discrete events. However, this representation did not take into account the numeric values of the laboratory test results. Escobar *et al.* [84] developed a model to predict non-ICU patient deterioration where a laboratory-based acute physiology score (LAPS) based on numeric values from 14 laboratory test results was used. Their study suggested that SuperAlarm might be improved by further considering ways to incorporate numeric values of laboratory test results. In addition, although 62 laboratory tests from 19 laboratory panels were integrated with monitor alarms here to build SuperAlarm set, only a subset (up to 35) of these 62 laboratory test results were part of the SuperAlarm patterns. On the other hand, there would be other laboratory tests that might be highly correlated with patient deterioration. Future work would also focus on investigation into whether different laboratory tests would improve the performance at predicting deterioration. Apart from integration of laboratory tests with monitor alarms, there is still a great volume of relevant data within an Electronic Medical Record (EMR) system that can be used to predict patient deterioration. Heldt *et al.* [85] suggested that an advanced patient monitoring system should integrate and analyze multi-dimensional clinical variables including alarms, waveforms, vital signs, laboratory tests and clinical notes to monitor the pathophysiological state of a patient. Huang *et al.* [86] reported that surveillance tools in modern hospitals may benefit from the integration of early warning scores with medications that are temporally associated with clinical deterioration to improve patient outcomes. By design, SuperAlarm is inherently a multivariate approach designed to recognize patient deterioration. Therefore, it meets the requirement of a patient-centered design of future patient alarm systems which should integrate patient data and assess clinical patterns of multiple alarms and associated vital signs holistically [34].

We employed the MAFIA algorithm to generate the SuperAlarm patterns in the present work. MAFIA was designed to discover patterns with maximal number of components that still satisfy the minimum support threshold [82]. This is a desirable characteristic because a long SuperAlarm pattern is less likely to be triggered by control patients. In our previous work [28], we extracted frequent itemsets(FI) and closed frequent itemsets(CFI) as SuperAlarm patterns, which may likely contain redundant SuperAlarm patterns leaving room for more frequent false triggers. However, the current algorithm will not recognize potentially useful patterns embedded in the occurring order of alarms and laboratory tests. Additional approaches such as Hidden Markov Model (HMM) [87], Bayesian Network [88] and String kernels [89] should be investigated as potential methodological improvement to the SuperAlarm framework.

In this study we were not able to implement other algorithms of detecting patient deterioration or compared their performance with that of SuperAlarm. This is partly due to the fact that our existing data set does not contain vital sign data or nursing notes that are needed in several existing patient deterioration detection algorithms [59, 65, 66, 84]. Nevertheless, we presented here the performance of these algorithms as reported in the original papers. In [59], the authors demonstrated that MEWS score ≥ 5 was associated with increased risk of death, ICU admission and high dependency unit (HDC) admission with odds ratio being 5.4, 10.9 and 3.3, respectively. In [65], the authors reported that the positive predictive value (PPV) of the Biosign alerts was 95%. In [66], the authors reported that RI predicted patient deterioration, 24-h mortality and 30-day readmissions with a c-statistics ≥ 0.92, ≥ 0.93 and $= 0.62$, respectively. In [84], Escobar *et al.* reported that their model predicted patient deterioration outside the ICU with a c-statistic value of 0.775 in the validation dataset and the work-up to detection ratio was 14.5 when identifying 15% of all transfers to the ICU. It is important to select appropriate performance metrics to help users evaluate patient deterioration detection systems. Conventional metrics as c-statistics undoubtedly have strong theoretical underpins. However, we argue that a patient deterioration detection system needs to be evaluated at a particular operating point on the ROC curve. At a chosen operation point, work-up to detection ratio is an excellent metric to gauge the extra work for a correct detection and can be readily communicated to clinical users.

In addition, sensitivity can help understand how many deterioration events can be potentially captured. For monitoring applications, sensitivity needs to be evaluated as a function of lead time. However, the concept of incorporating lead time in evaluating sensitivity has not been widely used. To better compare with monitor alarm frequency, the false SuperAlarm ratio is proposed. This metric has not been used in other works either. In summary, we acknowledge that a direct comparison of similar patient deterioration detection approaches needs to be done, preferably using a standard database, proper implementation, and appropriate performance metrics.

Finally, the discovered SuperAlarm patterns need further verifications by clinical knowledge. Given that these patterns were discovered from data of critically ill patients, it is very likely that some of these patterns may not add new knowledge per se but being able to track these patterns automatically in practice may alleviate the alarm fatigue problem.

2.5 Conclusion

The present study proposed novel approaches to integrate monitor alarms with laboratory test results to discover SuperAlarm patterns using maximal frequent itemsets mining technique. The performance of SuperAlarm patterns was assessed based on four metrics using an independent test data set: sensitivity with respect to prediction window, sensitivity with respect to lead time, false SuperAlarm ratio, and work-up to detection ratio. Results showed that both the SuperAlarm sets generated from *Ab lab + Alarm* data set and *Delta lab + Alarm* data set outperformed the SuperAlarm set consisting of only monitor alarms in terms of these metrics. Further performance gain may be achieved by using numeric values of laboratory test results, integrating metrics of raw physiological signals as additional "alarms", and incorporating sequential patterns of SuperAlarm triggers.

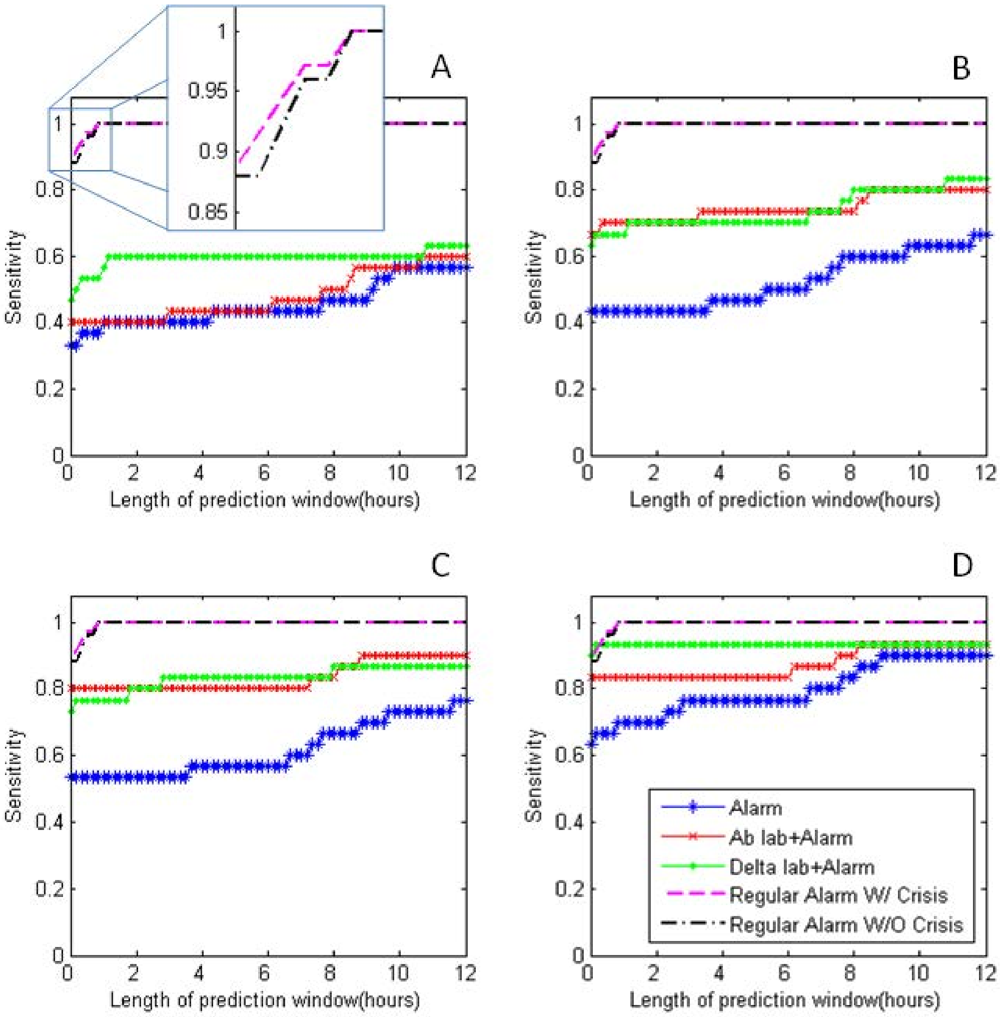


Figure 2.4: Sensitivity curves of the three final SuperAlarm sets with respect to prediction window (i.e., Sen^P@T). The sensitivity curves are obtained by applying the corresponding type of the final SuperAlarm sets to the independent test data set of *Alarm*(blue curve), *Ab Lab* + *Alarm* (red curve) and *Delta Lab* + *Alarm* (green curve) based on FPR$_{max}$=0.02 (A), FPR$_{max}$=0.05 (B), FPR$_{max}$=0.10 (C) and FPR$_{max}$=0.15 (D), respectively. The x-axis represents the length of prediction window preceding code blue events. The magenta curve and black curve represent the sensitivity of regular monitor alarms with respect to the prediction window with and without "crisis" alarms from the independent test data set, respectively.

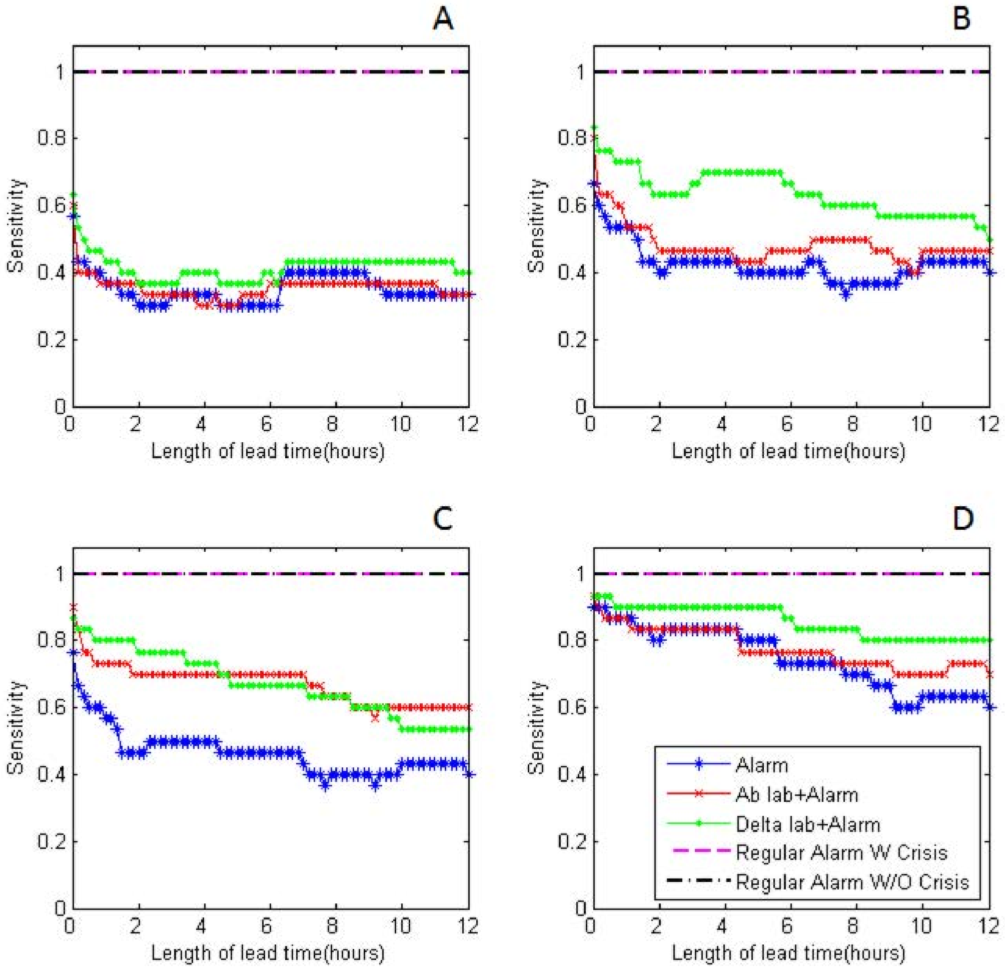


Figure 2.5: Sensitivity curves of the three final SuperAlarm sets with respect to lead time (i.e., Sen^{L}@T) based on FPR$_{max}$=0.02 (A), FPR$_{max}$=0.05 (B), FPR$_{max}$=0.10 (C) and FPR$_{max}$=0.15 (D), respectively. The x-axis represents the length of lead time preceding code blue events. The magenta curve and black curve represent the sensitivity of regular monitor alarms with respect to lead time with and without "crisis" alarms from the independent test data set, respectively.

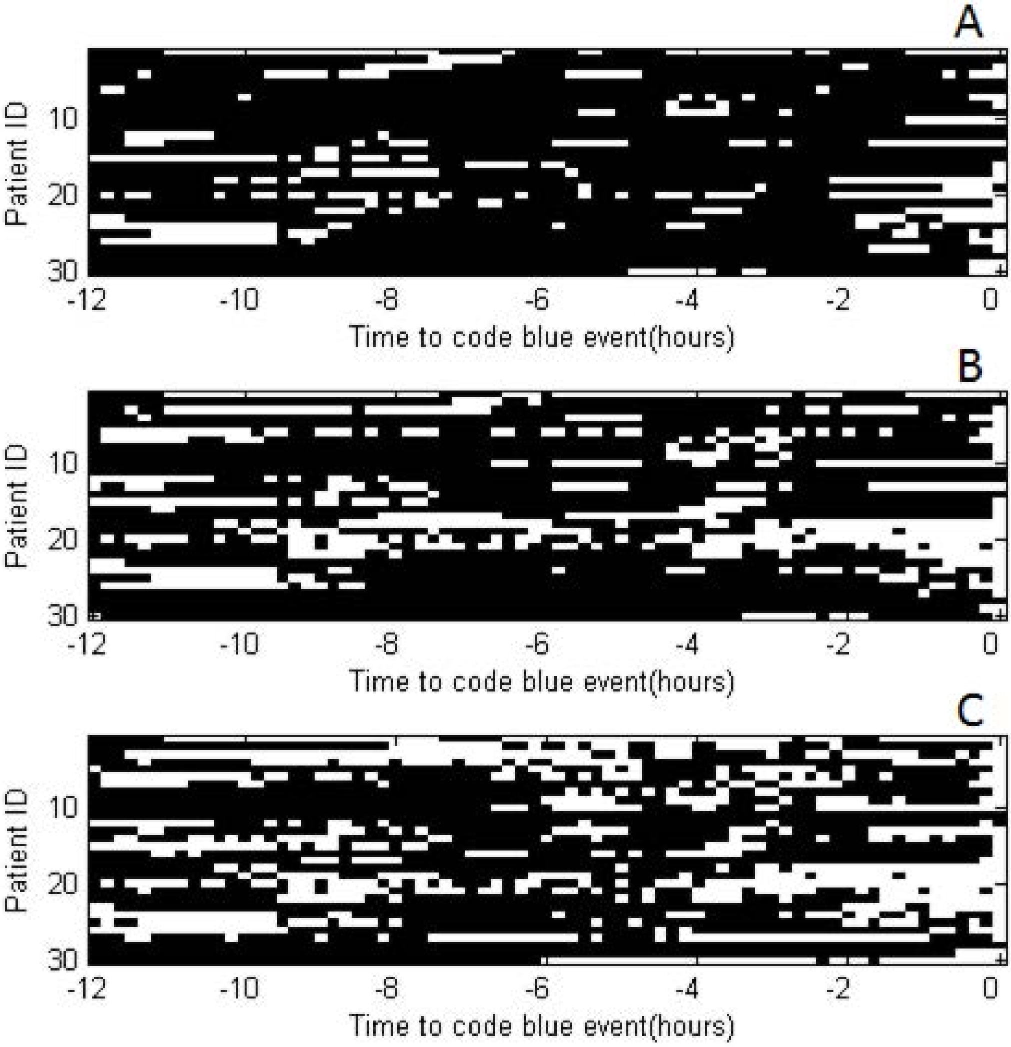


Figure 2.6: Sequences of SuperAlarm triggers within 12-hour window preceding code blue events from the independent test data set. The white point represents that at least one of the SuperAlarm patterns is triggered. Zero point on the x-axis represents the onset of code blue event. A, SuperAlarm triggers from the *Alarm* data set; B, SuperAlarm triggers from the *Ab Lab* + *Alarm* data set; C, SuperAlarm triggers from *Delta Lab* + *Alarm* data set.

CHAPTER 3

Is the sequence of SuperAlarm triggers more predictive than sequence of the currently utilized patient monitor alarms?

In the previous Chapter we developed a data fusion framework to identify SuperAlarm patterns. The SuperAlarm patterns can further be deployed to monitor patients in real time to detect the emerging ones (termed SuperAlarm triggers) that are intended to alert caregivers to the changes in the patient's status. However, it may not be favorable for caregivers to simply rely on the individual SuperAlarm triggers due to the potential redundancy (i.e., multiple disparate SuperAlarm patterns can be triggered at the same time point) and uninteresting SuperAlarm patterns(i.e., SuperAlarm patterns that are simply mined by the frequent itemset algorithm may not be clinically interpretable or relevant to code blue events). In this Chapter, we develop a sequence classifier to recognize temporal patterns in SuperAlarm sequences that are constructed by consecutive SuperAlarm triggers over time. The sequence classifier essentially functions as a filter of SuperAlarm triggers. In addition, we test the hypothesis that SuperAlarm sequences may contain more predictive temporal patterns than monitor alarms sequences. This Chapter is organized as follows. In Section 3.1 we introduce the background and objective of the study in this Chapter. We then describe the methods in Section 3.2, including sampling subsequences, utilization of the term frequency inverse document frequency (TFIDF) to represent the subsequences, use of the information gain (IG) to select the most relevant SuperAlatm patterns to the code blue events, and the weighted support vector machine (SVM) to perform classification. The results are demonstrated in Section 3.3. Finally, we present discussion and draw the conclusion on the study in this Chapter in Section 3.4 and 3.5, respectively.

3.1 Introduction

The trajectory of a patient's physiological state through hospitalization is dynamic, particularly for critically ill patients. Unfortunately, the ability to effectively and precisely detect and anticipate patient status changes using current patient monitors remains unsatisfactory as evidenced by the wide-spread alarm fatigue problems in hospitals [1,13]. A straightforward approach to handle alarm fatigue is to suppress false alarms by signal processing and machine learning approaches [47,52,53,71,90]. These approaches have shown some potentials for a few types of arrhythmia and intracranial pressure alarms, but additional research is needed to develop methods to remove false threshold-crossing parameter alarms [11,17,18]. In addition to false alarms, nuisance alarms are considered as a major contributor to alarm fatigue. Nuisance alarms reflect transient and sometimes minor deviations of monitored physiological variables but do not indicate major patient status changes and therefore are often not actionable. As a result, a trend in the community to address nuisance alarms is to adjust alarm limits to find optimal settings for these limits [19]. However, caution is necessary in excessively suppressing nuisance alarms because it is possible that certain patterns such as increasing frequency of these transient deviations of physiological variables may be the harbinger of some major events [91]. In our view, the number of alarms should not be the sole outcome for gauging the effectiveness of interventions for addressing alarm fatigue. Instead, a more comprehensive approach towards fulfilling the ultimate goal of patient monitoring needs to be taken.

In a recent position paper [34], the authors pointed out that future patient monitoring systems should shift focus from individual alarms to recognizing clinical patterns by integrating all patient-linked devices. This concept indeed supports our evolving approach [74, 92] to improve patient monitoring by identifying multivariate patterns hidden in data streams of patient monitor alarms, physiological signals, and data from electronic health record (EHR) systems. We refer to such a multivariate pattern as a SuperAlarm pattern. The term SuperAlarm was first introduced in our paper [74] to define a superset of patient monitor alarms that co-occur within a time window immediately preceding "code blue" events for more than a minimal percentage of coded patients but less than a maximal percentage of control patients without triggering any "code blue" calls. In our subsequent study [92], we further extended this approach by integrating laboratory test results from EHR system with monitor alarms to identify SuperAlarm patterns and demonstrated the improved performance in prediction of "code blue" events. As a consequence of this extension, a SuperAlarm pattern as referred to in the present work is a superset of co-occurring monitor alarms and laboratory test results. With a training dataset consisting of data from both coded and control patients, a set of SuperAlarm patterns can be identified. These patterns can then be deployed to monitor patients, and each detection of an emerging SuperAlarm pattern is termed a SuperAlarm trigger. A sequence of consecutive triggers is termed SuperAlarm sequence. As a next step to expand this SuperAlarm approach, we recently developed a sequence representation algorithm that uses fixed-dimensional vectors to represent SuperAlarm sequences that can have different number of triggers [93]. By exploiting a vectorization method for representing SuperAlarm sequences, there is the opportunity to use off-the-shelf machine learning approaches to recognize temporal patterns encoded by these sequences. However, it should be realized that various sequence representation methods exist and they can also be applicable to sequences of just monitor alarms directly. Therefore, an interesting question arises regarding whether it is beneficial to first identify SuperAlarm patterns and construct SuperAlarm sequence versus directly utilizing monitor alarm sequence.

The central objective of this work is to provide an answer for the above question by investigating three types of sequences: 1) sequences of raw monitor alarms; 2) sequences of modified monitor alarms where vital sign parameter alarms are preprocessed by discretizing their numeric values, e.g., systolic blood pressure alarms “systolic arterial blood pressure $>$ 135 mmHg” and “systolic arterial blood pressure $>$ 200 mmHg” will be treated as a different alarms if the values 135 and 200 are discretized into different bins; and 3) sequences of SuperAlarm triggers. The second sequence type is included because discretization of vital sign parameter alarms was also used as a preprocess step when identifying SuperAlarm patterns. To fairly compare these three types of sequences, we use the same sequence representation and machine learning algorithm. In particular, we use a sequence representation technique in document classification – term frequency inverse document frequency (TFIDF) [94] to convert sequences into fixed-dimensional vectors. Regarding the machine learning approach, we use the information gain (IG) technique [95] to conduct feature selection [96] and apply a modified support vector machine (SVM) called weighted SVM [97] as the classifier that incorporates different misclassification costs into the objective function to handle the imbalance training dataset.

3.2 Materials and Methods

3.2.1 Overview of a classification approach for sequences

Figure 3.1 illustrates the proposed algorithm to predict a clinical endpoint, e.g., “code blue” event, using a sequence of triggers. In this figure, we use SuperAlarm sequence as an example. The algorithm consists of three steps:

- Step 1, generation of SuperAlarm sequence. As shown in Figure 3.1 when an alarm “ABP Dia LO $<$ 45 mmHg” occurs at current time t_i, the algorithm first extracts all raw alarms and laboratory test results in a T_w-long time window (orange rectangle) preceding t_i. If any subset of these alarms and laboratory test

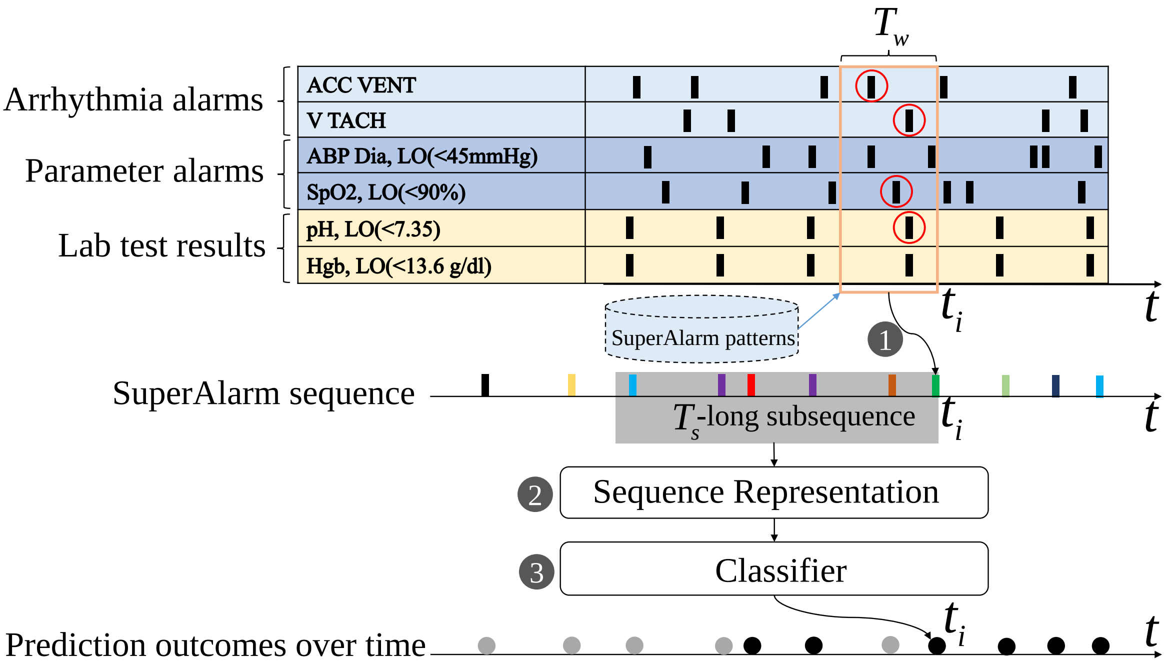


Figure 3.1: A graphic illustration of the proposed approach to predict "code blue" event in use of SuperAlarm sequences. This example illustrates 6 physiological variables out of the 3 groups occurring over time: arrhythmia alarms, parametric alarms, and laboratory test results. ACC VENT: accelerated ventricular; V TACH: ventricular tachycardia; ABP Dia, LO: diastole arterial blood pressure low; SpO2, LO: peripheral capillary oxygen saturation low; pH, LO: pH value low; Hgb, LO: hemoglobin low.

results matches a SuperAlarm pattern, a SuperAlarm trigger then occurs at t_i. By repeating this process whenever a new alarm or a new laboratory test result is received, a sequence of SuperAlarm triggers will be generated and they are depicted as vertical bars in different colors in Figure 3.1.

- Step 2, representation of SuperAlarm sequence. Assume at time t_i, we would assess the risk of impending "code blue" event by using all SuperAlarm triggers that are within a T_s-long window preceding t_i. A sequence representation approach is then used to convert this subsequence of SuperAlarm triggers into a fixed-dimensional vector so that it can be used as an input feature to a classifier.

We use TFIDF method in this work and investigate the effect of different choices of T_s.

- Step 3, classification. In this step, the feature vector from step 2 will be subjected to a feature selection process using the IG technique and then classified by an SVM model. This process will be repeated at every single t_i where there is at least one SuperAlarm trigger. Based on SVM output, some SuperAlarm triggers will be classified as negative — i.e., not associating with the clinical endpoint (depicted as gray dots) and others will be classified as positive (black dots). This classifier essentially functions as a filter of SuperAlarm triggers.

3.2.2 Monitor alarms, laboratory test results and superAlarm patterns

The present work uses the same set of SuperAlarm patterns that were identified in our previous study. Therefore, we provide a brief introduction of monitor alarms, laboratory test results, the process to identify SuperAlarm patterns as used in that study [92].

Monitor alarms were extracted from a central repository where data from patient monitors were continually archived by BedMasterEx system (Excel Medical Electronics, Inc, Jupiter, FL). A total of 100 distinct monitor alarms were used in our previous study including 14 ECG arrhythmia alarms and 86 parameter alarms that signal the deviation of vital signs outside preset upper or lower thresholds. These vital signs include heart rate, respiratory rate, pulse oximetry, systolic, diastolic, and mean arterial blood pressure, just to name a few. Crisis alarms including asystole, ventricular fibrillation (VFib) and no breath were excluded because our clinical endpoint — "code blue" is typically triggered when true crisis alarms occur. We also exclude technical alarms, e.g., "ECG LEADS FAIL" as they do not represent patient status. We further discretized the values of vital signs that triggered the corresponding parameter alarms using the algorithm described in [75]. After discretization, a total of 362 distinct types of discretized parameter alarms were obtained. Each parameter alarm can be uniquely mapped to a discretized parameter alarm.

From the EHR, we extracted laboratory test results based on a total of 62 conventional laboratory tests (e.g., arterial blood gas, complete blood count, blood chemistry). Numeric values of the laboratory test results were not utilized. Instead, we employed the abnormality flag reported for each laboratory test result by the EHR system. There are five flags available for each laboratory test indicating the deviation of the result from the reference range: HH (critically high), H (high), N (normal), L (low) and LL (critically low). In our previous work, we tested two approaches of encoding a laboratory test result as an equivalent lab "alarm". The present work uses the delta lab approach where we encode the pair of abnormality flags of the two most recent consecutive laboratory test results, e.g., the last two potassium test results are encoded as "Potassium N $\rightarrow$ L", which means that serum potassium level changed from normal to below normal.

A SuperAlarm pattern is identified following two steps. In the first step, we used the MAFIA frequent itemset mining algorithm [82] to identify combinations of monitor alarms and lab "alarms" that co-occurred in a T_w-hour long time window preceding a "code blue" event for more than *min_sup* percentage of coded patients. Those candidate patterns were then removed if they also occurred for more than FPR_{max} percentage of all T_w-hour windows that were consecutively selected from all control patients. Understandably, parameters including T_w, *min_sup*, and FPR_{max} control the number of final SuperAlarm patterns and the performance of the set of SuperAlarm patterns. In the present work, we use the SuperAlarm patterns that were identified based on the set of algorithm parameters that achieved highest sensitivity, which is the upper limit for the sequence classifier approach to achieve. The values for the three algorithm parameters are: $T_w = 0.5$, $min\_sup = 5\%$, and $\text{FPR}_{max} = 15\%$ and the total number of resultant SuperAlarm patterns is 428.

3.2.3 Sequence representation

We formulate a SuperAlarm sequence as follows. Let $\mathbf{\Sigma} = \{SA_1, SA_2, \ldots, SA_m\}$ be a set of m distinct SuperAlarm patterns. A SuperAlarm sequence $\mathbf{S}$ is denoted as $\mathbf{S} = \langle SA_{t_1}, SA_{t_2}, \ldots, SA_{t_n} \rangle$, where $SA_{t_i} \in \mathbf{\Sigma}$ is a SuperAlarm trigger occurring at time t_i. A T_s-long SuperAlarm subsequence is a segment of SuperAlarm sequence denoted as $\mathbf{s} = \langle SA_{t_a - T_s}, SA_{t_a - T_s + 1}, \ldots, SA_{t_a} \rangle$, where $t_1 \leq t_a - T_s, t_a \leq t_n$. We call SA_{t_a} the anchor SuperAlarm trigger for the subsequence $\mathbf{s}$.

Inspired by approaches developed for representing documents in the field of information retrieval [98], we treat each SuperAlarm pattern as a word in a vocabulary consisting of m distinct SuperAlarm patterns. Then each subsequence can be treated as a document written using words from this vocabulary. This analogy enables us to adopt the vector space model [99] to represent each subsequence as a vector where each component in the vector corresponds to a particular SuperAlarm pattern. In the vector space model, term frequency inverse document frequency (TFIDF) [94] is one of the well-known weighting schemes used to assign a weight to each component in the vector. TFIDF explores the importance of a given SuperAlarm pattern in a given subsequence and within the entire training dataset by evaluating the term frequency (TF) of the SuperAlarm pattern in the subsequence multiplied by its inverse document frequency (IDF) calculated over the entire training dataset [100]. As a result, a subsequence $\mathbf{s}_j$ represented as a numeric-valued vector using TFIDF can be written as $\mathbf{tfidf}_j = (tfidf_{1j}, tfidf_{2j}, \ldots, tfidf_{mj})^T$, where m is the total number of SuperAlarm patterns in $\mathbf{\Sigma}$. The component $tfidf_{ij}$ is defined as

$$tfidf_{ij} = tf_{ij} \times idf_i \tag{3.1}$$

where $tf_{ij} = \log(1 + n_{ij})$, n_{ij} is the number of the SuperAlarm trigger SA_i occurring in the subsequence $\mathbf{s}_j$, $idf_i = \log \frac{N}{1+df_i}$, $df_i = \sum_{j=1}^{N} \mathbb{I}(n_{ij} > 0)$ calculates the number of subsequences in the training dataset containing the SuperAlarm trigger SA_i, N is the total number of subsequences in the training dataset. Note that: 1) the logarithm transformation used for tf_{ij} is to reduce the effect of the SuperAlarm trigger SA_i occurring many times within the subsequence $\mathbf{s}_j$; 2) the IDF favors the rare SuperAlarm triggers, which means that common SuperAlarm triggers occurring in the majority of subsequences in the training dataset will have lower IDF values than uncommon ones.

In order to eliminate the impact of length of the subsequence (i.e., the amount of SuperAlarm triggers occurring in the subsequence), the cosine normalization is applied to the TFIDF vector $\mathbf{tfidf}_j$, which is defined as

$$x_{ij} = \begin{cases} \dfrac{tfidf_{ij}}{\|\mathbf{tfidf}_j\|_2}, & \text{if } \|\mathbf{tfidf}_j\|_2 \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

The vector $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{mj})^T$ is the final representation of subsequence $\mathbf{s}_j$.

3.2.4 Feature selection

The above normalized TFIDF representation approach will result in a high-dimensional sparse vector. For machine learning problems with high-dimensional sparse vectors, dimension reduction of features has proven to be a beneficial step [96, 101]. In this study, we adopt a feature selection method to find a subset of SuperAlarm patterns that are highly relevant to the prediction of “code blue” event.

In particular, we use information gain (IG) [95] as the feature selection method. In essence, IG measures the expected reduction in entropy of one random variable having knowledge of the other. IG generally exhibits a competitive performance in text classification in comparison with other approaches [102]. In addition to its wide use in text classification, IG has been successfully applied in bioinformatics [103] and medical diagnosis [104], which justifies its adoption in the present study.

In this study IG is used to evaluate the amount of information obtained for “code blue” event prediction by observing the presence or absence of a SuperAlarm trigger SA_i in subsequences from the training dataset. Let c_+ be the positive class, c_- the negative class, $SA_i = 1$ the presence of SuperAlarm trigger SA_i in a subsequence and $SA_i = 0$ the absence in the subsequence. The IG of SA_i is given by

$$\begin{aligned} IG(SA_i) = & - \sum_{c \in (c_+, c_-)} p(c, \mathbf{X}) \log p(c, \mathbf{X}) \\ & + p(SA_i = 0, \mathbf{X}) \sum_{c \in (c_+, c_-)} p(c, \mathbf{X}_{SA_i=0}) \log p(c, \mathbf{X}_{SA_i=0}) \\ & + p(SA_i = 1, \mathbf{X}) \sum_{c \in (c_+, c_-)} p(c, \mathbf{X}_{SA_i=1}) \log p(c, \mathbf{X}_{SA_i=1}) \end{aligned} \tag{3.3}$$

where $\mathbf{X}$ is the training dataset containing all subsequences; $\mathbf{X}_{SA_i=0}$ ($\mathbf{X}_{SA_i=1}$) is the subset of $\mathbf{X}$ in which SA_i is absent (present); $p(SA_i = 0, \mathbf{X})$ ($p(SA_i = 1, \mathbf{X})$) is the probability of subsequences in $\mathbf{X}$ that SA_i is absent (present); $p(c, \mathbf{X})$, $p(c, \mathbf{X}_{SA_i\,=\,0})$ and $p(c, \mathbf{X}_{SA_i\,=\,1})$ are the probabilities of subsequences in $\mathbf{X}$, $\mathbf{X}_{SA_i\,=\,0}$ and $\mathbf{X}_{SA_i\,=\,1}$ that belongs to the class c, respectively. Note that $p(SA_i = 0, \mathbf{X}) + p(SA_i = 1, \mathbf{X}) = 1, \forall SA_i \in \mathbf{\Sigma}$.

By applying (3.3) to each of the m distinct SuperAlarm patterns and ranking them in terms of the IG values in decreasing order, the top k SuperAlarm patterns with the highest IG values are selected for TFIDF representation.

3.2.5 Weighted support vector machine

The SVM has been extensively used in numerous real-world applications as it often exhibits highly competitive performance in comparison with other classification methods [105]. Therefore, we adopt it in this study.

Due to the imbalance of the training dataset in this study (more control patients than coded patients), the conventional SVM classifier tends to simply classify positive samples into the majority class (i.e., negative class) because the learned hyperplane is too close to the positive samples [106]. A strategy to handle this issue has been proposed by assigning different penalties of misclassification costs to each of classes, which is called weighted SVM [97]. In this way, the hyperplane will be pushed away from the positive samples and towards the negative ones [106]. The weighted SVM is defined as

$$\min \frac{1}{2}\|\mathbf{w}\|^2 \;+\; C^+ \sum_{i \in S^+} \xi_i \;+\; C^- \sum_{i \in S^-} \xi_i \tag{3.4}$$

subject to

$$y_i(\mathbf{w} \cdot \phi(\mathbf{x}_i) \;+\; b) \geq 1 \;-\; \xi_i$$

$$\xi_i \geq 0,\, i \;=\; 1, 2, ..., N$$

where ξ_i is the slack variable; C^+ and C^- are the penalty parameters of misclassification costs for positive samples S^+ and negative samples S^- in the training set, respectively; $y_i \in \{-1, +1\}$ is the label for $\mathbf{x}_i$ that indicates a positive ($y_i = +1$) or a negative ($y_i = -1$) sample; $\mathbf{w}$ is the normal vector to the hyperplane; $\phi(\mathbf{x}_i)$ is a function to map vector $\mathbf{x}_i$ into a new feature space; b is the bias; N is the total number of samples in the training dataset.

An empirical method has been provided for setting the penalty ratio to the inverse of the number of samples in each class by assuming that the number of misclassified samples from each class is proportional to the number of samples in each class [107]. The penalty ratio is given by

$$\frac{C^+}{C^-} \;=\; \frac{n^-}{n^+} \tag{3.5}$$

where n^+ and n^- are the amount of samples in positive class and negative class, respectively.

Let C be a parameter, ω^+ (ω^-) the weight of positive class (negative class). Suppose $C^+ = \omega^+ C$, $C^- = \omega^- C$, with Equation (3.5) we have $\frac{\omega^+}{\omega^-} = \frac{n^-}{n^+}$, that is, the overall weight of each class is equal (i.e., $\omega^+ \cdot n^+ \;=\; \omega^- \cdot n^-$). Let ω^- be fixed (e.g., $\omega^- \;=\; 1$), then we have

$$C^+ \;=\; \frac{n^- \omega^- C}{n^+} \;=\; \frac{n^- C}{n^+} \tag{3.6}$$

$$C^- \;=\; \omega^- C \;=\; C \tag{3.7}$$

Therefore, the penalty C^+ for positive samples in the minority class will become larger (higher weight) than the penalty C^- for negative samples in the majority class. Equations (3.6) and (3.7) allow leaving only one parameter (i.e., C) to be learned.

The solution to classify a new sample vector $\mathbf{x}_{new}$ using the weighted SVM classifier with optimal parameters $\mathbf{w}^*$ and b^* learned from objective function (3.4) is given by

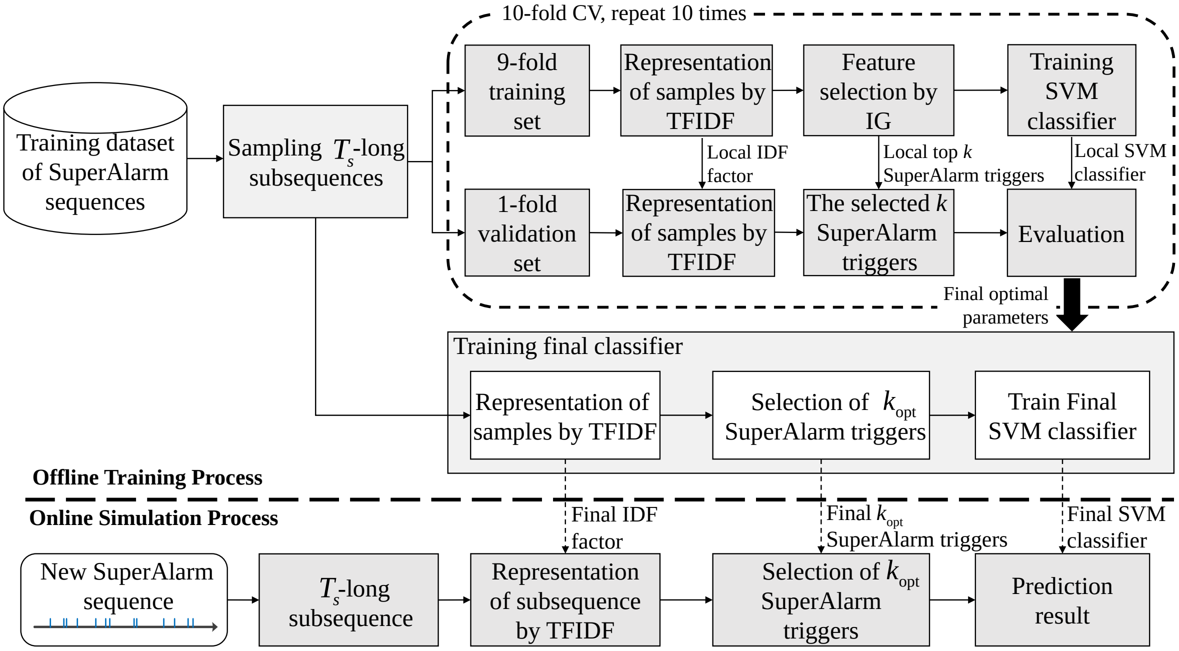
$$f(\mathbf{x}_{new}) = \begin{cases} +1, & \text{if } \mathbf{w}^* \cdot \phi(\mathbf{x}_{new}) + b^* > \text{threshold} \\ -1, & \text{if } \mathbf{w}^* \cdot \phi(\mathbf{x}_{new}) + b^* \leq \text{threshold} \end{cases} \tag{3.8}$$

where threshold is usually set equal to zero (i.e., default threshold).

In this study we will use the linear kernel (mapping function $\phi(\mathbf{x}) = \mathbf{x}$) for the following reasons: 1) the linear kernel measures the cosine similarity between samples in the original feature space; 2) the linear kernel can achieve better performance in comparison with other types of kernel functions when the original input vector is high-dimensional and the training set is large [108]; and 3) since input vector $\mathbf{x}$ is a normalized TFIDF vector, the linear kernel defined by inner product of two sample vectors can approximate the Fisher kernel [109]. We use the implementation of this algorithm as found in the LIBLINEAR library [110](v1.96, `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`).

3.2.6 Experiment and evaluation of results

Figure 3.2 provides an overview of the experiment to evaluate the proposed sequence classification approach. We use SuperAlarm sequence as an example in this figure as well as in the following description but the processes are applied to all three types of sequences. The experiment consists of two major processes: 1) offline training process, in which the SVM model with optimal parameters, the final set of relevant SuperAlarm patterns as determined by the IG method, and the IDF factor are obtained using the training dataset; and 2) online simulation process, in which evaluation of the SVM model is performed based on an independent test dataset.



10-fold CV, repeat 10 times
Training dataset of SuperAlarm sequences
Sampling T_s-long subsequences
9-fold training set
Representation of samples by TFIDF
Feature selection by IG
Training SVM classifier
Local IDF factor
Local top k SuperAlarm triggers
Local SVM classifier
1-fold validation set
Representation of samples by TFIDF
The selected k SuperAlarm triggers
Evaluation
Final optimal parameters
Training final classifier
Representation of samples by TFIDF
Selection of k_{opt} SuperAlarm triggers
Train Final SVM classifier
Offline Training Process
Online Simulation Process
Final IDF factor
Final k_{opt} SuperAlarm triggers
Final SVM classifier
New SuperAlarm sequence
T_s-long subsequence
Representation of subsequence by TFIDF
Selection of k_{opt} SuperAlarm triggers
Prediction result

Figure 3.2: The flowchart of the proposed framework of predicting code blue events using SuperAlarm sequences. $k = \lfloor r \cdot m \rfloor$, where r is feature selection ratio, m is the number of distinct SuperAlarm patterns in the dataset, $\lfloor r \cdot m \rfloor$ is referred to as the maximum integral number not greater than $r \cdot m$.

3.2.6.1 Sampling subsequences

As described in Section 3.2.3, the formulation of a SuperAlarm subsequence $\mathbf{s}$ is controlled by two parameters: the length of the subsequence T_s and the anchor SuperAlarm trigger SA_{t_a} that occurs at time t_a. T_s is an algorithm parameter that will be varied to study its effect. Many anchor triggers are randomly sampled for a given patient. A conventional technique to sample these anchor triggers is window-based, which extracts samples by sliding T_s-long window along the complete sequence [111]. However, based on an intuitive heuristic that subsequences closer to "code blue" events are more predictive, we propose to have a higher probability to select anchor triggers that are closer to "code blue" events. We use an exponential probability density function to model the probability of selecting a SuperAlarm trigger as illustrated in Figure 3.3(a).

Subsequences that are extracted from coded patients are treated as *positive samples*. For control patients, we select anchor triggers following a uniform distribution as illustrated in Figure 3.3(b). The orange vertical bars in Figure 3.3 represent the selected SuperAlarm triggers while the black vertical bars represent the SuperAlarm triggers that are not sampled.

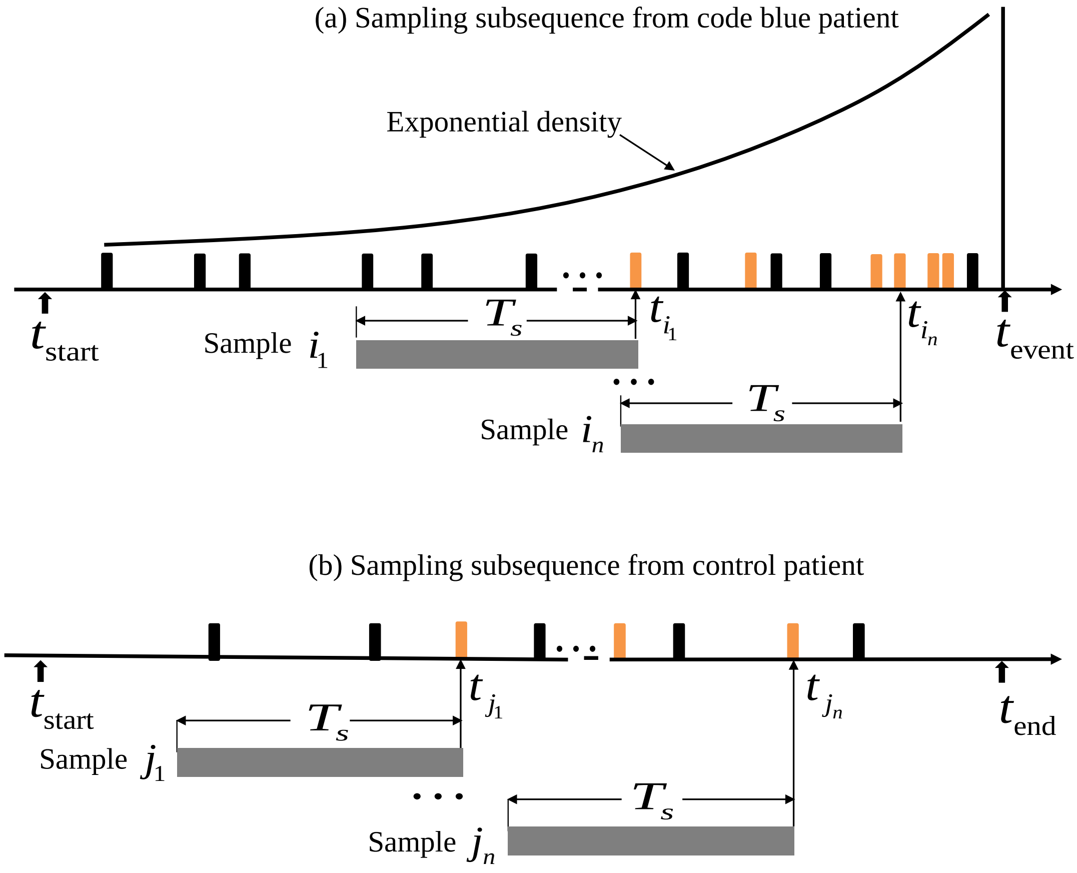


Figure 3.3: (a) Sampling subsequences from a coded patient, (b) sampling subsequences from a control patient.

3.2.6.2 Offline training process

The goal of the offline training process is to determine the optimal algorithm parameters for the final SVM classifier. To create the 10-fold cross-validation (CV) dataset, we randomly divide the positive samples and negative samples in the training dataset into 10 equal partitions. Samples in one partition are used for validation while the remainders for selecting features and training the classifier. This procedure is repeated 10 times and the optimal algorithm parameters can be determined by averaging a performance metric across 10 folds. In particular, we consider three algorithm parameters, including T_s, the cutoff of feature selection ratio r , and the parameter C in the SVM model. The F_1 score is used to determine the optimal parameters.

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3.9}$$

where $precision = \frac{TP}{TP+FP}$, $recall = \frac{TP}{TP+FN} = sensitivity$, TP (true positive) is the number of positive samples predicted correctly by the classifier, FP (false positive) is the number of negative samples predicted incorrectly, FN (false negative) is the number of positive samples predicted incorrectly. The reasons we utilize F_1 score are: 1) the F_1 score is a harmonic mean of precision and recall and conveys a trade-off measure between them; and 2) as a composite measure, the F_1 score weights more on positive samples, making it more likely to select parameter settings that lead to more sensitive classifiers. After determining the optimal parameters, the final SVM classifier is trained using the entire training dataset obtained by coalescing the 10-fold CV dataset into a single one. It should be noted that the IDF resulting from the TFIDF weighting scheme based on the entire training dataset (termed final IDF factor as shown in Figure 3.2) will be stored and used in the online simulation analysis.

3.2.6.3 Online simulation analysis

We employ an independent test dataset to simulate the application of the learned SVM classifier acting on a SuperAlarm sequence in real-time and assess the performance in predicting “code blue” event. At every single SuperAlarm trigger, a T_s-long subsequence immediately preceding this trigger will be evaluated using the learned SVM classifier. The T_s-long subsequence is first represented as a vector by the normalized TFIDF. Only those components in the vector that are retained based on IG criterion in the offline training phase will be used. To obtain a binary outcome, a threshold is specified and applied to the continuous-valued output of the learned SVM classifier. We derive an optimal threshold based on the receiver operating characteristic (ROC) analysis.

The following three metrics are employed to assess the performance of the SVM classifier based on the independent test dataset:

- Sensitivity of lead time ($\mathrm{Sen}^L@T$). This metric is computed in terms of percentage of coded patients predicted correctly at least once by the SVM classifier within a 12-hour window that is T hours ahead of the "code blue" event.
- Alarm Frequency Reduction Rate (AFRR). This metric is defined as AFRR $=$ $1-$FPR, where the FPR is the false positive ratio calculated as a ratio of the hourly rate of positive predictions from the SVM model to the hourly rate of monitor alarms among the control patients.
- Work-up to detection ratio (WDR). The WDR is defined as WDR $= \frac{a+b}{a}$, where a is the number of coded patients predicted correctly at least once (i.e., true positives, TPs) by the SVM classifier within a 12-hour window preceding "code blue" events, b is the number of control patients predicted incorrectly at least once (i.e., false positives, FPs) within window of the same length. The WDR measures how many FPs can be introduced using the SVM classifier when one TP is achieved.

3.2.7 Algorithm parameter evaluated

We studied seven T_s values with $T_s \in \{2, 4, 6, 8, 10, 12, \infty\}$ hours, where ∞ implies that a subsequence is sampled from the beginning of monitoring to a current anchor SuperAlarm trigger. Various values are specified for the SVM parameter $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{15}\}$ and feature selection ratio $r \in \{10\%, 20\%, \ldots, 100\%\}$. For each T_s, optimal values of r and C are determined by performing the 10-fold CV over a 2-D grid search in terms of F_1 score.

3.3 Results

3.3.1 Patient data

The same patient cohort as described in our previous study [92] was employed in this study. This cohort has a total of 254 adult patients experiencing at least one "code blue" event during their hospitalization between March 2010 and June 2012 at the University of California, Los Angeles (UCLA) Ronald Regan Medical Center and a total of 2213 control patients. Compared with a coded patient, control patients had same APR DRG (All Patient Refined Diagnosis Related Group) or Medicare DRG, the same age (±5 years), the same gender, and stayed in the same hospital unit within the same period as coded patients. Patient's gender was male for 54% and 68% of the coded and the control patients, respectively. Average age was 61.6 ± 18.2 years and 63.5 ± 14.6 years for the coded and control patients, respectively. The analysis of the patient data was approved by the Institutional Review Board with a waiver of patient consent. The training dataset was composed of monitor alarms and laboratory test results from randomly selected 80% of both coded and control patients. Data from the remaining 20% patients were used as the independent test dataset.

3.3.2 Characteristics of sampled subsequences in training dataset

After excluding patients without any SuperAlarm triggers, the training dataset used in this study consisted of 176 coded and 1766 control patients. The independent test dataset contained data from 30 coded and 440 control patients. This test dataset is identical to the one used in our previous study [92]. By applying the subsequence generation method described in Section 3.2.6.1 with a maximal number of sampled subsequences being 60 per each coded patient and 10 per each control patient, we obtain 7174 SuperAlarm positive samples (40.76 per each coded patient) and 12522 SuperAlarm negative samples (7.09 per each control patient) in the training dataset. We apply the same protocol as used for sampling SuperAlarm subsequences to the raw alarm sequences and discretized alarm sequences. The number of positive samples and negative samples in the training dataset are 7719 (43.86 per each coded patient) and 13709 (7.76 per each control patient) for the raw alarm sequences, and 7723 (43.88 per each coded patient) and 13709 (7.76 per each control patient) for the discretized alarm sequences, respectively.

3.3.3 Offline training results

For a given subsequence length T_s, each combination of the SVM parameter C and feature selection ratio r is applied to train SVM model and the average F_1 score across the 10-fold CV set is employed as a performance metric for assessing this parameter combination. In addition, we select the optimal C that corresponds to the largest average F_1 at each r. Table 3.1 reports these results for each of the three types of sequences. From Table 3.1, we can see that for a given T_s and r, average F_1 score for SuperAlarm sequence is consistently higher than that of both discretized alarm sequence and raw alarm sequence, but no difference in F_1 score could be seen between the discretized alarm sequence and raw alarm sequence. We can also observe that as T_s increases from 2 hours to ∞, r associated with the highest average F_1 is between [30–60%] for the SuperAlarm sequence, [10–80%] for discretized alarm sequence, and [10–100%] for raw alarm sequence, respectively.

Based on the results shown in Table 3.1, a two-way analysis of variance (2-way ANOVA) of sampling window ($T_s \in \{2, 4, 6, 8, 10, 12, \infty\}$) and feature selection ratio ($r \in \{10\%, 20\%, \ldots, 100\%\}$) on average F_1 score is conducted for each of the three types of sequences. The results of the 2-way ANOVA test show that for each of the three types of sequences, the main effect of feature selection ratio on average F_1 score is not significant ($p > 0.05$) while significant main effect of the length of sampling window on average F_1 score exists ($p < 0.05$). The interaction effect between these two factors is not significant ($p > 0.05$).

Table 3.1: Results of the F_1 score for the three types of sequences at various feature selection ratio r based on disparate sampling window T_s (hours) in the training phase. The bold numbers indicates the highest F_1 score for a given type of sequence based on a given sampling window.

		F_1 score (%, mean±std)									
T_s	Sequence Type	$r = 10\%$	$r = 20\%$	$r = 30\%$	$r = 40\%$	$r = 50\%$	$r = 60\%$	$r = 70\%$	$r = 80\%$	$r = 90\%$	$r = 100\%$
2	SuperAlarm	61.89±8.74	67.13±11.63	**70.12 ± 10.93**	68.06±10.56	68.99±9.80	67.07±9.89	65.40±9.27	63.71±8.44	60.56±10.14	61.38±10.93
	Discretized Alarm	51.72±5.94	**52.85 ± 3.55**	52.56±3.16	52.11±3.14	51.60±3.78	51.43±3.72	51.48±3.96	51.48±3.96	51.48±3.96	51.48±3.96
	Raw Alarm	**52.90 ± 6.82**	50.11±6.15	50.71±6.27	49.79±5.99	51.27±6.42	50.93±6.68	52.00±6.00	52.16±5.41	52.54±5.12	52.54±5.12
4	SuperAlarm	68.99±12.32	62.24±8.39	66.55±9.84	69.47±7.25	**69.92 ± 8.50**	66.47±8.59	65.61±9.64	63.91±10.41	61.78±11.37	62.31±10.06
	Discretized Alarm	51.72±4.56	**53.55 ± 4.00**	52.91±4.57	51.71±5.44	51.89±5.13	51.67±2.87	52.32±2.41	52.32±2.41	52.32±2.41	52.32±2.41
	Raw Alarm	**55.19 ± 10.55**	51.72±7.64	50.28±7.30	51.12±8.02	51.82±8.41	51.51±8.51	50.58±8.53	50.58±7.64	52.65±7.29	52.65±7.29
6	SuperAlarm	64.96±12.61	58.72±12.85	67.19±4.95	**69.67 ± 10.41**	69.63±8.45	67.78±8.68	66.20±9.89	63.62±9.55	60.59±10.42	61.81±10.69
	Discretized Alarm	50.70±4.98	51.23±5.14	51.90±3.86	51.18±3.69	51.76±3.53	51.96±3.22	51.91±2.68	**52.03 ± 2.52**	52.03±2.52	52.03±2.52
	Raw Alarm	**56.97 ± 10.38**	50.33±8.21	50.31±6.70	51.02±7.35	51.82±7.89	51.33±7.69	51.81±7.94	52.61±8.43	52.26±7.64	52.26±7.64
8	SuperAlarm	65.08±9.46	58.86±12.37	61.85±8.85	70.94±7.02	**71.64 ± 6.52**	69.10±8.82	66.95±8.98	65.35±8.65	61.09±10.35	62.18±9.77
	Discretized Alarm	52.62±4.45	51.43±4.74	52.20±4.74	50.70±5.18	51.18±4.50	51.54±3.32	52.63±2.84	**52.64 ± 2.48**	52.64±2.48	52.64±2.48
	Raw Alarm	**56.57 ± 8.75**	54.49±10.14	50.61±7.77	51.70±7.83	51.33±6.56	51.54±7.55	52.17±7.00	53.13±6.59	52.01±7.99	51.84±7.84
10	SuperAlarm	66.21±11.69	61.61±12.06	66.42±8.94	71.03±7.18	**72.42 ± 7.40**	70.63±7.92	67.09±9.41	65.89±8.51	62.95±10.58	62.12±7.90
	Discretized Alarm	52.59±6.34	49.45±5.51	52.04±4.75	**52.85 ± 3.55**	51.33±3.75	52.44±3.80	51.53±3.45	51.75±2.54	51.75±2.54	51.75±2.54
	Raw Alarm	52.53±13.56	**55.21 ± 10.57**	51.98±8.03	50.76±6.99	52.76±7.41	51.71±6.22	50.89±6.91	51.78±5.03	51.54±8.14	51.85±8.04
12	SuperAlarm	67.98±12.71	64.60±11.97	68.35±7.89	71.72±5.81	**71.98 ± 7.39**	71.39±8.86	66.89±8.32	66.42±10.94	64.02±10.43	63.63±8.59
	Discretized Alarm	**52.04 ± 5.32**	51.03±5.87	50.58±5.34	51.99±5.07	50.89±4.84	50.91±3.68	51.05±3.13	51.09±2.58	51.09±2.58	51.09±2.58
	Raw Alarm	51.53±13.42	52.33±10.73	**52.42 ± 9.05**	51.42±8.65	51.04±7.95	51.21±6.96	52.04±5.84	51.61±7.08	51.39±7.37	51.41±8.00
∞	SuperAlarm	63.15±10.69	63.07±11.24	63.13±10.10	61.81±9.72	62.96±10.12	**65.36 ± 9.98**	64.47±12.00	64.14±9.31	61.79±12.34	60.60±9.98
	Discretized Alarm	**52.18 ± 9.41**	49.23±9.19	50.03±8.72	47.20±8.35	46.96±7.35	48.45±8.59	45.98±10.94	45.31±11.86	44.83±10.37	44.29±10.56
	Raw Alarm	51.44±9.26	49.40±8.42	49.43±7.62	46.89±9.74	48.68±8.79	50.20±8.73	51.11±9.59	50.09±8.96	50.68±7.47	**51.78 ± 7.09**

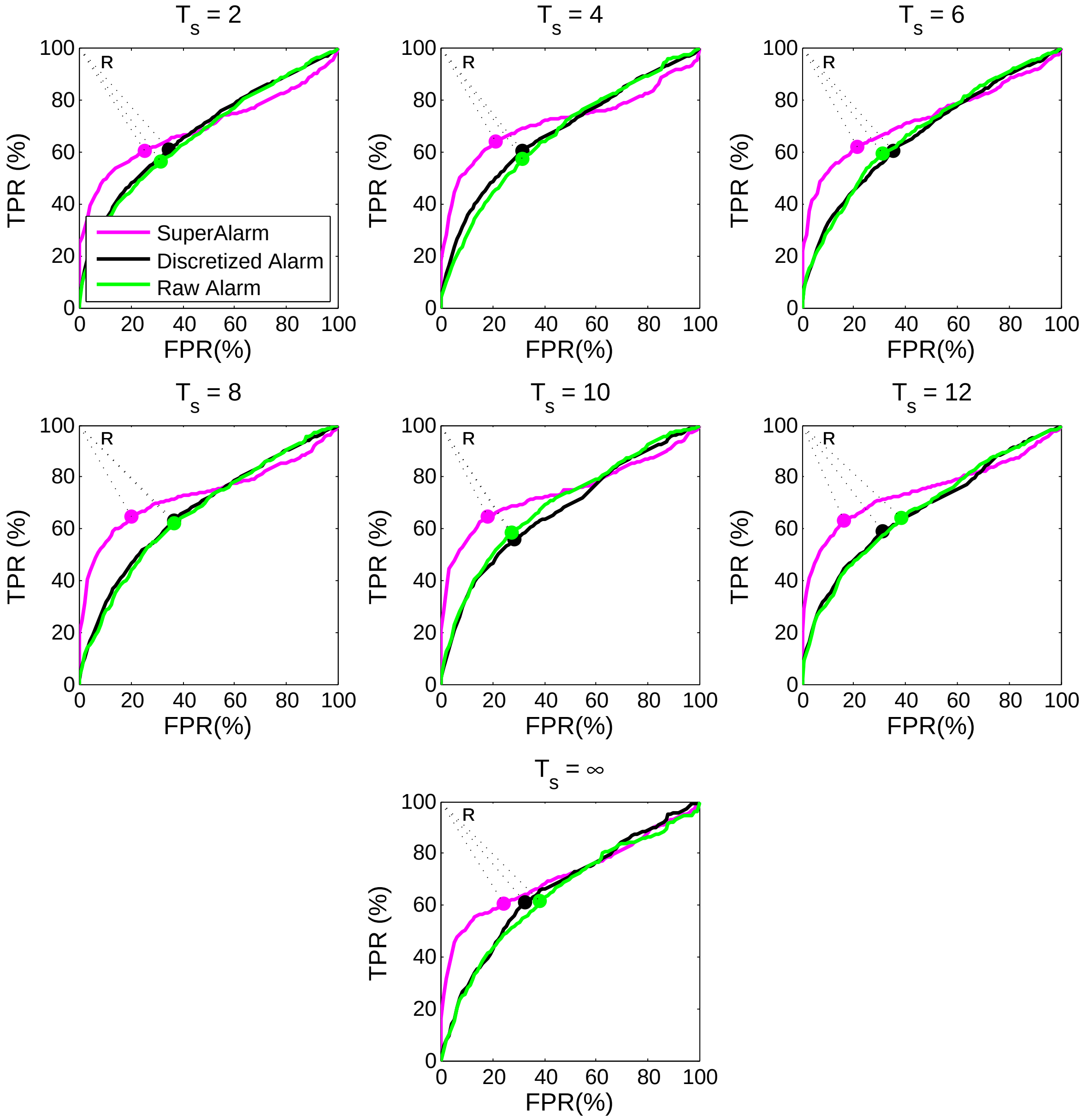


Figure 3.4: Average ROC curves for each of the three types of sequences based on 10-fold CV set. The optimal operating point on each curve corresponds to the point closest to the reference point R.

We vary the threshold to dichotomize SVM output and generate receiver operating characteristic (ROC) curves (Figure 3.4) for the three types of sequences using the 10-fold CV set under their corresponding optimal algorithm parameters. We then find the optimal operating point on each ROC curve that is closest to the reference point on the upper left corner of the unit square (i.e., point R in the Figure 3.4). The corresponding threshold at this optimal operating point is used as optimal SVM output threshold in the online analysis, as shown in Equation (3.8).

3.3.4 Online simulation results

Table 3.2 lists the three performance metrics based on the optimal SVM thresholds for classification. For instance, when varying T_s between 2 hours and ∞, the values of Sen^L@2 (sensitivity of 2-hour lead time) for SuperAlarm sequences, discretized alarm sequences and raw alarm sequences are [53.33–80.00%], [76.67–90.00%] and [56.67–83.33%], respectively. Meanwhile, the ranges of AFRR are [88.42–96.20%], [57.53–70.70%] and [51.64–86.25%], respectively, while the values of WDR are [1.94–2.93], [7.53–11.68] and [6.41–10.12], respectively.

Based on results in Table 3.2 alone, it is impossible to compare the performance of the three sequences because the three performance metrics are related and yet the values of these metrics are not controlled at the same levels for comparison. To address this issue, we plot sensitivity metric against AFRR and WDR, respectively, by varying thresholds for dichotomizing SVM output. The curve thus created is termed SvA curve for sensitivity versus AFRR, and SvW curve for sensitivity versus WDR. As an examples, we create these two types of curves for Sen^L@2 for SuperAlarm sequences under the subsequence length $T_s = 12$ hours. Here the reason we choose the 12-hour subsequence length is because that according to Table 3.2, given the 2-hour lead time, the SuperAlarm sequences has the highest sensitivity under the 12-hour sampling window. In addition, we plot similar curves for the raw alarm sequences and the discretized alarm sequences but under all studied sampling windows to offer a complete compar-

Table 3.2: Results of Sen^L@T, AFRR and WDR based on SVM classifiers under the optimal thresholds obtained from Figure 3.4.

		Sen^L@T (%)						
T_s	Sequence Type	0.5 hr	1 hr	2 hrs	6 hrs	12 hrs	AFRR (%, mean±std)	WDR (mean±std)
2	SuperAlarm	53.33	53.33	53.33	50.00	46.67	96.20±15.84	2.32±0.19
	Discretized Alarm	93.33	90.00	90.00	83.33	80.00	57.53±23.41	11.68±0.26
	Raw Alarm	93.33	90.00	83.33	76.67	76.67	70.73±23.92	10.12±0.28
4	SuperAlarm	70.00	63.33	63.33	60.00	53.33	95.32±15.11	2.30±0.18
	Discretized Alarm	93.33	90.00	86.67	80.00	70.00	61.16±25.46	10.56±0.29
	Raw Alarm	66.67	60.00	60.00	56.67	56.67	83.80±20.16	7.88±0.38
6	SuperAlarm	76.67	70.00	70.00	70.00	63.33	95.03±16.65	2.02±0.15
	Discretized Alarm	96.67	93.33	86.67	86.67	66.67	59.55±24.39	10.74±0.27
	Raw Alarm	63.33	60.00	60.00	50.00	43.33	86.05±18.74	6.81±0.38
8	SuperAlarm	70.00	66.67	66.67	63.33	56.67	95.46±14.78	2.16±0.17
	Discretized Alarm	93.33	90.00	83.33	83.33	63.33	61.38±25.68	10.45±0.29
	Raw Alarm	60.00	56.67	56.67	50.00	43.33	86.25±18.69	6.41±0.37
10	SuperAlarm	73.33	66.67	66.67	63.33	53.33	95.80±14.72	1.94±0.15
	Discretized Alarm	93.33	90.00	86.67	83.33	66.67	68.84±26.73	8.28±0.29
	Raw Alarm	90.00	83.33	76.67	70.00	70.00	67.27±28.91	8.36±0.29
12	SuperAlarm	86.67	80.00	80.00	80.00	66.67	88.42±24.00	2.93±0.16
	Discretized Alarm	90.00	86.67	80.00	73.33	63.33	70.70±27.20	7.53±0.30
	Raw Alarm	90.00	86.67	86.67	76.67	76.67	58.97±30.54	9.11±0.28
∞	SuperAlarm	73.33	70.00	70.00	66.67	60.00	90.71±23.56	2.85±0.16
	Discretized Alarm	76.67	76.67	76.67	70.00	73.33	62.03±40.46	7.99±0.25
	Raw Alarm	90.00	86.67	86.67	86.67	86.67	51.64±41.58	8.60±0.23

ison. The results of the SvA and SvW curves are shown in Figure 3.5(a) and Figure 3.5(b), respectively, where the colored band corresponds to SvA and SvW curves for the raw alarm sequences and discretized alarm sequences. Comparing SvA curves, we can see that sensitivity of SuperAlarm sequence is the highest for the desirable range of high AFRR. The sensitivity of SuperAlarm sequence also remains the highest for the desirable range of low WDR. From this curve, one can see that the optimal SVM threshold could have been determined corresponding to the circle point on the SvA or SvW curve where sensitivity first reaches a plateau as AFRR decreases or WDR increases. The optimal SVM threshold as determined from the training data does not exactly match the optimal choice based on the testing data, which may not be obtainable.

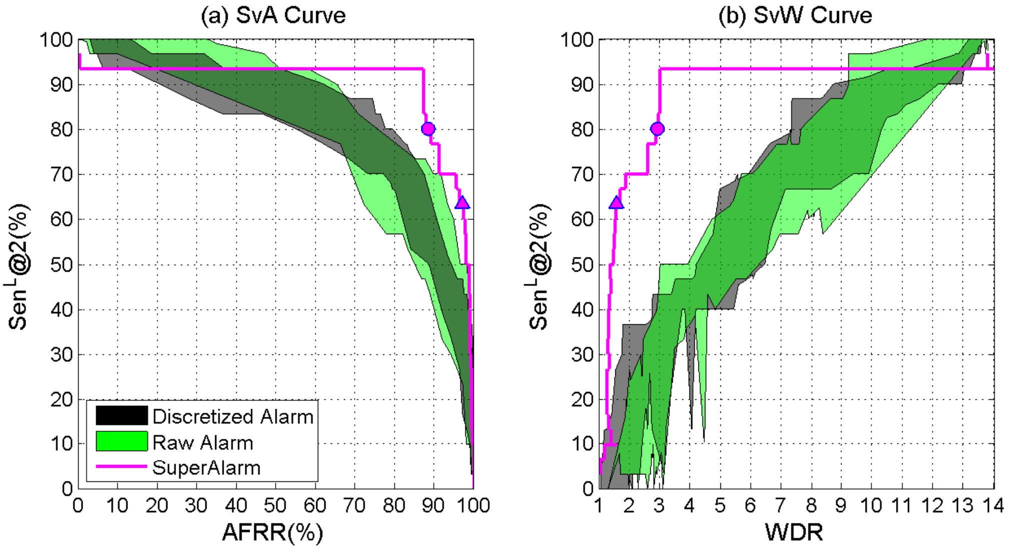


Figure 3.5: (a) SvA curve: Sen^L@2 versus AFRR.(b) SvW curve: Sen^L@2 versus WDR. The ranges are displayed for discretized alarm sequences (black) and raw alarm sequences (green) under all specified subsequence lengths (from 2 hours to ∞). The curves for SuperAlarm sequences (magenta) are created based on the 12-hour subsequence length. The circle on the curve represents the pair of values obtained using the optimal SVM threshold while the triangle on the curve represents that obtained using the default SVM threshold (i.e., zero).

3.4 Discussion

This study compares prediction of in hospital code blue events using sequences of SuperAlarm triggers, monitor alarms, and discretized monitor alarms. Identical sequence representation and machine learning model are used to build the classifer. Based on the results from an independent test dataset, highest sensitivity with respect to 2-hour lead time (Sen^L@2) is obtained by using the SuperAlarm sequence under a desirable range of high alarm frequency reduction rate (AFRR) or low work-up to detection ratio (WDR). Particularly, Figure 3.5 shows that SuperAlarm sequence achieved 93.33% $Sen^L@2$ while keeping AFRR = 87.28% and WDR = 3.01. This performance is also better than what we achieved in our previous study of using individual SuperAlarm triggers to predict code blue events where Sen^L@2 was 90.0% with AFRR = 85.2% and WDR = 6.5 [92] under the same training and test datasets.

Figure 3.5 clearly shows the advantage of SuperAlarm sequences to predict code blue events as compared to the raw alarm sequences and discretized alarm sequences by having a higher sensitivity under a desirable range of high AFRR or low WDR. One likely explanation for this better performance is the inherent multivariate nature of SuperAlarm patterns because each pattern is a combination of different monitor alarms and laboratory test results that can better characterize a patient's physiological status than what a single variable can do. One could argue that sequences of raw or discretized monitor alarms also embed multivariate patterns. However, these sequences are more susceptible to false alarms. As discussed in our previous work, SuperAlarm patterns are less influenced by false alarms. For instance, a false ECG arrhythmia alarm is unlikely to have a co-occurring blood pressure alarm while clinically significant ECG arrhythmia may also compromise hemodynamic status and cause a co-occurring blood pressure alarm [112].

Comparing the triangle and circle points in Figure 3.5 we can observe that adjusting SVM threshold in decision function, as shown in Equation (3.8), can have a significant impact on the binary prediction performance. In this study, we seek to select an optimal SVM threshold by choosing the point on the ROC curve that is closest to the ideal predictor with 100% sensitivity and 100% specificity. Nevertheless, as illustrated in Figure 3.5 this optimal threshold determined from the training dataset does not exactly match the optimal choice that could have been determined using the independent test dataset. The likely contributor to this discrepancy might be due to the lack of sufficient training data that are representative of the characteristics of the testing data.

Consequently, the SVM classifier with optimal threshold obtained from training data may not guarantee the same performance as that on the testing data. Determining the optimal SVM threshold in attempts to achieve the excellent prediction performance based on an imbalanced dataset remains a challenging problem. Few studies reported the adjustment of decision thresholds for machine learning algorithms based on the ROC analysis [113–115] and therefore further studies are needed.

All patient monitor alarms analyzed in this study were audible and contributed to the alarm fatigue problem. Since we do not include patient crisis alarms such as ventricular fibrillation (VFib) and asystole as part of SuperAlarm patterns, it would be possible to lower criticality levels of the alarms (an intervention to address alarm fatigue that has been reported) that are part of SuperAlarm patterns so that they are not audible or have fewer number of beeps on the bedside monitors but are transmitted in real-time to a backend system running SuperAlarm sequence classifier to detect patient deterioration. At a sensitivity of 93.3% of predicting "code blue" event, the alarm frequency of such a backend system would be only about 13% of the monitor alarms presumably offering a significant alleviation of alarm fatigue. Even though the current algorithm's sensitivity is still well below 100%, it should be pointed out that the current algorithm can be easily augmented by adding back crisis alarms to offer greater sensitivity. Indeed, such a hypothetical algorithm could be adopted by the primary monitor if sensitivity is close to 100%. Another potential use case is to have this system function as a secondary patient monitor to provide additional safety net in situations where some of these non-crisis alarms may not be noticed by bedside caregivers or when their criticality levels are further lowered as an intervention to address the alarm burden.

Future work is needed to study different approaches to represent sequences, perform feature selection, and select appropriate classifier model because how these factors influence ultimate performance in detecting patient deterioration is not the focus of this work. However, the main conclusion regarding the improved performance of SuperAlarm sequence over raw and discretized monitor sequences will still hold if the improved approaches are applied in an equal fashion to these three sequences. Future studies also need to be conducted in a real-time and prospective manner to evaluate feasibility of streaming data analytics, true predictive power of the SuperAlarm approach.

3.5 Conclusion

We have studied the prediction of code blue events using the sequence of SuperAlarm triggers. We proposed a new method to sample subsequences from the compete sequences. We employed term frequency inverse document frequency method to represent the sequences as fixed-dimension numerical-value vectors. Information gain was used to select most relevant SuperAlarm patterns as a preprocessing step in the training phase. We applied weighted support vector machine to build the prediction model. Three metrics were assessed based on the independent test dataset in the simulation online analysis: sensitivity at different lead time choices, alarm frequency reduction rate and work up to detection ratio. Results have demonstrated that sequence of SuperAlarm triggers is more predictive than sequence of monitor alarms as it has higher sensitivity under a desirable range of high alarm frequency reduction rate or low work-up to detection ratio. Therefore, the proposed SuperAlarm sequence classifier may assist in predicting patient deterioration and reducing alarm burden.

CHAPTER 4

Development of a comprehensive database for SuperAlarm study

A large-scale and comprehensive patient dataset is required for the development and evaluation of advanced SuperAlarm algorithms. In this Chapter, we report a research database, called SuperAlarm study database II that we have developed primarily for this purpose by aggregating a large volume of temporal physiologic and clinical data from both UCLA and UCSF Medical Centers. These two centers are involved because our study started at UCLA but then continued at both campuses. The SuperAlarm study database II so far includes patient demographics, admission-discharge-transfer (ADT) information, monitor alarms, laboratory test results, physiologic waveforms and vital signs collected from a large number of identified adult coded and control patients.

4.1 Introduction

In Chapters 2 and 3 we have described the SuperAlarm approach to predict code blue events and address alarm fatigue problem. The key of the SuperAlarm approach is to identify multivariate patterns hidden in the data streams of patient monitor alarms, physiologic waveforms, and data from the electronic health record (EHR) systems (e.g., laboratory test results). The development and proper evaluation of SuperAlarm algorithms require a large database that covers a wide variety of temporal physiologic and clinical data.

The SuperAlarm study database II contains temporal data from a large number of identified adult coded and control patients admitted to the ICUs in UCLA and UCSF Medical Centers. We extract two categories of such temporal data for establishing the database: clinical data (e.g., patient demographics, patient ADT information, laboratory test results) from the EHR system, and physiologic data (e.g., monitor alarms, vital sign measurements, high-resolution physiologic waveforms) from a central repository where all the available physiologic data generated by bedside monitors are archived. The primary purpose of establishing such database includes twofold: (1) The ongoing SuperAlarm study can be greatly facilitated by this comprehensive database when developing new algorithms; and (2) The high-volume, readily available patient data can also support a diverse array of analytic studies (e.g., physiologic signal analysis). The effort of developing the SuperAlarm study database II is ongoing: in addition to continuously collecting the aforementioned data from the two institutions, we also extracted other types of patient data such as medication administration records (MARs) as well as the data obtained beyond coded and corresponding control patients. The SuperAlarm study database II is intended to be disseminated for research groups at both UCLA and UCSF.

Although the raw physiologic and clinical data are readily available at both UCLA and UCSF Medical Centers, several challenges are encountered when using these data to create a comprehensive, aggregated and well-documented research database for the SuperAlarm study. First, patients are uniquely identified by local medical record numbers (MRNs) in the UCLA and UCSF hospitals, respectively. The MRNs and other protected health information should be de-identified in compliance with Health Insurance Portability and Accountability Act (HIPAA) to allow research groups at UCLA and UCSF to access to the SuperAlarm study database II. Second, the data, for instance, of alarms and laboratory test results extracted from the two institutions may have different names even for the same physiologic variables. Therefore, thorough understanding of the local proprietary data format and naming scheme is needed when integrating data from these two institutions. Furthermore, the same laboratory tests may have disparate units of measure between the UCLA and UCSF Medical Centers. In order to overcome these difficulties, we explore several technologies and strategies to collect the patient data from the two institutions in an automated manner.

This Chapter is organized as follows. In the next section we describe the methods for automated mapping of monitor alarms and laboratory test results extracted from the two institutions. In addition, we report an software application that was developed to extract physiologic waveforms and vital signs from the flat files archived by the BedMaster systems (Excel Medical Electronics, Inc, Jupiter, FL), and save them into binary files following a file format that is publicly available. Section 4.3 provides the characteristics and statistical summaries of the patient data collected in the database. Section 4.4 discusses our experiences and lessons learned from developing this database. The Chapter ends with conclusion presented in section 4.5.

4.2 Materials and Methods

Figure 4.1 illustrates the physiologic and clinical data archiving architecture in order to create the SuperAlarm study database II. We use a retrospective data collection design to extract physiologic and clinical data from patients admitted to intensive care unites (ICUs). The ICUs in both UCLA and UCSF Medical Centers are equipped with GE physiologic monitors (GE Healthcare, Milwaukee, WI) to continuously acquire and process patient physiologic data including monitor alarms, physiologic waveforms, vital sign measurements. BedMaster systems are installed at UCLA and UCSF Medical Centers, respectively. The BedMaster system serves as a central repository that stores monitor alarms and alarm settings into the SQL server database, and waveforms and vital signs in flat files but in a proprietary format. The reader can refer to a detailed description of UCSF hospital infrastructure used for automated storage of all physiologic monitor waveform and alarm date [1]. Meanwhile, patient demographics, admit-discharge-transfer (ADT) information, and laboratory test results are extracted from the EHR systems that are also implemented in UCLA and UCSF medical centers, respectively.

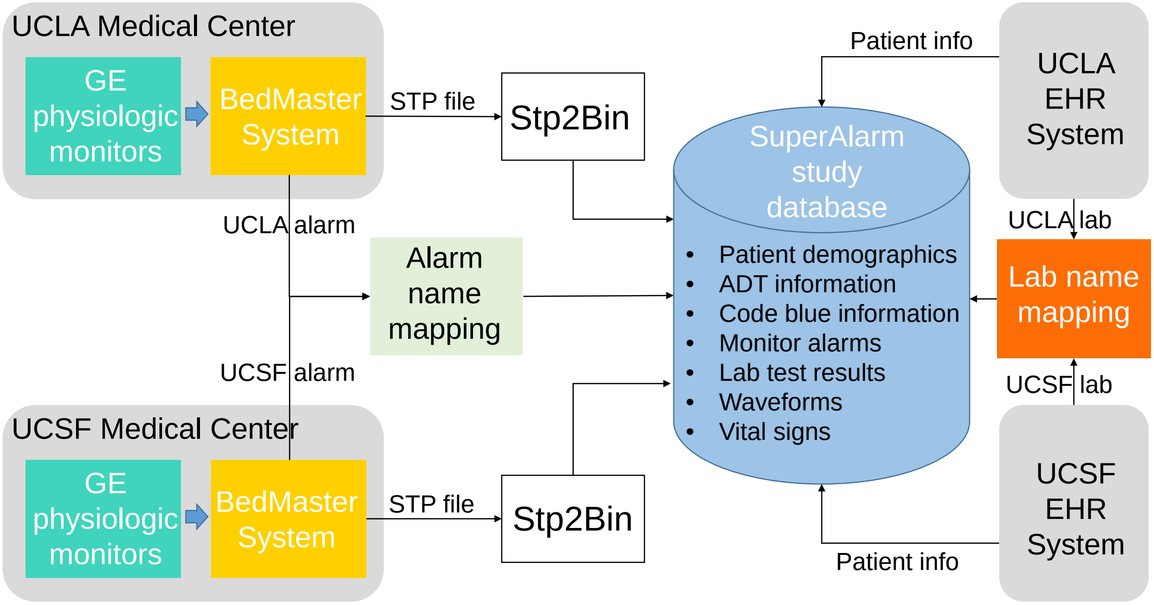


Figure 4.1: The physiologic and clinical data collection flowchart for establishing the SuperAlarm study database II.

4.2.1 Patient data

The acquisition of physiologic and clinical data from the two institutions did not impact the patient's routine clinical care. Therefore, this study has been approved by the UCLA Institutional Review Board and the UCSF Committee on Human Research with a waiver of patient consent, respectively. At the moment of the study, the clinical endpoint that will be predicted using the SuperAlarm approach is code blue event. Hence, all physiologic and clinical data collected for creating the SuperAlarm study database II at present are extracted from coded patients who were identified by quality management services at UCLA and UCSF Medical Centers and corresponding control patients.

The physiologic data including monitor alarms, waveforms (such as ECG, SpO_2 and ABP) and vital signs (such as HR and RR) are extracted from the BedMaster systems. The BedMaster systems store these data with timestamps and alarm settings (such as thresholds of parameter alarms) that are generated by GE bedside monitors for critically ill adult patients admitted to 120-bed ICUs (neurosurgical, cardiac, medical, and medical/surgical) at UCLA Medical Center and 77-bed ICUs (neurological/neurosurgical, medical/surgical, and cardiac critical care) at UCSF Medical Center, respectively. The clinical data for each of coded and control patients is obtained from EHR systems, which encompasses patient demographics (such as age at admission, gender, race and ethnicity), admit-discharge-transfer (ADT) information and laboratory test results (such as arterial blood gas, complete blood count and blood chemistry).

4.2.2 Patient deidentification

The protected health information (PHI) stored in physiologic and clinical databases should be de-identified to preserve patient anonymity and comply with the HIPAA before dissemination. Due to the structured data sources from the two institutes, it is straightforward to remove the protected health information (e.g., patient name, date of birth). Furthermore, we provide an encoding scheme for automated de-identification of MRNs. The scheme accommodates all MRNs in this study obtained from the two institutions and encodes each MRN uniquely.

4.2.3 Alarm mapping

The monitor alarms include arrhythmia alarms, vital sign parameter alarms. An arrhythmia alarm will be activated when a change in cardiac rhythm is detected by arrhythmia detection algorithms implemented in the physiologic monitoring devices. A parameter alarm will be triggered when its corresponding vital sign measurement falls outside the predefined alarm thresholds. In addition to monitor alarms, technical alarms are also generated by physiologic monitors. A technical alarm typically reflects device malfunction, e.g., “ECG LEADS FAIL”, and it does not represent patient status. These timestamped alarms are extracted from the SQL server database deployed along with the BedMasterEx system for all patients during their admission to the ICUs. However, UCLA and UCSF Medical Centers have disparate terminologies defined in their local physiologic monitors for some of the same alarms. These discrepancies cause the major impediments to automated collection and aggregation of alarms. In order for the SuperAlarm study database II to consolidate alarms obtained from the two institutions, an alarm codebook is established for mapping alarms from the two institutions.

With the alarm codebook, alarms are made agnostic to the port number of monitor devices to which the sensors for the same physiological variables are attached [74]. Alarms signaling abnormalities in invasive arterial blood pressure (ART), non-invasive arterial blood pressure (NIBP) and femoral (FEM) are treated equivalently and merged to name as blood pressure (BP). This is because these names are used for differing pressure connectors that the arterial pressure lines are plugged. In addition, we neglect the “PaceMode” features that are associated with some arrhythmia alarms for patients with ventricular pacemakers. Furthermore, numeric values of vital signs that trigger the corresponding parameter alarms are extracted from the raw alarm messages. The vital sign parameter alarms are then further named as alarm messages and polarities of HI (high) or LO (low). The polarity represents whether the value of the physiologic parameter is greater than an upper bound threshold (HI) or less than the lower bound threshold (LO) of the predefined alarm setting. For the alarm with detected numeric value equal to the predefined thresholds, the polarity of the alarm will be determined as LO if the distance between this value and the average value of all preset lower bound thresholds from the same patient is closer than that between the value and the average value of all upper bound thresholds, otherwise, it is determined as HI. This is because the alarm thresholds can be adjusted by caregivers during the patient’s admission in the ICUs. Moreover, we catalog “crisis” alarms that are “life-threatening” and indicative of progressive patient deterioration. The crisis alarms include asystole, ventricular fibrillation (VFib) and apnea, and prompt interventions are required when any of these crisis alarms are activated. In addition, we mark the technical alarms as they are device-related alarms without any clinical relevance. Table 4.1 shows examples of mapped monitor alarms extracted from UCLA and UCSF.

Table 4.1: Examples of consolidating raw monitor alarms extracted from UCLA and UCSF Medical Centers.

Raw Alarm Message	Alarm Type	Data Source	Consolidated Alarm Message	Alarm Numeric Value	Alarm Threshold	
					Low	High
ATRIAL FIB	Arrhythmia	UCLA	ATRIAL FIB	N/A	N/A	N/A
AFib or AFib(PaceMode1)		UCSF				
ACC VENT	Arrhythmia	UCLA	ACC VENT	N/A	N/A	N/A
ACC vent or ACC vent (PaceMode2)		UCSF				
V TACH	Arrhythmia	UCLA	V TACH	N/A	N/A	N/A
VTach or VTach(PaceMode1)		UCSF				
ART1 D HI 153	Parameter	UCLA	BP DIA HI	153	50	90
NBP D HI 110		UCLA		110		
ART Dia 113>90		UCSF		113		
Nbp Dia 102>90		UCSF		102		
HR LO 10	Parameter	UCLA	HR LO	10	50	130
HR 50=50		UCSF		50		
SpO2 LO 77	Parameter	UCLA	SPO2 LO	77	90	105
SpO2 81<90		UCSF		81		

4.2.4 Laboratory test mapping

Results of conventional clinical laboratory tests, marked with the timestamps when the samples were drawn, are extracted from the EHR systems for the patients included in this study. The basic data requirements for the laboratory test results include such items as test or procedure identifier (e.g., name of laboratory test), specimen type, reference range, unit of measure, and numeric-valued result. Unfortunately, problematic issues are encountered when attempting to aggregate the laboratory test results due to the different nomenclatures used for coding the similar data at the two institutions. One of the standard laboratory test identifier system is LOINC (logical observation identifiers names and codes) [116]. However, retrospective mapping local laboratory test names to the LOINC is labor-intensive [117]. Furthermore, some of the same laboratory tests have different units defined by the two institutions. Such a mismatch in unit of measure cannot be handled by simply using LOINC. Therefore, a laboratory test codebook comprised of uniform laboratory test names, units, and reference ranges is needed. It should be noted that for some of the same laboratory test, different gender has different reference range, which has been reflected in the laboratory test codebook.

By employing the laboratory test codebook, the names and units that were used differently by UCLA and UCSF Medical Centers for measuring the same laboratory tests are mapped and unified. In addition, non-numeric data are eliminated in places where numeric values of laboratory test results are expected. For example, signs such as ">" and "<" along with the values are removed, and texts such as "Neg" representing the negative values are converted into the corresponding mathematical form. Furthermore, the polarity of HI (high), LO (low) or NR (normal) for each of laboratory test results is determined by comparing its value against the corresponding reference range. Moreover, the laboratory test of the "Anion Gap" (AG) was not explicitly ordered for the UCLA patients, we therefore derive the AG (mmol/L) using the formula:

$$\text{AG} = \text{Na}^{+} - \text{Cl}^{-} - \text{HCO}_3^{-} \tag{4.1}$$

where Na^{+}, Cl^{-} and HCO_3^{-} represent clinical measurements of sodium (mmol/L), chloride (mmol/L) and arterial bicarbonate (mmol/L), respectively. To calculate the AG correctly, the same timestamps of Na^{+}, Cl^{-} and HCO_3^{-} being used are required. Table 4.2 lists examples of mapped laboratory test results extracted from UCLA and UCSF.

Table 4.2: Examples of consolidating raw laboratory test results extracted from UCLA and UCSF Medical Centers.

Raw Laboratory Test Name	Data Source	Raw Unit	Raw Result	Mapped Laboratory Test Name	Mapped Unit	Mapped Numeric Value Result	Mapped Reference Range	Polarity
GLUCOSE, POC	UCLA	mg/dL	>600	GLUCOSE	mg/dL	600	[70, 125]	HI
HEMOGLOBIN, POC	UCLA	g/dL	11.1	HEMOGLOBIN	g/dL	11.1	Male: [13.8, 17.2] Female: [12.1, 15.1]	LO
MAGNESIUM, PLASMA	UCLA	mEq/L	1.6	MAGNESIUM	mg/dL	1.95	[1.7, 2.2]	NR
BASE EXCESS	UCSF	mmol/L	Neg 4.0	VENOUS BASE EXCESS	mmol/L	-4	[-2, 2]	LO
LACTATE, PLASMA	UCSF	mmol/L	0.8	LACTATE	mg/dL	7.21	[4.5, 19.8]	NR
N/A	UCLA	N/A	N/A	ANION GAP	mmol/L	32.5	[8, 16]	HI

4.2.5 Physiologic waveforms and vital signs

Physiologic waveforms and vital sign measurements generated by the GE bedside monitors are continuously archived into a central repository by the BedMaster systems and saved into flat files using a proprietary format (we call them "STP" files according to the file extensions). These STP files are not accessible without BedMaster software and hence cannot be readily utilized by researchers to perform further studies such as physiological signal processing. As manual conversion of a large amount of STP files is prohibitively expensive and time-consuming, it is helpful to explore an automated method for extracting timestamped waveforms and vital signs from the STP files and save them following a new format so that they can be easily used by researchers. We therefore develop an application, called "Stp2Bin". The Stp2Bin is a C#-based software application that extracts waveforms and vital signs from the STP files in an automated manner. It uses the publicly available format from AD Instrument (Dunedin, New Zealand) and a self-defined format (Table 4.3) to save waveforms and vital signs into binary files, respectively.

The Stp2Bin application provides two modes to accept the user's specific input:

Table 4.3: The format for saving the vital sign measurements into binary files.

File Header	
Type	*Name*
char[16]	Name of the vital sign
char[8]	UOM
char[8]	Unit
char[4]	Bed
int	Start year
int	Start month
int	Start day
int	Start hour
int	Start minute
double	Start second
Data Body: repeat the following for the vital sign	
Type	*Name*
double	Vital sign value
double	Offset to the start time in seconds
double	Alarm low limit
double	Alarm high limit

the "Patient MRN" mode and the "STP File" mode. The difference between the two modes is that the "Patient MRN" mode has additional procedures to find proper STP files based on the user-specified patient MRNs. In this mode the application accepts a list of patient MRNs and looks up the ADT table to query the locations that each patient has stayed at (e.g., the time of transfer in (ADT_IN) and the time of transfer out (ADT_OUT) for a given ICU bed), which are then used to locate the corresponding STP files in the central repository. The reason that we use the patient's bed location and the stay duration to locate the STP files rather than directly comparing the user-specified MRN with the MRNs recorded in the BedMaster system is because errors exist in the manually supplied MRNs in patient monitors. For example, the nurse may forget to modify the MRN when a new patient is admitted to a bed. In this case, the BedMaster system will store the new patient's data under an improper MRN (i.e., the MRN of the previous admitted patient). Therefore, the MRNs recorded by the BedMaster system are not reliable and the STP files will be located incorrectly if we directly use the MRNs in the BedMaster system. Fortunately, the BedMaster system offers an table that records additional information about each STP file such as bed location, the start and end of recording times (STP_START and STP_END). We provide a method for locating STP file by matching the patient's ADT information and the STP file information. In particular, the proper STP file is determined by detecting whether there has the time overlapping between the stay duration from the ADT table and the STP recording interval from the BedMaster table for the given bed location. Figure 4.2 depicts the four circumstances under which the proper STP file can be located using the "Patient MRN" mode when running the Stp2Bin application. It should be noted that the use of "Patient MRN" mode requires pre-authorization to access the database because this process will access to the local database where protected health information is saved.

With the STP files located and available, the core module of the Stp2Bin application is running to generate binary files for waveforms and vital signs. The module first calls the software utility provided by BedMaster system to extract the waveforms and vital signs from the STP file and save them into XML (extensible markup language) file. The module further parses the XML file and calibrates the data (e.g., detection of the gaps in the data stream due to measurement error). Finally, all of the available channels of the physiologic waveforms are assembled into a binary file. The module is capable of detecting changes in the configuration of channels (e.g., increase or decrease in the number of channels or the monitored physiologic variables). In this case, a new waveform binary files will be generated as long as any of such changes are detected. All

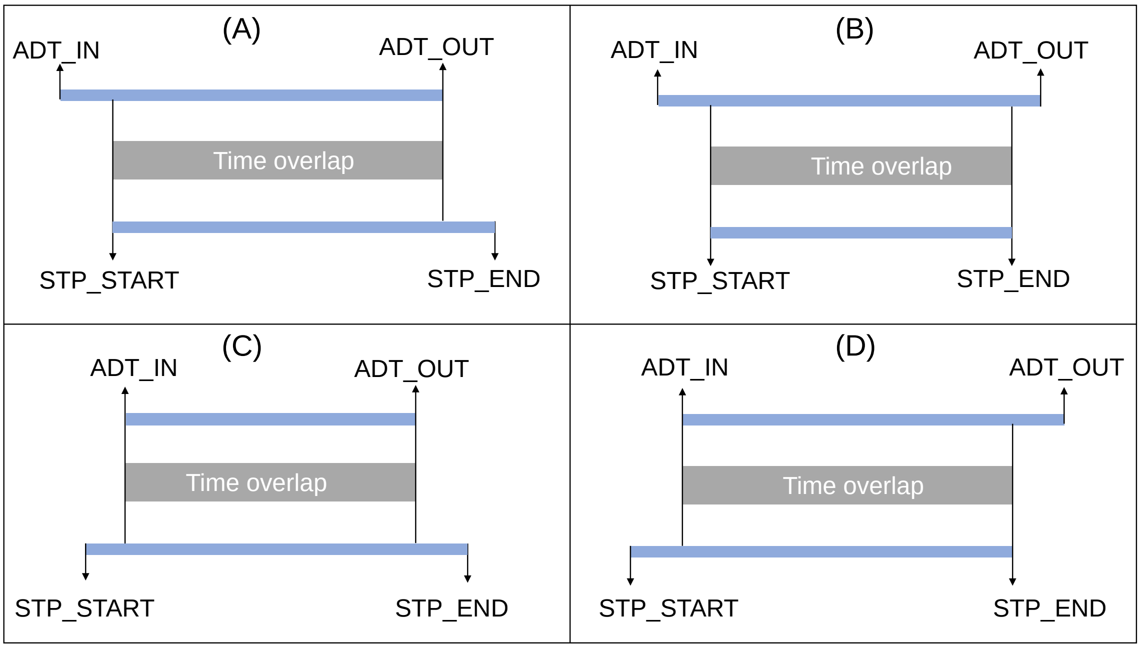


Figure 4.2: Location of proper STP files under the four circumstances. (A) ADT_IN $\leq$ STP_START and STP_START $\leq$ ADT_OUT $\leq$ STP_END ; (B) ADT_IN $\leq$ STP_START and STP_END $\leq$ ADT_OUT ; (C) STP_START $\leq$ ADT_IN and ADT_OUT $\leq$ STP_END ; (D) STP_START $\leq$ ADT_IN $\leq$ STP_END and STP_END $\leq$ ADT_OUT.

of the physiologic waveforms are saved at a sampling rate of 240 Hz. In the meanwhile, all the available vital signs are saved into separate binary files at a sampling rate of 1/2 Hz. The module matches the timestamps between physiologic waveforms and vital signs. The protected health information (PHI) is also removed from the binary file for the purpose of de-identification. All of the processes aforementioned are performed in a parallel manner to speed up the data extraction.

4.3 Results

4.3.1 Patient characteristics

Adult patients (age >18 years) included in this study were admitted to the UCLA ICUs between January 2010 to June 2014 and the UCSF ICUs between March 2013 to March 2015. Quality management services at UCLA Medical Center and UCSF Medical Center provided a listing of the patients with at least one code blue call during the admission, respectively. Control patients without any code blue call or unplanned ICU transfer were selected for each of coded patients under the following additional criteria: same age (± 5 years) and same gender; and admission to the same hospital unit within the same month.

Table 4.4 shows characteristics of the patients included in the SuperAlarm study database II. For UCLA Medical Center, there are a total of 403 coded patients [57.8% male, age at admission: 62.6 ± 16.5, average monitor duration: 23.3 days (median: 13.0, IQR: 4.3 to 27.2)] and 4667 controls patients [62.6% male, age at admission: 63.1 ± 14.0, average monitor duration: 12.8 days (median: 6.4, IQR: 2.8 to 13.1)].

For UCSF Medical Center, there are 152 coded patients (54.6% male, age at admission: 61.6 ± 15.2, average monitor duration: 13.2 days (median: 7.6, IQR: 2.5 to 18.4)] and 1115 control patients (63.7% male, age at admission: 62.8 ± 11.0, average monitor duration: 7.0 days (median: 3.4, IQR: 1.7 to 7.2)].

4.3.2 Alarms

Alarms including monitor alarms and technical alarms are extracted from both coded patients and control patients admitted to UCLA and UCSF Medical Centers. A total of 106 distinct alarms are included in the SuperAlarm study database II (Table 4.5). It should be noted that only those alarms that precede the first code blue call are collected in the cases where the patients had multiple code blue calls.

Table 4.4: Patient characteristics in the SuperAlarm study database.

	UCLA		UCSF	
	Coded Patients	Control Patients	Coded Patients	Control Patients
Total Number	403 †	4667	152 ‡	1115
Age at Admission(std)	62.6(16.5)	63.1(14.0)	61.6(15.2)	62.8(11.0)
Average Monitor Duration, days (median, IQR §)	23.3 (13.0, 4.3 to 27.2)	12.8 (6.4, 2.8 to 13.1)	13.2 (7.6, 2.5 to 18.4)	7.0 (3.4, 1.7 to 7.2)
Gender				
Female(%)	170(42.18)	1747(37.43)	69(45.39)	405(36.32)
Male(%)	233(57.82)	2920(62.57)	83(54.61)	710(63.68)
Ethnicity				
Hispanic or Latino(%)	80(19.85)	797(17.08)	18(11.84)	147(13.18)
Not Hispanic or Latino(%)	322(79.90)	3868(82.88)	121(79.61)	905(81.17)
Unknown(%)	1(0.25)	2(0.04)	13(8.55)	63(5.65)
Race				
White or Caucasian(%)	291(72.21)	3437(73.65)	72(47.37)	623(55.88)
Black or African American(%)	38(9.43)	492(10.54)	14(9.21)	98(8.79)
Asian(%)	35(8.68)	411(8.81)	25(16.45)	140(12.55)
Other(%)	34(8.44)	324(6.94)	28(18.42)	192(17.22)
Unknown(%)	5(1.24)	3(0.06)	13(8.55)	62(5.56)

† 62 UCLA coded patients with more than one code blue call.

‡ 26 UCSF coded patients with more than one code blue call.

§ IQR: interquartile range.

4.3.2.1 Crisis alarms and technical alarms

We briefly report crisis alarms and technical alarms as they are not utilized in this dissertation. Three distinct crisis alarm and 15 distinct technical alarms are collected in the SuperAlarm study database after being mapped (see Table 4.5).

For UCLA patients, there are 36044 crisis alarms from coded patients (average: 8.3 per patient per day, median: 1.4, IQR: 0.3 to 5.5) while 132239 from control patients (average: 2.4 per patient per day, median: 0.3, IQR: 0.0 to 1.7). Moreover, a total of 275520 technical alarms are from coded patients (average: 48.9 per patient per day, median: 24.0, IQR: 13.9 to 43.3) and 1623761 from control patients (average: 27.0 per

Table 4.5: Alarms available in the SuperAlarm study database.

ASYSTOLE †	VFIB/VTAC †	APNEA †	V TACH
VT >2	V BRADY	BRADY	ATRIAL FIB
R ON T	ACC VENT	TACHY	PAUSE
BIGEMINY	IRREGULAR	COUPLET	TRIGEMINY
CHECK ADAPTER	PVC	NO BREATH	HR HI
HR LO	PVC HI	BP DIA HI	BP MEAN HI
BP SYS HI	BP DIA LO	BP MEAN LO	BP SYS LO
RESP HI	RESP LO	SPO2 HI	SPO2 LO
SPO2 RATE LO	SPO2 RATE HI	ST-AVF HI	ST-AVF LO
ST-AVL HI	ST-AVL LO	ST-AVR HI	ST-AVR LO
ST-I HI	ST-I LO	ST-II HI	ST-II LO
ST-III HI	ST-III LO	ST-V1 HI	ST-V1 LO
ST-V2 HI	ST-V2 LO	ST-V3 HI	ST-V3 LO
ST-V4 HI	ST-V4 LO	ST-V5 HI	ST-V5 LO
ST-V6 HI	ST-V6 LO	ST-V HI	ST-V LO
ST-dV2 HI	ST-dV2 LO	ST-dV3 HI	ST-dV3 LO
ST-dV4 HI	ST-dV4 LO	ST-dV6L HI	ST-dV6L LO
CVP MEAN HI	CVP MEAN LO	ICP MEAN HI	ICP MEAN LO
LAP MEAN HI	LAP MEAN LO	RAP MEAN HI	RAP MEAN LO
PA DIA HI	PA DIA LO	PA MEAN HI	PA MEAN LO
PA SYS HI	PA SYS LO	SP MEAN HI	SP MEAN LO
BP RATE HI	BP RATE LO	CO2 RSP HI	CO2 RSP LO
EXP CO2 HI	EXP CO2 LO	INSP CO2 HI	ARTIFACT ‡
ARRHY SUSPEND ‡	ECG LEADS FAIL ‡	BP FAIL ‡	RR LEADS FAIL ‡
SPO2 FAIL ‡	NO TELEMETRY ‡	NO ECG ‡	CVP DISCONNECT ‡
ICP DISCONNECT ‡	LAP DISCONNECT ‡	RAP DISCONNECT ‡	SENSOR FAIL ‡
SP DISCONNECT ‡	PA DISCONNECT ‡		

† “crisis” alarms defined in the database.

‡ technical alarms defined in the database.

patient per day, median: 17.3, IQR: 10.1 to 30.7).

For UCSF patients, a total of 30655 crisis alarms are from coded patients (average: 21.3 per patient per day, median: 9.6, IQR: 4.4 to 25.0) and 101645 are from control patients (average: 13.3 per patient per day, median: 4.7, IQR: 1.6 to 12.8). For technical alarms, 499369 are from coded patients (average: 324.5 per patient per day, median: 227.0, IQR: 130.4 to 414.2) while 2212903 are from control patients (average 299.9 per patient per day, median: 220.7, IQR: 128.3 to 382.9).

4.3.2.2 Monitor alarms

After exclusion of crisis alarms and technical alarms, 88 distinct monitor alarms (14 arrhythmia alarms and 74 vital sign parameter alarms) are included in this study. Figure 4.3 displays the distributions of monitor alarms extracted from coded patients and controls patients in the two institutions, respectively.

For UCLA patients, a total of 1566133 monitor alarms (average: 278.4 per patient per day, median: 158.6, IQR: 93.5 to 281.9) preceding code blue events are extracted from coded patients. Meanwhile, a total of 7341089 monitor alarms (average: 124.1 per patient per day, median: 83.0, IQR: 37.0 to 152.5) are extracted from control patients. The Wilcoxon Rank-Sum test result shows that the number of monitor alarms between the coded patients and control patients is significant different ($p \ll 0.01$).

For UCSF patients, 1815160 monitor alarms (average: 1107.2 per patient per day, median: 670.4, IQR: 376.7 to 1349.3) preceding code blue events are extracted from coded patients. And a total of 5901531 (average: 689.7 per patient per day, median: 363.7, IQR: 211.2 to 670.0) monitor alarms are extracted from control patients (The Wilcoxon Rank-Sum test: $p \ll 0.01$).

The Wilcoxon Rank-Sum test result also shows that the number of monitor alarms from the UCLA patients and UCSF patients is significant different ($p \ll 0.01$).

Figure 4.4 shows the top 20 frequent monitor alarms from coded patients and con-

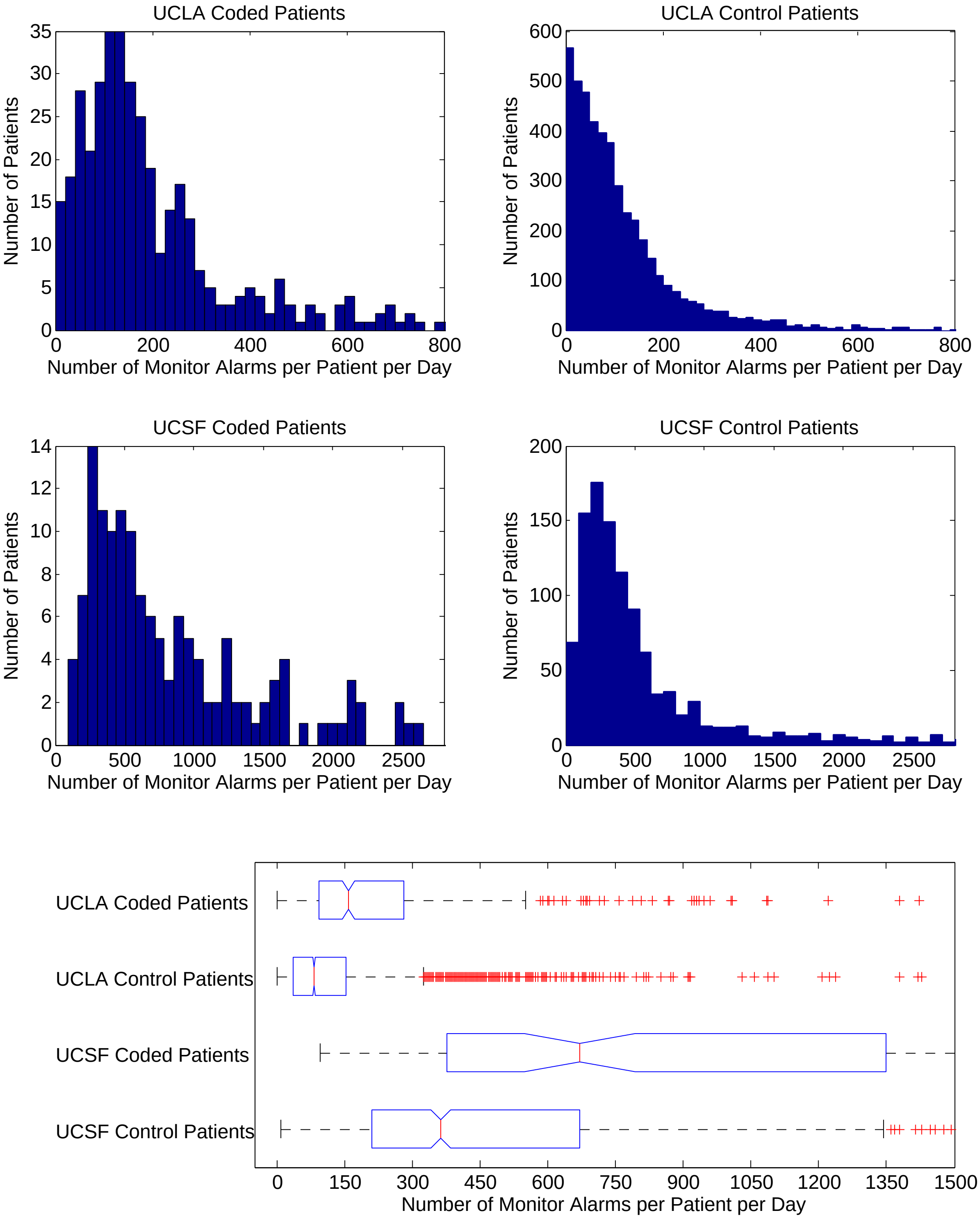


Figure 4.3: The distributions of monitor alarms collected from coded patients and control patients.

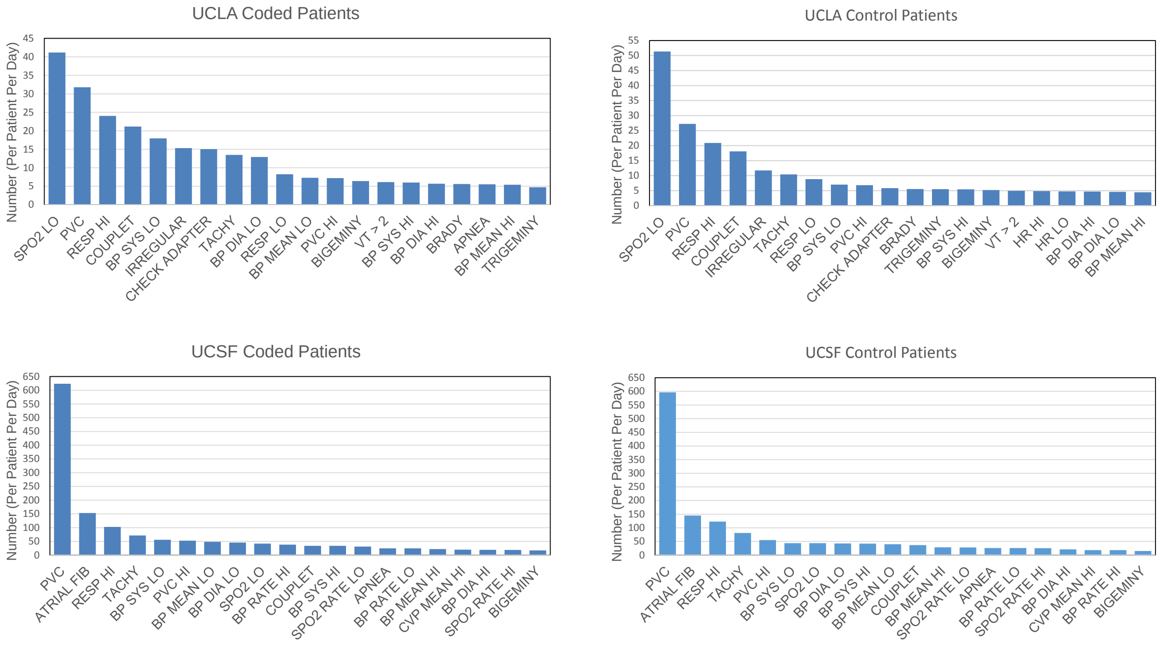


Figure 4.4: The top 20 frequent monitor alarms from coded patients and control patients.

trols patients in the two institutions, respectively. It is observed that the first four most frequent monitor alarms (per patient per day) for coded patients and control patients from UCLA Medical Center are $\mathrm{SpO_2}$ LO (41.2 vs 51.3), PVC(31.8 vs 27.3), RESP (Respiration Rate) HI (24.0 vs 20.9), and Couplet (21.2 vs 18.1), respectively. Whereas, the first four are PVC (633.4 vs 596.6), Atrial Fib (153.2 vs 145.2), RESP HI (102.5 vs 122.9), and Tachy (71.7 vs 81.1) for coded patients and control patients from UCSF Medical Center, respectively. We can also see that the monitor alarm of PVC has the dominant frequency among coded patients and control patient in the two institutions. The monitor alarm of “Atrial Fib” from the UCSF Medical Center occurs far more frequently than that from the UCLA Medical Center.

4.3.3 Laboratory test results

A total of 58 distinct laboratory tests that are routinely ordered by clinicians are included and mapped in the SuperAlarm study database (Table 4.6).

Figure 4.5 illustrates the distributions of laboratory test results collected from coded patients and controls patients at UCLA and UCSF Medical Centers, respectively. For UCLA patients, 534734 (average count 110.7 per patient per day, median 74.0, IQR 45.1 to 125.1) and 2590318 (average count 78.2 per patient per day, median 42.3, IQR 29.6 to 62.8) laboratory test results are extracted from coded patients and control patients, respectively. The Wilcoxon Rank-Sum test result shows the number of laboratory test results between the coded patients and control patients is significant different ($p \ll 0.01$).

For UCSF patients, a total of 216265 (average count 127.3 per patient per day, median 85.3, IQR 42.4 to 161.5) and 847730 (average count 95.7 per patient per day, median 56.6, IQR 32.6 to 115.2) laboratory test results are extracted from coded patients and control patients, respectively ($p \ll 0.01$).

The Wilcoxon Rank-Sum test results also demonstrate that the number of laboratory test results is not significantly different for coded patients between the two institutions ($p = 0.3310$) while the difference of the number of laboratory test results for control patients between the two institutions is significant ($p \ll 0.01$).

Figure 4.6 shows the distributions of of polarities of high (HI), normal (NR) and low (LO) that are obtained by comparing numeric values of laboratory test results against the corresponding reference ranges for coded patients and control patients at the UCLA and UCSF Medical Centers. We can see that laboratory tests with the normal (NR) results have the highest frequency among coded patients and control patients from the two institutions.

Table 4.6: Laboratory tests available in the SuperAlarm study database.

ARTERIAL BASE EXCESS	ARTERIAL BICARBONATE
ARTERIAL PCO2	ARTERIAL PO2
ARTERIAL O2SAT	ARTERIAL PH
AMMONIA	AMYLASE
BNP	HEMATOCRIT
HEMOGLOBIN	PLATELET COUNT
WBC	RBC
ABSOLUTE EOS COUNT	CORRECTED IONIZED CALCIUM
MAGNESIUM	PHOSPHORUS
BUN	CHLORIDE
TOTAL CO2	CREATININE
GLUCOSE	POTASSIUM
SODIUM	INR
PROTHROMBIN TIME	APTT
GFR EST. FOR AFRICAN AMERICAN	GFR EST. FOR NON-AFRICAN AMERIC
TACROLIMUS (FK-506)	SIROLIMUS
URINE TOTAL PROT/CREAT RATIO	LIPASE
ALBUMIN	ALKALINE PHOSPHATASE
ALT (SGPT)	AST (SGOT)
CONJUGATED BILIRUBIN	TOTAL BILIRUBIN
LACTATE DEHYDROGENASE	TOTAL PROTEIN
CK-MB	TOTAL CK
TROPONIN I	CSF WBC
CSF GLUCOSE	CSF PROTEIN
PHENOBARBITAL	VENOUS BASE EXCESS
VENOUS BICARBONATE	VENOUS PCO2
VENOUS PO2	VENOUS O2SAT
VENOUS PH	CALCIUM
LACTATE	ANION GAP

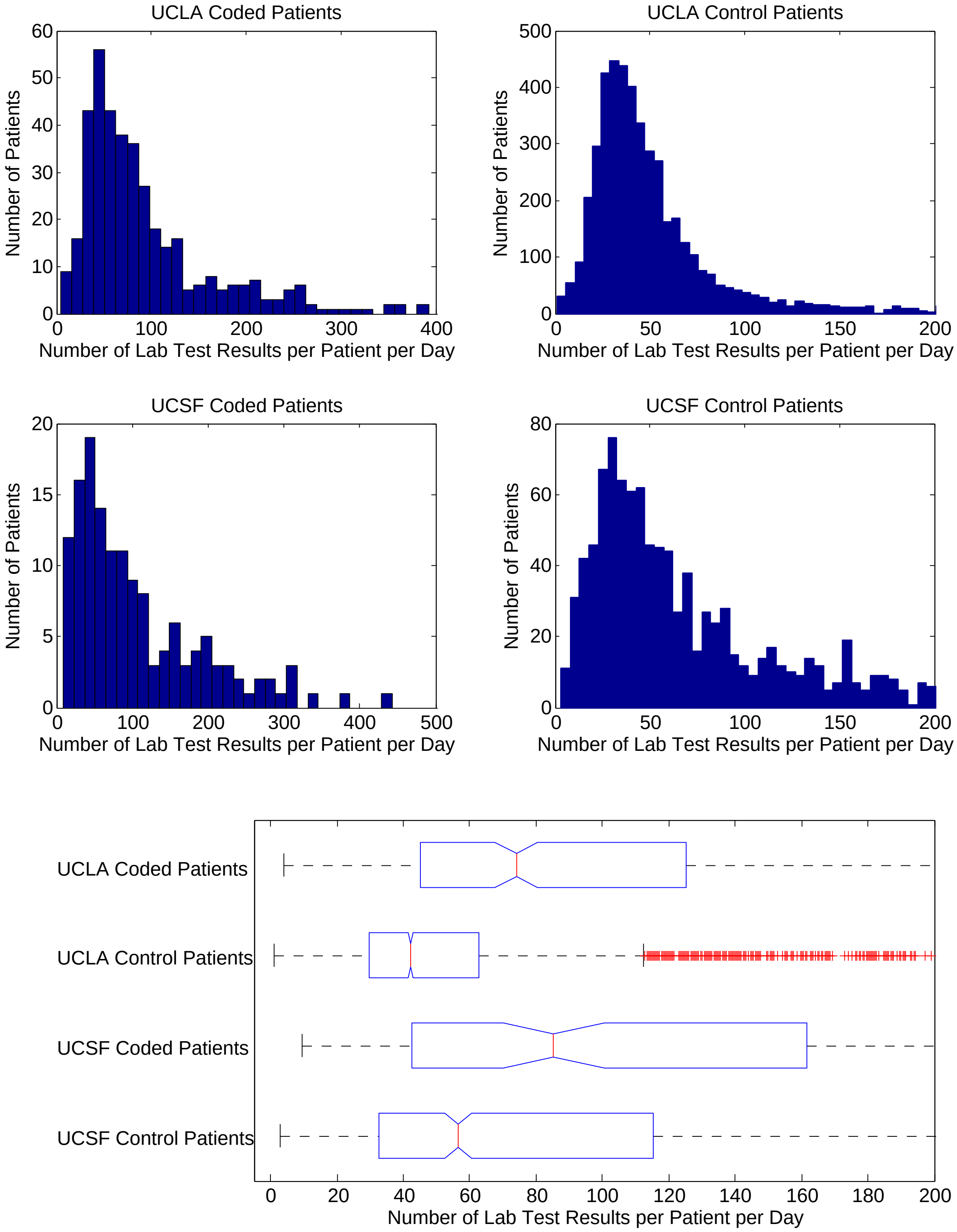


Figure 4.5: The distributions of laboratory test results collected from coded patients and control patients.

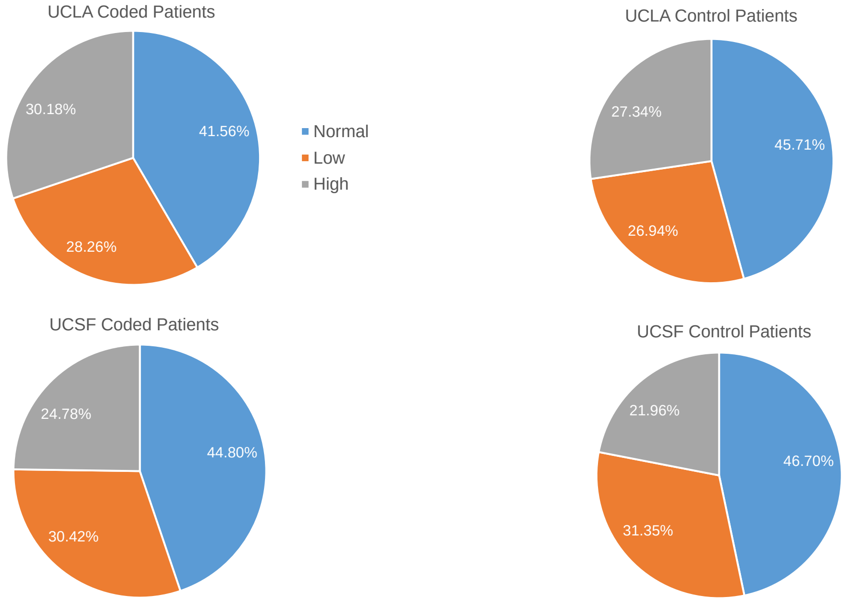


Figure 4.6: The distributions of high (HI), normal (NR) and low (LO) laboratory test results from coded patients and control patients.

4.3.4 Physiologic waveforms and vital signs

By running the Stp2Bin application, physiologic waveforms and vital signs with timestamps are extracted and saved into the binary files that can be readily used for further analysis. As physiological signal process is beyond the scope of this dissertation, we only show some examples of those waveforms and vital signs that are available in the SuperAlarm study database as follows.

Figure 4.7 displays ECG and ABP waveforms in a 15-second window preceding and after a VTach alarm. Here we plot the ECG waveforms of leads I, II, III and V. In addition to ECG and ABP, other physiologic waveforms are also available in the SuperAlarm study database such as SpO_2 and Respiration Rate (RR). We can see that the morphological characteristics of those waveforms exhibit great deviation in approximately 5 seconds before the VTach alarm.

Table 4.7 lists the available vital signs in the SuperAlarm study database. Figure

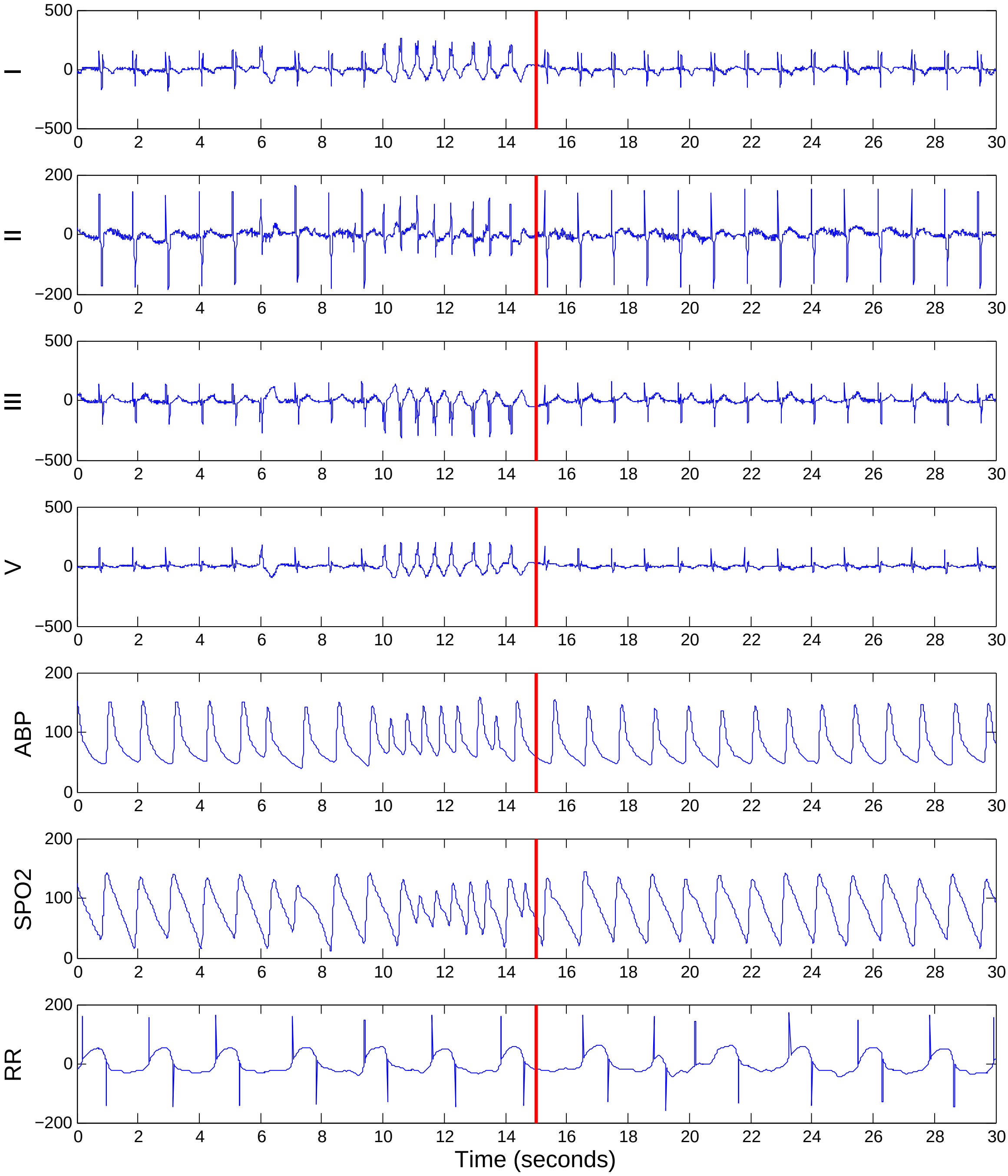


Figure 4.7: An example of ECG and ABP waveforms (sampling rate = 240 Hz) in a 15-second window preceding and after a VTach alarm. The vertical red lines represent the timestamp when the V Tach alarm occurred.

4.8 further demonstrates an example of time series of vital signs (sampling rate = 1/2 Hz) in a 30-minute window preceding and after the same VTach alarm as shown in Figure 4.7.

Table 4.7: Vital signs available in the SuperAlarm study database.

Heart Rate(HR)	Respiration Rate (RR)
Oxygen Saturation (SpO2)	Oxygen Saturation Rate (SpO2 Rate)
Arterial Blood Presure (Systolic, Diastolic, Mean)	Non-invasive Blood Pressure (Systolic, Diastolic, Mean)
Femoral(Systolic, Diastolic, Mean)	Temperature
Pulmonary Artery Pressure (Systolic, Diastolic, Mean)	ST (ECG lead I, II, III, V1, V2, V3)
Premature Ventricular Contraction(PVC)	CUFF
Central Venous Pressure(CVP)	Cerebral Perfusion Pressure (CPP)
Intracranial pressure (ICP)	Cardiac Output(CO)
Pulmonary Artery Wedge Pressure(PAWP)	Pulse Rate
Left Atrial Pressure (LAP)	Right atrial pressure (RAP)

4.4 Discussion

We have developed a large-scale, comprehensive research database for the SuperAlarm study, called the SuperAlarm study database II. The database aggregates and consolidates monitor alarms, laboratory test results, physiologic waveforms and vital signs that are obtained from coded and control patients admitted to UCLA and UCSF Medical Centers. We provide two naming codebooks for mapping monitor alarms and laboratory tests extracted from the two institutions, respectively. We develop a software application to convert flat files that were saved by the BedMaster system into binary files so that the vital signs and the high resolution waveforms can be readily reviewed and utilized for further analysis. The protected health information (PHI) has been de-identified in compliance with the Health Insurance Portability and Accountability Act (HIPAA). All these processes are performed in an automated and parallel manner. This database supports further research to develop and evaluate new SuperAlarm algorithms.

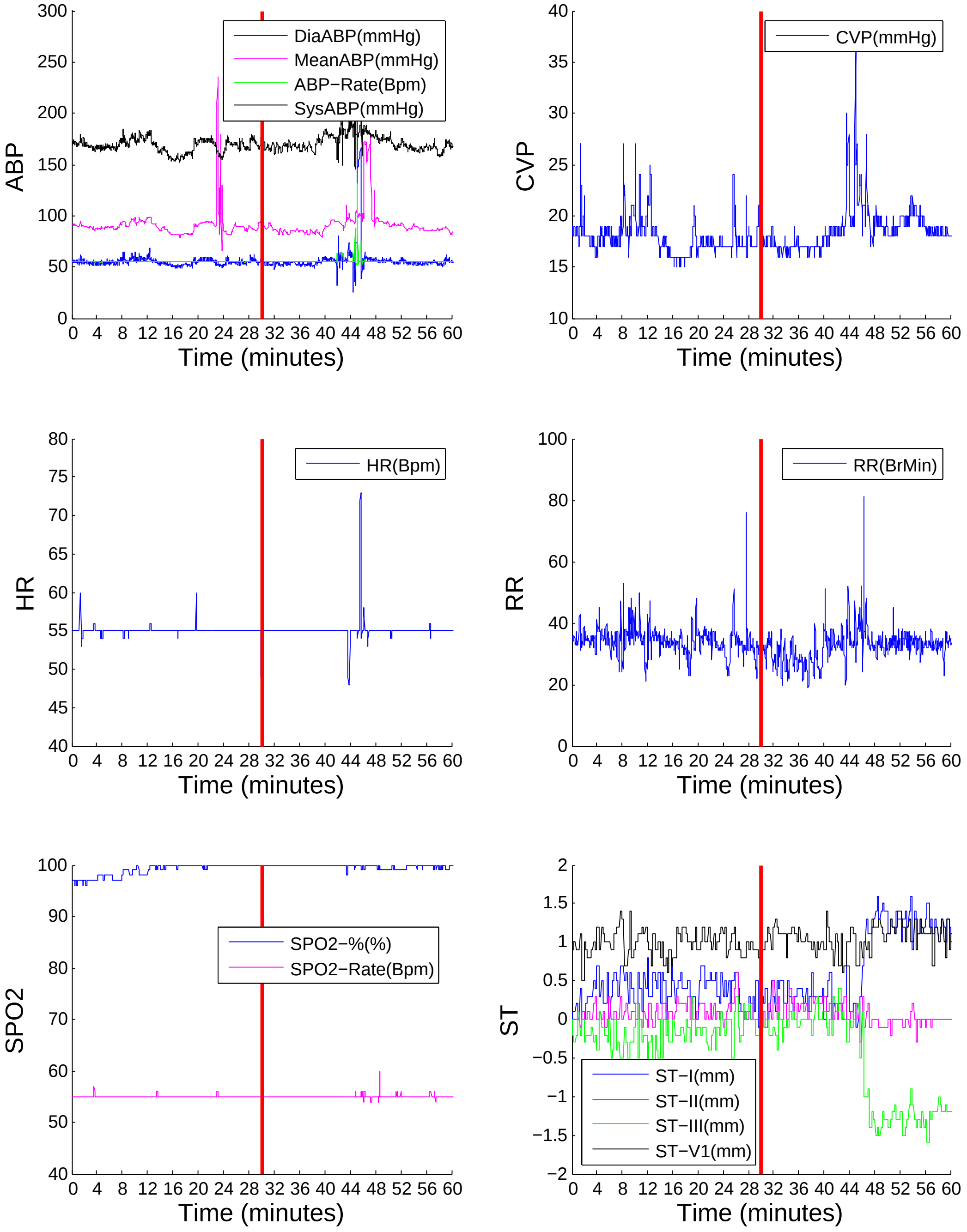


Figure 4.8: An example of physiologic vital signs (sampling rate = 1/2 Hz) in a 30-minute window preceding and after the same VTach alarm from the same patient. The vertical red lines represent the timestamp when the VTach alarm occurred.

The diversity of nomenclatures of monitor alarms and laboratory tests complicates large-scale data integration between different institutions. Even with the standard nomenclature such as LOINC, manual mapping of the large volume of data is needed but prohibitively labor-intensive. To allow evaluation of SuperAlarm algorithms on the basis of the comprehensive data collected from the two institutions, we provide two naming codebooks for automated mapping of monitor alarms and laboratory tests, respectively.

Comprehensive monitor alarms and laboratory test results are collected from the two institutions and merged into the SuperAlarm study database II with the aid of the proposed alarm naming codebooks. The significantly different number of monitor alarms and laboratory test results between coded patient and control patients suggests that coded patients may have had physiologic abnormalities more often than control patients and also may have had more diagnostic laboratory tests. The reason for the significantly different number of monitor alarms between the two institutions is because that different equipment was used for acquiring data from GE monitors at UCSF to allow the BedMaster system not only collect audible alarms but also message-level ones(inaudible) that just are displayed on the bedside monitors as text messages. It is believed that the inaudible alarms also distract nurses [1].

Physiologic waveforms are saved into the binary files following the publicly available format from AD Instrument [118]. Furthermore, these files can be opened and reviewed using a free software application “LabChart Reader” from AD Instrument. In addition, the binary files can also be readily available for further analysis by programmatically loading them into analytics computation platforms (e.g., MATLAB) and exploring advanced algorithms to develop the metrics that characterize a patient’s physiologic status. Although physiological signal processing is beyond the scope of this dissertation, the use of extracted waveforms has led to several studies, including reduction of false alarm [50], development of non-monitored metrics (e.g., R-R interval) to predict bradycardiac arrest [119] and prediction of outcome of external ventricular drainage (EVD) weaning trial using intracranial pressure (ICP) pulses [120].

4.5 Conclusion

We have reported the development of a large-scale, comprehensive research database for supporting the development and validation of SuperAlarm. The SuperAlarm study database II contains patient demographics, monitor alarms, laboratory test results, physiologic waveforms and vital signs extracted from coded patients and control patients admitted to UCLA and UCSF Medical Centers. The protected health information is de-identified in compliance with HIPAA. Two naming codebooks for automated mapping of monitor alarms and laboratory tests are designed. We have developed an application to extract waveforms and vital signs, and save them into binary files so that they can be readily available for researchers for further analysis. The establishment of such database enables researchers to develop and evaluate advanced and robust SuperAlarm algorithms based on large amount of real patient data.

CHAPTER 5

Prediction of patient deterioration using SuperAlarm sequence: a time weighted supervised sequence representation method

In Chapter 3, we have developed a sequence classifier to recognize temporal patterns in SuperAlarm sequences so as to predict code blue events and reduce alarm burden, and demonstrated the improved performance in comparison with the use of raw monitor alarm sequences. In particular, we employed the term frequency inverse document frequency (TFIDF) representation method to convert the sequences into fixed-dimensional vectors. The simplicity of TFIDF representation method leads to limitations. For example, TFIDF weights a SuperAlarm trigger simply by computing its frequency regardless where it occurs within a SuperAlarm sequence. It is intuitive that a SuperAlarm trigger should carry more weight to measure its importance when it approaches the endpoint (e.g., code blue event). To overcome such limitations, in this Chapter we propose a novel representation method to convert SuperAlarm sequences into fixed-dimensional vectors, called time weighted supervised sequence representation (TWSSR). The TWSSR is not only a supervised weighting scheme that takes into account the distribution of sequences between coded patients and control patients, it also considers the impact of time on the weight of a SuperAlarm trigger that occurs in a SuperAlarm sequence. Monitor alarms and laboratory test results in the SuperAlarm study database II as described in Chapter 4 are used to mine SuperAlarm patterns and further generate the SuperAlarm sequences. Support vector machine based recursive feature elimination (SVM-RFE) algorithm is applied to perform classification in conjunction with feature selection. The results demonstrate that the performance of the sequence classifier based on the TWSSR representation method is higher than that based on TFIDF method.

5.1 Introduction

Critically ill patents in intensive care units (ICUs) are surrounded by an impressive array of multi-parameter, alarm-equipped physiologic monitors that are deployed to alert caregivers to hemodynamic instabilities and facilitate prompt therapeutic interventions to prevent unexpected adverse events. However, the proliferation of these sophisticated technologies plagues caregivers with a myriad of false alarms and nuisance (false positive) alarms, leading to alarm fatigue that can detrimentally affect patient safety [1, 13, 19]. Numerous efforts have been directed towards addressing the alarm fatigue problem by reducing false alarms through the secondary analysis of physiologic waveforms that are related to specific alarms [47, 49, 52, 53], and suppressing nuisance alarms via optimizing alarm settings [19, 121, 122] or introducing alarm delays [123]. Despite some potentials, most of these approaches typically focus on suppression of individual alarms, and thereby further studies are required because certain patterns such as increasing frequency of changes in physiological variables may be signs and indicators of some adverse events [91].

On the other hand, to assist caregivers in early recognition of patient deterioration in a data fusion manner, a multitude of "track and trigger" systems (TTSs) have been developed [55, 56]. However, most of the TTSs deriving patient severity scores from vital signs alone are inadequate to detect patient deterioration appropriately [61]. Other studies have also been conducted to identify patient deterioration by leveraging additional clinical data available in the electronic health record (EHR) system such as laboratory test results [66,84]. Nevertheless, these approaches focus on identification of patient deterioration without offering direct relief of the existing alarm fatigue problem.

In analysis of the context of patient monitoring practice, researchers have pointed out that future monitoring systems should recognize clinical patterns in a data fusion manner by integrating data available from all patient-linked devices [34]. This concept indeed supports the approach we are developing [74, 92] to improve patient monitoring by using "SuperAlarm" patterns. The SuperAlarm pattern [74] was originally defined as a superset of patient monitor alarms that co-occurred in a time window immediately preceding "code blue" events for more than a minimal percentage of coded patients but for less than a maximal percentage of control patients. In a subsequent work [92], we further enriched the SuperAlarm patterns by integrating laboratory test results from the EHR system with monitor alarms and demonstrated the improvement of performance in prediction of code blue events and reduction of monitor alarm frequency. These SuperAlarm patterns can be deployed to monitor patients and a detection of an emerging SuperAlarm pattern in data streams is termed a SuperAlarm trigger. A sequence of consecutive triggers over time is termed a SuperAlarm sequence.

Advancing these endeavors, in this study we seek to predict code blue events using SuperAlarm sequences. The exploitation of method for representing sequences is of importance because it offers the opportunity to use off-the-shelf machine learning approaches to recognize temporal patterns encoded by these sequences. As reported in our preliminary study [93], representing a SuperAlarm sequence as a fixed-dimensional vector and subsequently classifying this vector has the potential to further reduce SuperAlarm frequency without compromising the sensitivity. Therefore, in the present work we develop a novel sequence representation method, which is called time weighted supervised sequence representation (TWSSR). Compared to the approach in our preliminary study [93] that weights the temporal closeness of a SuperAlarm trigger to the current time when counting the occurrence rate of this trigger in the sequence, TWSSR further considers the importance of the SuperAlarm patterns between coded patients and control patients so that a greater weight is assigned to SuperAlarm patterns with higher occurrence rate among coded patients but lower rate among control patients.

Furthermore, the present study significantly expand, by integrating data from both UCLA and UCSF Medical Centers, the database for training and evaluating the proposed approach as compared to the database used in our prior studies [74, 92, 93]. As a baseline sequence representation approach, we compare the prediction performance in use of TWSSR method with that using term frequency inverse document frequency (TFIDF) method, a well-known representation scheme that is widely used in community of information retrieval [94].

5.2 Methods

Figure 5.1 illustrates the flowchart of the proposed framework to predict code blue events. It is composed of two major steps: (1) offline training, where the optimal values for all algorithm parameters used in this study are determined based on a training dataset; (2) online testing, where the performances in prediction of code blue events are evaluated using an independent test dataset. The processing blocks depicted in Figure 5.1 are described in details as follows.

5.2.1 Data source

This retrospective study uses the monitor alarms and laboratory test results stored in the SuperAlarm study database II that contains the data from coded patients and control patients admitted to University of California, Los Angeles (UCLA) Ronald Regan Medical Center and University of California, San Francisco (UCSF) Medical Center. The ICUs and acute care areas in the two institutions are equipped with bedside monitors (GE Healthcare, Milwaukee, WI) to acquire and process patient physiological data. Continuous physiological waveforms, vital signs and monitor alarms are archived by BedMasterEx system (Excel Medical Electronics, Jupiter, FL) into a central repository at each institution. Timestamped monitor alarms were extracted from the SQL server database deployed along with the BedMasterEx system. Laboratory test results

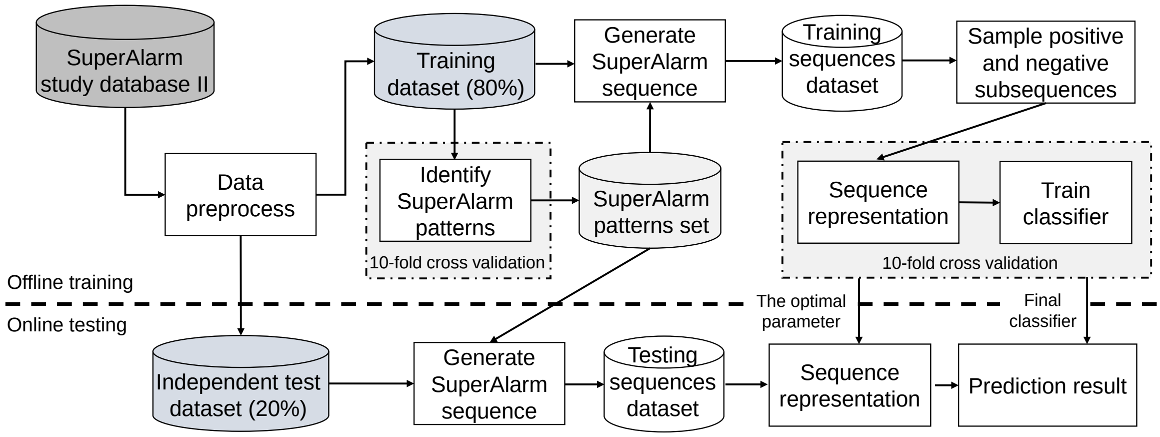


Figure 5.1: Flowchart of the proposed SuperAlarm sequence classification approach to predict code blue event events.

with taken timestamps were collected from the EHR systems implemented in the two institutions, respectively. The diversity in these data such as the disparate names of monitor alarms and different units of measure for the same laboratory tests between the two institutions was further unified by the SuperAlarm study database II using mapping codebooks. The reader can refer to a detailed description of the SuperAlarm study database II in Chapter 4.

5.2.2 Data preprocessing

The data preprocessing includes two steps as we did in our previous study [92]. The first step is to exclude the technical alarms and the “crisis” alarms. The second step is to exclude coded patients with abnormally small number of monitor alarms in a user-specified T_w-long time window preceding code blue events using the Non-Homogenous Poisson Process (NHPP) model. The NHPP model estimates the amount of monitor alarms occurring in the T_w-long time window and the minimum alarm count threshold is determined in terms of the floor value of 5% quantile of the estimated alarm count in this window. Patients with fewer monitor alarms within the T_w-long time window prior to the code blue event than this threshold will be excluded.

5.2.3 Identification of SuperAlarm patterns

Here we briefly introduce the processes of identifying SuperAlarm patterns because we apply the same framework as we did in our previous study [92].

After excluding technical alarms, "crisis" alarms as well as coded patients with fewer monitor alarms than the minimum alarm count threshold, the vital sign parameter alarms from coded and control patients in the training dataset are encoded by discretizing the extracted numeric values using the class-attribute contingency coefficient (CACC) algorithm [75]. With the discretization scheme, numeric values of vital sign parameter alarms are discretized and mapped into the corresponding intervals. For example, three parameter alarms "systolic arterial blood pressure (SysBP) HI 160 mmHg, SysBP HI 145 mmHg, and SysBP HI 130 mmHg" will be treated as two different alarms if their numeric values fall into two intervals after discretization: $150 <$ SysBP HI < 180 and $120 <$ SysBP HI < 150.

We then define a delta laboratory trigger as the difference of a given laboratory test between the first taken result post admission and the result taken at current time. For instance, if the first arterial carbon dioxide (PaCO2) result from a patient after admission to the ICU is normal and becomes high at current time t, then the delta laboratory trigger of PaCO2 for the patient at time t is denoted as "PaCO2 NR→HI". The delta laboratory triggers reflect changes in laboratory test results during a patient's ICU course. This definition is different from that in our previous work where a delta laboratory trigger was defined as the difference of a given laboratory test between the last two results taken in a T_w-long time window. This is because clinicians often examine changes in diagnostic tests at current time by comparing those at the time of admission to assess the trajectory of a patient's status. It is noted that here we only consider the polarity of a laboratory test result that is provided by the SuperAlarm study database II and ignore the numeric values.

The arrhythmia alarms and encoded vital sign parameter alarms within the T_w-long window, as well as the delta laboratory triggers within an edging threshold of 24-hour window preceding code blue events are extracted from the training dataset to mine SuperAlarm candidates using the maximal frequent itemset algorithm (MAFIA) [82]. The MAFIA algorithm is governed by the parameter of minimum support value (min_sup) that is defined as a minimum percentage of coded patients in the training dataset containing the given SuperAlarm candidates. The false positive rate of a SuperAlarm candidate k, denoted as $FPR_{SA\_C_k}$, will then be calculated in terms of the percentage of T_w-long windows that are consecutively selected from all control patients in the training dataset trigger the SuperAlarm candidate k. Given a specified false positive rate threshold FPR_{thre}, a SuperAlarm candidate will be removed if its $FPR_{SA\_C_k} > FPR_{thre}$. The reminder of SuperAlarm candidates create the final set of SuperAlarm patterns.

5.2.4 SuperAlarm sequence and sampling subsequence

The final set of SuperAlarm patterns is applied on both the coded and control patients to generate SuperAlarm sequences. For each patient at time t when an alarm or a delta laboratory trigger occurs during the monitoring in ICU, any arrhythmia alarms, encoded vital sign parameter alarms and delta laboratory triggers within $[t - T_w, t]$ window are extracted to determine whether any of the final SuperAlarm patterns are triggered at time t. As time evolves, the consecutive SuperAlarm triggers generate a SuperAlarm sequence. Let $\mathbf{\Sigma} = \{SA_1, SA_2, \ldots, SA_K\}$ be a set of the K distinct SuperAlarm patterns. A SuperAlarm sequence $\mathbf{S}$ is denoted as $\mathbf{S} = \langle SA_{t_1}, SA_{t_2}, \ldots, SA_{t_n}\rangle$, where $SA_{t_i} \in \mathbf{\Sigma}$ is a SuperAlarm trigger occurring at time t_i. A T_s-long SuperAlarm subsequence is a segment of SuperAlarm sequence denoted as $\mathbf{s}_{T_s}^{t_a} = \langle SA_{t_a - T_s}, SA_{t_a - T_s + 1}, \ldots, SA_{t_a}\rangle$, where $t_1 \leq t_a - T_s, t_a \leq t_n$. We call SA_{t_a} the anchor SuperAlarm trigger at t_a for the subsequence $\mathbf{s}_{T_s}^{t_a}$.

We follow the approach proposed in Chapter 3 to sample subsequences. In particular, it is intuitive that the subsequences that are closer to code blue events will be more informative and predictive. Therefore, the anchor SuperAlarm trigger SA_{t_a} will have higher probability to be selected if it is closer to the events. This probability can be modeled by an exponential probability density function to select anchor SuperAlarm triggers in order to sample subsequences. The subsequences sampled from coded patients are treated as *positive samples*. On the other hand, we select anchor SuperAlarm triggers based on a uniform distribution from control patients to create *negative samples*. This is because we treat the same importance of each anchor SuperAlarm triggers in a SuperAlarm sequence from a control patient.

5.2.5 Time weighted supervised sequence representation (TWSSR)

The technique of sequence representation is to build a function f by which the subsequence $\mathbf{s}_{T_s}^{t_a}$ can be mapped into a K-dimensional numeric vector $\mathbf{x}^{\mathbf{s}} \in \mathbb{R}^K$, that is, $f : \mathbf{s}_{T_s}^{t_a} \rightarrow \mathbf{x}^{\mathbf{s}}$, where $\mathbf{x}^{\mathbf{s}} = [x_1^{\mathbf{s}}, x_2^{\mathbf{s}}, \ldots, x_K^{\mathbf{s}}]^T$, K is the size of the set of SuperAlarm patterns $\mathbf{\Sigma}$. The element $x_k^{\mathbf{s}}$ in the vector $\mathbf{x}^{\mathbf{s}}$ represents the importance of the associated SuperAlarm trigger in the subsequence $\mathbf{s}_{T_s}^{t_a}$. To create such vector, we propose the time weighted supervised sequence representation (TWSSR) that consists of three factors: local weighting factor, global weighting factor and normalization factor.

5.2.5.1 Local weighting factor

A typical function to measure the importance of each SuperAlarm trigger in the subsequence $\mathbf{s}_{T_s}^{t_a}$ (i.e., local weight) is to calculate its frequency such as tf in term frequency inverse document frequency (TFIDF) method [94]. However, this function does not take into account the time effect on the importance of the SuperAlarm trigger in the subsequence $\mathbf{s}_{T_s}^{t_a}$. An intuitive heuristic is that a SuperAlarm trigger in the subsequence $\mathbf{s}_{T_s}^{t_a}$ should carry more weight to measure its importance when it approaches the anchor SuperAlarm trigger SA_{t_a}.

Let $w^{\mathbf{s}}(t)$ be the weight of time t in the subsequence $\mathbf{s}_{T_s}^{t_a}$, which is given by

$$w^{\mathbf{s}}(t) = (1 - \eta)\eta^{t_a - t} \tag{5.1}$$

where $t \in [t_a - T_s, t_a], \eta \in (0, 1)$. The Equation 5.1 is also called exponential trace memory [124].

Let

$$h_k^{\mathbf{s}}(t) = \begin{cases} 1, & SA_k \text{ triggered at } t \text{ in subsequence } \mathbf{s}_{T_s}^{t_a}, SA_k \in \mathbf{\Sigma} \\ 0, & \text{otherwise} \end{cases} \tag{5.2}$$

then, the weight of the SuperAlarm trigger SA_k occurring at time t in $\mathbf{s}_{T_s}^{t_a}$ is given by

$$w_k^{\mathbf{s}} = \sum_{t \in [t_a - T_s, t_a]} w^{\mathbf{s}}(t) h_k^{\mathbf{s}}(t) \tag{5.3}$$

As a result, the local weight of the SuperAlarm trigger SA_k in $\mathbf{s}_{T_s}^{t_a}$ is defined as

$$f_{\text{local}}^{\mathbf{s}}(k) = \begin{cases} 0, & \sum_{t \in [t_a - T_s, t_a]} h_k^{\mathbf{s}}(t) = 0 \\ \log_2(w_k^{\mathbf{s}} + 1), & \text{otherwise} \end{cases} \tag{5.4}$$

The logarithm transformation in Equation 5.4 is used for reducing the effect of the SuperAlarm trigger SA_k occurring many times in the subsequence $\mathbf{s}_{T_s}^{t_a}$.

5.2.5.2 Global weighting factor

The global weighting factor is to measure the weight of each SuperAlarm trigger occurring in all subsequences in the entire training dataset. As a supervised weighting method, the global weighting factor of TWSSR incorporates the class information of coded patients and controls patients. Many supervised weighting schemes have been proposed, especially in text classification, among which the relevance frequency (rf) weighting scheme has demonstrated the competitive improvement of performance in text classification in comparison with other supervised weighting methods [125]. We follow the rf weighting scheme in this study to measure the global weight of a SuperAlarm trigger.

Let tp_k be the number of subsequences in the coded patients (positive class) in the training dataset having the SuperAlarm trigger SA_k occurred, fp_k the number of subsequences in the control patient group (negative class) containing the SuperAlarm trigger SA_k, the the global weighting factor is then defined as

$$f_{\text{global}}(k) = \log_2 \left(\frac{tp_k}{\max(1, fp_k)} + 2 \right) \tag{5.5}$$

According to the definition of the global weighting factor $f_{\text{global}}(k)$, the more subsequences in the positive class that contain the SuperAlarm trigger SA_k than that in the negative class, the more weight the SuperAlarm trigger SA_k will gain.

5.2.5.3 Normalization factor

The purpose of normalization is to eliminate the impact of a subsequence size (the number of SuperAlarm triggers in the subsequence). The cosine normalization (ℓ_2-norm) is a general normalization method that is defined as $\frac{w_{kj}}{\sqrt{\sum_k w_{kj}^2}}$, where w_{kj} is the weight of the SuperAlarm trigger SA_k in the subsequence j, and therefore $\frac{1}{\sqrt{\sum_k w_{kj}^2}}$ is the cosine normalization factor. However, the cosine normalization may loss the capability of preserving relative weight among different SuperAlarm triggers in as subsequence due to its non-linear transformation. Inspired by [126], the normalization factor in this study is given by

$$f_{\text{norm}}^{\mathbf{s}} = \frac{1}{\log_2(n^{\mathbf{s}} + 2)} \tag{5.6}$$

where $n^{\mathbf{s}}$ is the total number of SuperAlarm triggers in subsequence $\mathbf{s}_{T_s}^{t_a}$.

To this end, the weight of the SuperAlarm trigger SA_k in the subsequence $\mathbf{s}_{T_s}^{t_a}$ measured by the proposed TWSSR scheme is defined as

$$x_k^{\mathbf{s}} = \begin{cases} 0, & \sum_{t \in [t_a - T_s, t_a]} h_k^{\mathbf{s}}(t) = 0 \\ f_{\text{local}}^{\mathbf{s}}(k) \cdot f_{\text{global}}(k) \cdot f_{\text{norm}}^{\mathbf{s}}, & \text{otherwise} \end{cases} \tag{5.7}$$

Substitute Equation 5.4, 5.5 and 5.6 into Equation 5.7, we have

$$x_k^{\mathbf{s}} = \begin{cases} 0, & \sum_{t \in [t_a - T_s, t_a]} h_k^{\mathbf{s}}(t) = 0 \\ \log_{2+n^{\mathbf{s}}}(w_k^{\mathbf{s}} + 1) \cdot \log_2 \left(\frac{tp_k}{\max(1, fp_k)} + 2 \right), & \text{otherwise} \end{cases} \tag{5.8}$$

5.2.6 Baseline approach

The term frequency inverse document frequency (TFIDF) representation scheme [94] is employed as a baseline approach to convert the subsequence $\mathbf{s}_{T_s}^{t_a}$ to the K-dimensional numeric vector $\mathbf{x}'^{\mathbf{s}} = [x_1'^{\mathbf{s}}, x_2'^{\mathbf{s}}, \ldots, x_K'^{\mathbf{s}}]^T$. The element $x_k'^{\mathbf{s}}$ is given by

$$x_k'^{\mathbf{s}} = \begin{cases} 0, & \text{tfidf}_k^{\mathbf{s}} = 0 \\ \dfrac{\text{tfidf}_k^{\mathbf{s}}}{\sqrt{\sum_{SA_k \in \mathbf{\Sigma}} (\text{tfidf}_k^{\mathbf{s}})^2}}, & \text{otherwise} \end{cases} \tag{5.9}$$

where $\text{tfidf}_k^{\mathbf{s}} = \log_2(n_k^{\mathbf{s}} + 1) \cdot \log_2 \dfrac{N}{1 + df_k}$, $n_k^{\mathbf{s}} = \sum_{t \in [t_a - T_s, t_a]} h_k^{\mathbf{s}}(t)$ is the number of SuperAlarm trigger SA_k occurring in the subsequence $\mathbf{s}_{T_s}^{t_a}$, N is the total number of subsequences in the entire training dataset, $df_k = \sum_{\mathbf{s}} \delta(n_k^{\mathbf{s}} > 0)$ measures the total number of subsequences in the entire training dataset that contains the SuperAlarm trigger SA_k, $\delta(x) = 1$ if x is true, $\delta(x) = 0$ otherwise.

5.2.7 Support vector machine and SuperAlarm pattern selection

We adopt support vector machine (SVM) as the classifier to predict code blue events since the choice of classification algorithms is not the focus in this study and SVM often exhibits highly competitive performance in contrast to other classification methods

[105], and also shows comparable performance in computational biology problems [127]. In essence, SVM is a machine learning algorithm to identify an optimal hyperplane (decision boundary) that separates the positive data from the negative data with a maximum margin.

The linear kernel (i.e., mapping function $\phi(\mathbf{x}) = \mathbf{x}$) is employed in this study for the following reasons: (1) the linear kernel measures the cosine similarity between subsequences in the original feature space; (2) the linear kernel can yield better performance in comparison with other types of kernel functions in the case where the dimension of the original feature vector is high and the amount of subsequences in the training dataset is large [108]; and (3) the line kernel offers additional capability for SVM to perform feature selection in an embedded manner.

Due to sparseness of the importance vectors with high dimensionality we encounter in this study, feature selection (i.e., SuperAlarm pattern selection) is a beneficial and fundamental step not only to reduce the risk of "overfitting" and improve the prediction performance using machine learning algorithms [96,101], but also to reduce SuperAlarm triggers redundancy and yield a more compact and informative subset of SuperAlarm patterns [128]. The resultant subset of SuperAlarm patterns could be clinically more relevant to code blue events. Despite its ability to handle the highly dimensional data by "kernel trick", SVM also benefits from the feature selection [129].

We adopt SVM-based recursive feature elimination (SVM-RFE) algorithm to select subset of relevant features. SVM-RFE is a well-known embedded method that incorporates the search of optimal feature subset as part of the linear SVM training process [130]. The basic idea is that SVM-RFE, starting with initial set of all features, removes the feature that is least effective on classification iteratively in a backward elimination manner until the desired number of features to select is reached. The selection criterion of SVM-RFE was derived based on optimal brain damage (OBD) [131] that approximates the change of the linear SVM objective function $J = \mathbf{w}^T\mathbf{w}/2$ caused by removing the feature caused by removing the feature k by expanding the J in Taylor series to second order, which is given by

$$\Delta J(k) = \frac{\partial J}{\partial w_k} \Delta w_k + \frac{1}{2} \frac{\partial^2 J}{\partial w_k^2} \left(\Delta w_k\right)^2 \tag{5.10}$$

At the optimum of J, the first order can be neglected, yielding

$$\Delta J(k) = \frac{\left(\Delta w_k\right)^2}{2} \tag{5.11}$$

Discarding the feature k can be viewed as replacing its weight by zero equivalently, which means $\Delta w_k = w_k$. Therefore, w_k^2 is used as the feature selection criterion of SVM-RFE (i.e., SVM-RFE score). Removal of feature with the smallest w_k^2 (i.e., the associated coefficient of weight vector of the linear SVM) is due to its least effect on classification.

The procedure of SVM-RFE algorithm to select relevant and informative subset of SuperAlarm triggers is described as follows (Table 5.1).

Table 5.1: SVM-RFE algorithm.

```
Algorithm: SVM-RFE

Input: the training dataset: ${\mathbf{X}, \mathbf{y}}$;
       the initial set of SuperAlarm patterns: $\mathbf{\Sigma}$;
       the number of SuperAlarm patterns to be selected: $\ell$.
Output: the subset of SuperAlarm patterns with size $\ell$.

  1. Initialize the feature set $S = \mathbf{\Sigma}$;
  2. Train a linear SVM using SuperAlarm patterns in the set $S$;
  3. Calculate the weight vector $\mathbf{w}$ of the line SVM;
  4. Find the SuperAlarm pattern $SA_k$ with the smallest SVM-RFE score: $k = \underset{i}{\operatorname{argmin}}\, w_i^2$;
  5. Update: $S = S - SA_k$;
  6. Repeat steps 2–5 until $|S| = \ell$;
  7. Return the set $S$ with selected SuperAlarm patterns;
```

For computational efficiency, we remove 1% original SuperAlarm patterns set $\mathbf{\Sigma}$ (i.e., $0.01K$) in each iteration.

5.2.8 Evaluation

The performance of the proposed algorithm is assessed in phases of offline training and online testing. This is done by randomly selecting 80% of coded patients and control patient to compose a training dataset and the remainder of patients to constitute an independent test dataset. The training dataset is used in offline phase to determine the optimal algorithm parameters while the independent dataset is employed in online phase to evaluate the prediction performance under the optimal parameters.

5.2.8.1 Offline analysis

(1) Determination of optimal algorithm parameters to identify SuperAlarm patterns.

In the stage of identifying SuperAlarm patterns, two parameters are required to be optimized: the length of time window T_w and the value of minimum support *min_sup*. This is done by applying a 10-fold cross validation (CV) set on the training dataset. Data from the first 9 folds of the 10-fold CV set is used to generate SuperAlarm candidates in use of MAFIA algorithm under a given *min_sup*. The SuperAlarm candidates are then applied to control patients in the first 9 folds of 10-fold CV set to calculate $FPR_{SA\_C_k}$ for each SuperAlarm candidate. The SuperAlarm candidates with $FPR_{SA\_C_k}$ greater than the threshold FPR_{thre} will be filtered out and the rest of qualified SuperAlarm patterns are applied to the remaining one fold of the 10-fold CV set (i.e., validation set). This process then leads to a pair values of true positive rate (TPR) and false positive rate (FPR), where TPR here is defined as the percentage of coded patients in the validation set who trigger at least one SuperAlarm candidate while FPR here is referred to as percentage of T_w-long long windows that are consecutively selected from all control patients in the validation set that trigger any SuperAlarm patterns.

Varying the threshold of FPR_{thre} results in various pairs of values of TPR and FPR, and a receiver operation characteristic (ROC) curve is generated. The above process repeats 10 times and an average ROC curve can be drawn. Given an acceptable user-specified FPR_{max}, the optimal values for the length of time window T_w and the minimum support min_sup can be determined by selecting the corresponding operating point FPR_{thre} on the ROC curve where the maximal TPR is achieved across all the combinations of T_w and min_sup. Under the optimal T_w and min_sup, the MAFIA algorithm is applied again on the entire training dataset to mine the final SuperAlarm candidates by coalescing the 10-fold CV set into a single one. The final SuperAlarm patterns are eventually identified by removing theSuperAlarm candidates whose $FPR_{SA\_C_k}$ are greater than FPR_{thre}.

(2) Determination of the optimal parameters to recognize temporal patterns.

In the stage of recognizing temporal patterns, the optimal values for the following parameters are determined under a given T_s-long subsequence: the decay parameter η in Equation 5.1, the cutoff of feature selection ratio r for retaining subset of SuperAlarm patterns, and the hyperparameter C for the linear SVM. This is achieved by creating another 10-fold CV set based on the positive and negative T_s-long subsequences in the training dataset. The positive and negative T_s-long subsequences in the first 9 folds of the 10-fold CV set are used to train the linear SVM classifier under the each combination of specified values for parameters of η, r and C. The learned SVM classifier is then applied on the remaining one fold of the 10-fold CV set to assess the classification performance evaluated by the area under the curve (AUC). This process repeated 10 times and an average value of AUC is obtained for each combination of parameters. As a result, the values for parameters η, r and C that lead to the highest AUC are determined as the optimal ones. The final linear SVM classifier is then trained under the optimal parameters of η, r and C based on the entire training dataset of the positive and negative T_s-long subsequences obtained by coalescing the 10-fold CV set into a single one.

It should be noted that the parameter η is not required in use of TFIDF representation. In the training phase based on the 10-fold CV set, the IDF factor is calculated based on the 9 folds and it is then passed to the remaining one fold to evaluate the performance. However, the final IDF factor, which is eventually used in online analysis phase is obtained based on the entire training dataset.

5.2.8.2 Online analysis

The SuperAlarm sequences in the independent test dataset are employed to simulate the application of the learned SVM classifier to predict code blue events. At the moment of a SuperAlarm trigger occurring in a SuperAlarm sequence at time t_i, a T_s-long subsequence preceding t_i is first extracted and converted into a K-dimensional numeric vector by the TWSSR method with the optimal η or the TFIDF scheme with final IDF factor. The numeric vector is then input into the learned SVM classifier under the optimal feature selection ratio r and hyperparameter C, producing a binary prediction outcome of positive or negative at t_i by applying a SVM-threshold to the continuous-valued output of the SVM classifier. As time evolves, this procedure will create a sequence of prediction outcomes. We will optimize the value of SVM-threshold through ROC analysis.

Six metrics are employed to assess the prediction performance based on the learned classifier using SuperAlarm sequences in the independent test dataset. We first define the following measures which will be used in the metrics: (1) $TP^L@T$: true positives with respect to T-long lead time, which is the number of coded patients in the independent test dataset predicted correctly by the classifier (i.e., positive outcome) at least once in a 12-hour window preceding a specified T-long lead time. A lead time T is referred to as a time interval immediately preceding code blue events during which interventions and treatments can be performed. Hence, $TP^L@0$ means the true positives in a 12-hour window immediately preceding the code blue events; (2) $FN^L@T$ false negatives with respect to T-long lead time, which is defined as $FN^L@T = N - TP^L@T$, where N is the number of coded patients in the independent test dataset; (3) FP: false positives, which is the number of control patients in the independent test dataset predict incorrectly (i.e., positive outcome) at least once in a 12-hour window. This is computed as follows. The 12-hour window is randomly selected over the whole monitoring time and this process is repeated 100 time for each control patient. As a result, the average number of incorrect prediction outcomes in a 12-hour window for each control patient is obtained, denoted as $\mu_i = \frac{\sum_{j=1}^{100} T_j}{100}$, where $T_j = 1$ if the jth 12-hour window had at least one incorrect prediction outcome, otherwise $T_j = 0$. Therefore, $FP = \sum_{i=1}^{M} \mu_i$, where M is the total number of control patients in the independent test dataset; (4) TN: true negatives, which is defined as $TN = M - FP$. Consequently, the six evaluation metrics are defined as follows:

1. Sen^L@T, sensitivity with respect to -long lead time that is computed in terms of Sen^L@$T = \frac{TP^L@T}{N}$.

2. Spe, specificity which is calculated by Spe $= \frac{TN}{M}$.

3. PPV, positive predictive value which is obtained by PPV $= \frac{TP^L@0}{TP^L@0 + FP}$.

4. NPV, negative predictive value which is computed by NPV $= \frac{TN}{TN + FN^L@0}$.

5. AFRR, alarm frequency reduction rate which is defined as AFRR$= 1-$FPR, where FPR is the false positive ratio which is calculated as a ratio of hourly number of the incorrect prediction outcomes for control patients in the independent dataset to that of monitor alarms and laboratory test results. The AFRR measure how many false SuperAlarm triggers occurring in control patients can be compressed by the SVM classifier.

6. WDR, work-up to detection ration which is calculated by WDR $= \frac{TP^L@0 + FP}{TP^L@0}$. The WDR measures how many false positives can be introduced when one true positive occurs.

5.2.9 Experimental setup

The experiments in this study are conducted under the following conditions. Four T_w–long time windows (in minutes) are specified: $T_w \in \{30, 60, 90, 120\}$. Three minimum support values are given: $min\_sup \in \{0.05, 0.10, 0.20\}$. The FPR_{max} used to identify the final SuperAlarm pattern is set to 30% to ensure a minimum of true SuperAlarm triggers are missed. Four subsequence lengths (in hours) are specified: $T_s \in \{4, 8, 12, \infty\}$, where ∞ implies the subsequence is extracted from the beginning of monitor to the current time when a SuperAlarm trigger occurs. The decay parameter η in Equation 5.1 is given by $\eta \in \{0.50, 0.80, 0.90, 0.95, 0.98\}$. The feature selection ratio r and the hyperparameter C in the SVM classifier are specified as $r \in \{0.01, 0.02, \ldots, 0.10\}$ and $C \in \{2^{-8}, 2^{-7}, \ldots, 2^1\}$, respectively.

5.3 Results

5.3.1 Patient data

This study has been approved by the UCLA Institutional Review Board and the UCSF Committee on Human Research with a waiver of patient consent. Adult patients (age $>$ 18 years) included in this study were admitted to 120 beds in ICUs (neurosurgical, cardiac, medical, medical/surgical) at UCLA Medical Center between January 2010 to June 2014, and 77 beds in ICUs (neurological/neurosurgical, medical/surgical, and cardiac critical care)at UCSF Medical Center between March 2013 to March 2015. Quality management services at UCLA Medical Center and UCSF Medical Center provided a listing of patients with at least one code blue call during the admission, respectively. A cohort of control patients without any code blue call or unplanned ICU transfer were selected for each of coded patients under the following additional criteria: (1) same age (±5 years); (2) same gender; (3) admission to the same hospital unit; and (4) admission in the same month.

As a result, a total of 403 coded patients (57.8% male, 72.2% White or Caucasian, age at admission 62.6 ± 16.5, average monitor duration 23.3 days (median 13.0, IQR 4.3 to 27.2)) and 4667 controls patients (62.6% male, 73.7% White or Caucasian, age at admission 63.1 ± 14.0, average monitor duration 12.8 days (median 6.4, IQR 2.8 to 13.1)) from UCLA Medical Center were evaluated.

For UCSF Medical Center, 152 coded patients (54.6% male, 47.4% White or Caucasian, age at admission 61.6 ± 15.2, average monitor duration 13.2 days (median 7.6, IQR 2.5 to 18.4)) and 1115 control patients (63.7% male, 55.9% White or Caucasian, age at admission 62.8 ± 11.0, average monitor duration 7.0 days (median 3.4, IQR 1.7 to 7.2)) are included in this study.

5.3.2 Monitor alarms and laboratory test results

After exclusion of crisis alarms and technical alarms, 88 distinct monitor alarms 14 arrhythmia alarms and 74 vital sign parameter alarms) are included in this study. For UCLA patients, A total of 1566133 monitor alarms (average 278.4 per patient per day, median 158.6, IQR 93.5 to 281.9) preceding code blue events are extracted from coded patients. There are 62 coded patients with more than one code blue call; only those monitor alarms that precede the first call are extracted. Meanwhile, a total of 7341089 monitor alarms (average 124.1 per patient per day, median 83.0, IQR 37.0 to 152.5) are extracted from control patients.

For UCSF patients, 1815160 monitor alarms (1107.2 per patient per day, median 670.4, IQR 376.7 to 1349.3) preceding code blue events are extracted from coded patients. A total of 26 patients had more than one code blue call; only monitor alarms preceding the first call are extracted. There are 5901531 (average 689.7 per patient per day, median 363.7, IQR 211.2 to 670.0) monitor alarms extracted from control patients.

A total of 58 distinct laboratory tests included in this study. For UCLA patients, 534734 (41.6% NR(normal), 28.3% LO(low), 30.1% HI(high), average count 110.7 per patient per day, median 74.0, IQR 45.1 to 125.1) and 2590318 (45.7% NR, 26.9% LO, 27.4% HI, average count 78.2 per patient per day, median 42.3, IQR 29.6 to 62.8) laboratory test results are extracted from coded patients and control patients, respectively.

For UCSF patients, a total of 216265 (44.8% NR, 30.4% LO, 24.8% HI, average count 127.3 per patient per day, median 85.3, IQR 42.4 to 161.5) and 847730 (46.7% NR, 31.3% LO, 22.0% HI, average count 95.7 per patient per day, median 56.6, IQR 32.6 to 115.2) laboratory test results laboratory test results are extracted from coded patients and control patients, respectively.

The Wilcoxon Rank-Sum test results shows that the number of monitor alarms from the patients between the two institutions are significantly different ($p \ll 0.01$). However, the number of laboratory test results is not significantly different for coded patients between the two institutions ($p = 0.3310$) while the difference of the number of laboratory test results for control patients between the two institutions is significant ($p \ll 0.01$).

By applying NHPP model based on the four specified T_w-long windows, that is, 30 minutes, 60 minutes, 90 minutes and 120 minutes, the minimum alarm count thresholds are 4, 8, 11 and 13 respectively. Accordingly, the number of excluded all coded patients are 126, 138, 139 and 134, respectively.

5.3.3 Results of identification of SuperAlarm patterns

Figure 5.2 shows the ROC curve based on each combination of available T_w-long windows and min_sup values. Based on the specified $FPR_{max} = 30\%$, the maximal TPR of 90.83±8.07% is obtained under $T_w = 60$ minutes and $min\_sup = 0.05$. The final SuperAlarm patterns are obtained under the optimal T_w and min_sup by filtering out the ones whose $FPR_{SA\_C_k}$ are greater than the associated operating point FPR_{thre} on the ROC curve. As a result, 6224 SuperAlarm patterns are obtained. The length of the SuperAlarm patterns ranges from 1 to 10. An example of such SuperAlarm

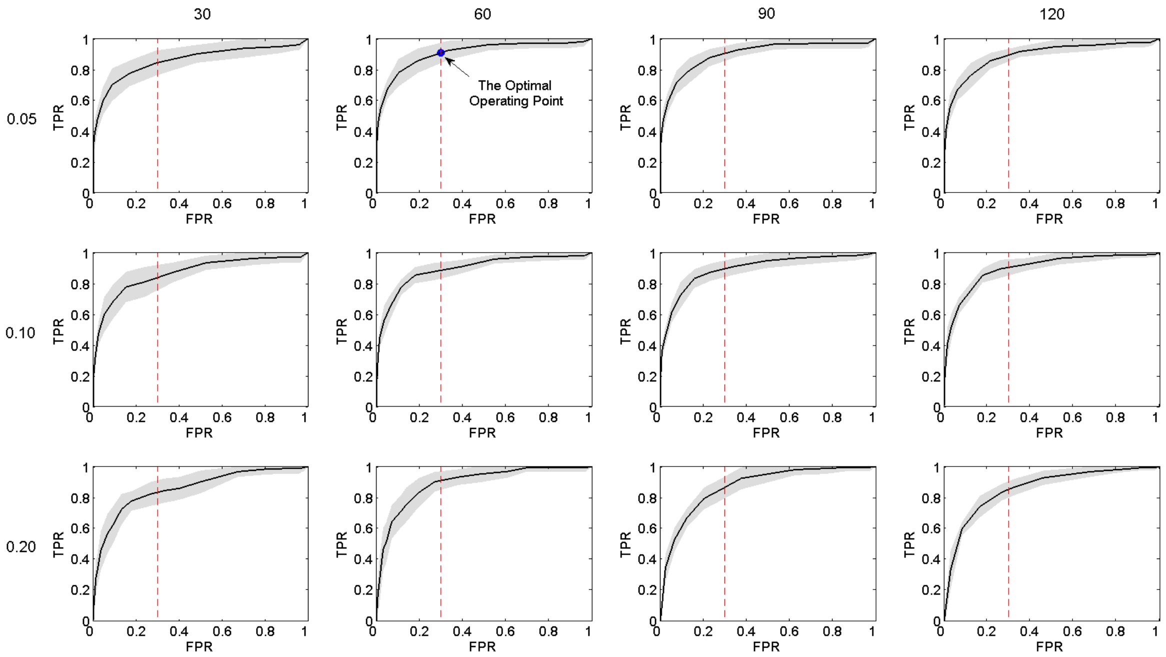


Figure 5.2: Determination of the optimal parameters of T_w-long time window and minimum support min_sup for identification of finial SuperAlarm patterns under $FPR_{max} = 30\%$ based on ROC analysis. The rows and columns represent the specified values of min_sup and T_w(in minutes), respectively.

pattern “POTASSIUM NR→NR, 19.5<BP MEAN LO<52.5, 27.5<BP SYS LO<75.5, 61.5<SPO2 LO<85.5, PVC” can be explained as follows: if the clinical events “potassium remains normal, mean blood pressure being low between 19.5 to 52.5 mmHg, systolic blood pressure being low ranging from 27.5 to 75.5, SpO2 being low between 61.5 to 85.5 and premature ventricular contraction (PVC)” co-occur in a 60 minutes time window, then the patient could be at risk. Please note that delta laboratory test result (e.g., potassium remains normal in this case) in the SuperAlarm patterns can occur in 24-hour window (i.e., edging threshold as aforementioned in Section 5.2.3).

5.3.4 Positive and negative samples in the training dataset

By applying the method mentioned in Section 5.2.4 with a maximal number of sampled subsequences being 60 per each coded patient and 3 per each control patient in order to make a balanced training dataset, we obtain a total of 12017 positive samples and 12161 negative samples after discarding the subsequences without any SuperAlarm triggers.

5.3.5 Offline analysis results

Table 5.2 lists AUC values with respect to feature selection ratio r resulting from the offline training analysis. For each specified subsequence length T_s, the optimal values for parameters for the methods of TWSSR (i.e., decay parameter η and hyperparameter C) and TFIDF (i.e., hyperparameter C) are also given in the table, respectively. It is observed that for a given representation method and T_s, the AUC value first grows to a maximal point when the feature selection ratio r increase, and then the AUC value starts to drop as the feature selection ratio r continues increasing. Consequently, the optimal feature selection ratios are determined as the one that is associated with the highest AUC. We can also see that TWSSR only retains a few SuperAlarm triggers after applying feature selection process because of the optimal ratios ranging from 0.01 to 0.03 based on the all of the four specified subsequence lengths, whereas TFIDF can lead the optimal ratio to 0.08 when $T_s = \infty$.

Figure 5.3 displays an example of all 4-hour long subsequences in the training dataset that are converted into numeric vectors by TWSSR method under the optimal decay parameter $\eta = 0.95$. The columns represent the selected SuperAlarm patterns (features) with optimal $r = 0.02$, while the rows represent positive samples (upper) and negative samples (bottom). Intensity of each pixel in the image represents the importance of the selected SuperAlarm pattern reflected by the numeric value in the vector, which are scaled into [0, 1] in this plot. Visually, the selected SuperAlarm patterns are indicative of positive and negative classes.

Table 5.2: AUC values for methods of TWSSR and TFIDF based on each specified feature selection ratio r and subsequence length T_s. The bold numbers indicate the highest AUC for given method based on specified r and T_s under the optimal parameters of η and C.

T_s(hrs)	Method	AUC (%, mean±std)									
		r=0.01	r=0.02	r=0.03	r=0.04	r=0.05	r=0.06	r=0.07	r=0.08	r=0.09	r=0.1
4	TWSSR ($\eta = 0.95, C = 2^{-7}$)	69.73±3.12	**70.24 ± 4.46**	69.57±3.17	69.09±2.97	68.75±3.30	68.39±3.63	68.43±3.98	68.45±3.98	68.20±3.53	68.84±3.15
	TFIDF ($C = 2^{-6}$)	**70.02 ± 4.49**	69.7±4.38	69.20±4.70	68.82±3.75	68.93±3.62	68.91±3.43	68.56±3.45	68.16±3.44	67.67±3.6	67.65±3.73
8	TWSSR ($\eta = 0.98, C = 2^{-6}$)	70.05±3.28	69.89±3.74	**70.97 ± 2.96**	70.48±3.49	70.07±3.84	69.84±4.00	69.75±4.06	68.50±4.14	68.54±3.73	68.60±3.24
	TFIDF ($C = 2^{-6}$)	69.54±3.11	70.43±4.31	**71.03 ± 3.84**	70.04±3.81	69.94±3.98	69.42±4.31	69.07±4.08	68.98±3.95	69.21±3.99	69.07±4.25
12	TWSSR ($\eta = 0.98, C = 2^{-7}$)	69.77±3.43	**70.76 ± 3.59**	70.10±3.34	69.92±3.41	69.64±3.32	69.27±3.97	68.94±3.97	68.98±3.93	68.78±3.55	68.51±3.59
	TFIDF ($C = 2^{-5}$)	70.45±3.31	**71.00 ± 4.82**	69.94±4.24	69.27±4.71	69.15±4.69	68.93±4.19	68.65±4.25	68.45±4.44	68.73±4.16	68.53±4.39
∞	TWSSR ($\eta = 0.98, C = 2^{-7}$)	**72.78 ± 3.02**	72.04±4.42	70.73±3.9	70.11±3.36	70.16±4.24	70.11±3.51	69.25±4.13	69.63±4.02	69.31±4.47	69.82±5.17
	TFIDF ($C = 2^{-7}$)	71.62±3.96	70.41±3.15	71.35±3.65	71.95±3.18	72.04±4.03	72.13±4.19	72.23±3.99	**72.39 ± 4.05**	72.37±4.49	72.31±4.29

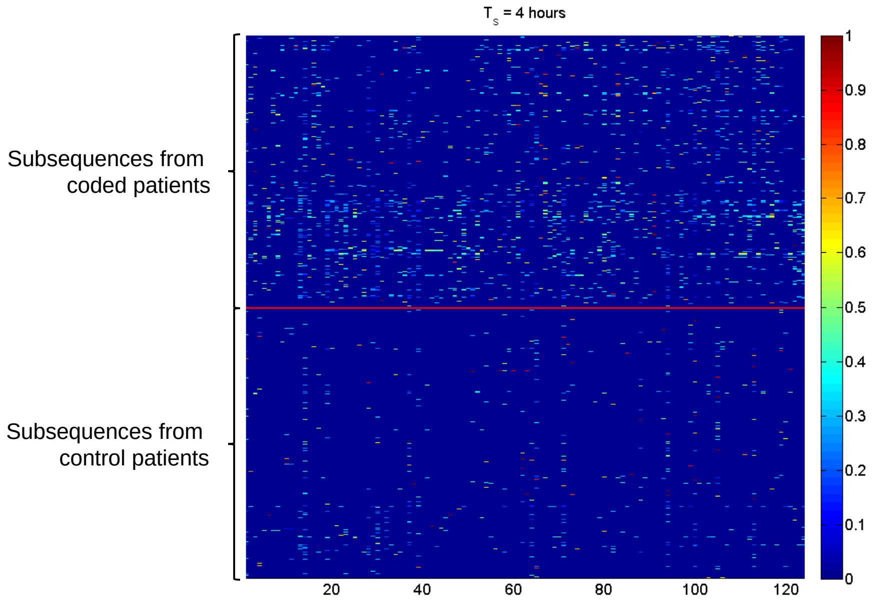


Figure 5.3: An example of illustration of all 4-hour long SuperAlarm sequences in the entire training dataset. The sequences are converted into numeric vectors by TWSSR method under the optimal decay parameter $\eta = 0.95$ with the retained SuperAlarm triggers after performing feature selection (the optimal feature selection ratio $r = 0.02$ in this example).

5.3.6 Online analysis results

We vary thresholds to dichotomize the SVM outputs and generate ROC curves in order to find the optimal ones for the SVM classifiers based on methods of TWSSR and TFIDF under the optimal algorithm parameters for each specified T_s. ROC curves are drawn using the positive and negative samples in the entire training dataset (Figure 5.4). The optimal threshold for a given method and T_s is determined by finding the operating point on the curve that is closest to the reference point on the upper left corner of the unit square (point R in Figure 5.4).

By applying the optimal SVM thresholds, Table 5.3 presents results of Sen^L@T when lead time T are set for half an hour, 1 hour, 2 hours, 6 hours and 12 hours, as well as the

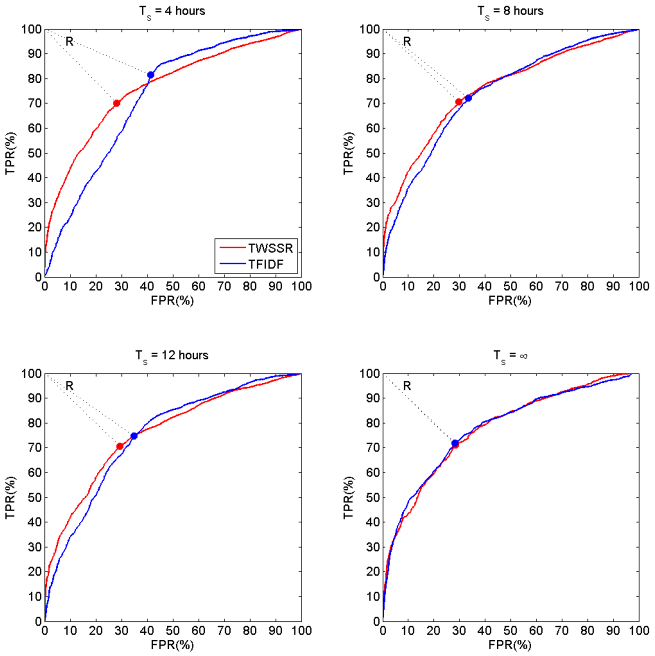


Figure 5.4: Determination of the optimal thresholds for SVM classifiers based on the entire training dataset. For a given T_s the optimal operating point on each ROC curve is selected the one closest to the reference point R.

results of Spe, PPV, NPV, AFRR and WDR. In general, we can see that for all available T_s, the values of Sen^L@T obtained using TWSSR method is higher than that obtained using TFIDF method with only two exceptions (i.e., Sen^L@1 and Sen^L@6 under $T_s = \infty$). For T_s between 4 hours to ∞, the values of Sen^L@1 (sensitivity with respect to 1-hour lead time), for instance, range from [71.60 - 83.95%] and [64.20 - 80.25%] by using TWSSR and TFIDF, respectively. Meanwhile, the values of AFRR, WDR, Spe, PPV and NPV resulting from TWSSR are [86.62 - 89.82%], [4.77 - 5.63], [61.16 - 67.07%], [17.75 - 20.97%] and [97.67 - 99.15%], respectively, while the values of these metrics obtained based on TFIDF method are [86.78 - 94.09%], [3.73 - 5.90], [61.05 - 81.10%], [16.94 - 26.83%] and [97.27 - 98.47%], respectively. In particular, T_s = 8 hours results in the highest Sen^L@T for both TWSSR and TFIDF, compared to other specified T_s values. For example, Sen^L@1 can yield 83.95% and 80.25% using TWSSR and TFIDF when T_s = 8, respectively. The corresponding AFRR, WDR, SPE, PPV and NPV are 88.21%, 5.55, 61.16%, 18.01% and 98.90% for TWSSR method, and 89.12%, 5.17, 65.90%, 19.35% and 98.47%, respectively. In addition, Table 5.3 also presents the Sen^L@T resulting from the individual SuperAlarm triggers as well as AFRR, WDR, SPE, PPV and NPV, which are represented as the boundary of performance by using SuperAlarm sequence classifier approach in this study.

Table 5.3: Results of the six performance metrics evaluated based on the independent test dataset under the optimal thresholds of the classifier. The optimal thresholds are obtained using ROC analysis shown in Figure 5.4.

T_s	Method	Sen^L@T (%)					AFRR(%,	WDR	SPE(%,	PPV(%,	NPV(%,
(hrs)		0.5 hr	1 hr	2 hrs	6 hrs	12 hrs	mean±std)	(mean±std)	mean±std)	mean±std)	mean±std)
4	TWSSR	85.19	81.48	75.31	69.14	51.85	89.82±0.61	4.77±0.13	67.07±1.16	20.97±0.58	99.15±0.01
	TFIDF	67.90	64.20	61.73	55.56	41.98	94.09±0.49	3.73±0.13	81.10±0.92	26.83±0.96	97.27±0.03
8	TWSSR	86.42	83.95	75.31	70.37	56.79	88.21±0.64	5.55±0.14	61.16±1.18	18.01±0.45	98.90±0.01
	TFIDF	83.95	80.25	72.84	70.37	56.79	89.12±0.62	5.17±0.14	65.90±1.15	19.35±0.53	98.47±0.02
12	TWSSR	79.01	76.54	71.60	65.43	55.56	88.81±0.62	5.22±0.14	65.75±1.17	19.16±0.53	98.10±0.02
	TFIDF	76.54	72.84	67.90	60.49	48.15	88.89±0.61	4.96±0.14	69.52±1.11	20.17±0.59	97.74±0.03
∞	TWSSR	79.01	71.60	69.14	60.49	58.02	86.62±0.62	5.63±0.13	63.36±1.00	17.75±0.40	97.67±0.02
	TFIDF	77.78	74.07	67.90	64.20	56.79	86.78±0.56	5.90±0.09	61.05±0.74	16.94±0.27	97.57±0.03
Individual SuperAlarm trigger		97.53	96.30	90.12	83.95	79.01	61.83±0.41	7.53±0.15	49.32±0.21	13.28±0.29	100.00±0.00

Table 5.3 only presents one choice of the threshold (i.e., the optimal threshold) and hence lack the whole picture of how the metrics vary with thresholds. To address this issue, we produce the plots of sensitivity vs. AFRR, called SvA curve, and sensitivity vs. WDR, called SvW curve based on different thresholds. As an example, curves of Sen^L@1 vs. AFRR and Sen^L@1 vs. WDR are illustrated in Figure 5.5(a) and Figure 5.5(b), respectively. For TWSSR method, we plot curves of SvA and SvW under $T_s = 8$ hours due to its ability to offer the highest sensitivity according to Table 5.3. On the other hand, we plot SvA and SvW curves for TFIDF method under all available T_s in this study so that for a given AFRR or WDR, a range from the lowest Sen^L@1 to the highest can be obtained. The cycle points on the curves resulting from TWSSR method represent the corresponding values obtained based on the optimal threshold while the

triangle points represent that based on default threshold (i.e., zero). We can see that the TWSSR method can achieve higher Sen^L@1 under $T_s = 8$ hours than TFIDF method under all available T_s under a desirable range of high AFRR or low WDR. It can be also seen that the optimal threshold we determined using training dataset does not match the one based on the independent test dataset, which may not be obtainable. Under all of the SVM-thresholds, we further conduct the paired t-test on metrics of Sen^L@1, AFRR, WDR that are obtained based on methods of TWSSR with $T_s = 8$ hours and TFIDF with all available T_s. The results show significant differences on these metrics between the TWSSR method and the TFIDF method for each T_s ($p \ll 0.01$).

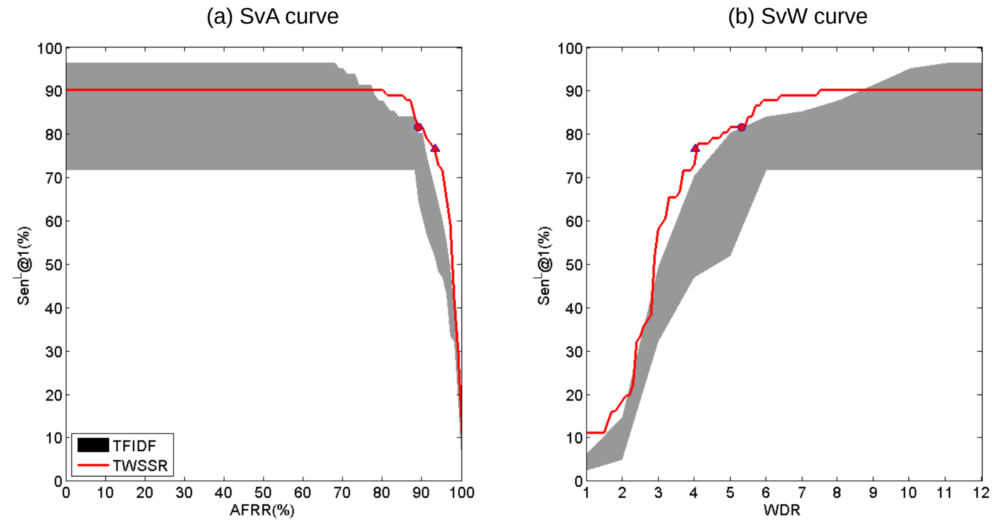


Figure 5.5: (a) SvA curve: Sen^L@1 vs. AFRR, (b) SvW curve: Sen^L@1 vs. WDR. The ranges displayed for the TFIDF method are obtained under all specified T_s in this study (from 4 hours to ∞), while the red lines for the TWSSR method represent the SvA and SvW curves under $T_s = 8$ hours.

According to Table 5.3, there are 13 (out of 81) coded patients and 421 (out of 1134) control patients in the independent test dataset predicted incorrectly. As an example, we display data profiles (monitor alarms and laboratory test results) and continuous-valued classification outcomes for four cases: a TP (true positive), a FN (false negative), a FP (false positive) and a TN (true negative) from the independent test dataset in Figure 5.6 to Figure 5.9, respectively. For clinical use, high sensitivity and high specificity for prediction of patient deterioration are critical for acuity monitoring and prompt interventions. Therefore, we further analyze the cases of FN (Figure 5.7) and FP (Figure 5.8) by performing chart review, respectively. For the FN case, the patient's terminal arrhythmias started as bradycardia, then ventricular fibrillation, and finally asystole that only occurred in the last 10 minutes preceding the event. In addition, blood pressure was plummeting prior to the arrhythmias (as shown in Figure 5.7). However, only a few of types of clinical alarms in the 12-hour time window preceding the 1-hour lead time frequently occurred, which result in negative prediction. For the FP case (Figure 5.8), this patient was admitted with an ST elevation myocardial infarction, and transferred into the ICU following stent placement to the right coronary artery, and for hemodialysis. Despite no code blue call, the condition of this patient was concerning, i.e., it is not a total surprise to have SuperAlarm sequences from this patient be classified as positive.

5.4 Discussion

This study has reported a systematically effort towards predicting code blue events and reducing alarm burden by recognizing temporal patterns in SuperAlarm sequences. Invented by Hu *et al.* in [74], the "SuperAlarm pattern" was originally referred to as a super set of monitor alarms and further extended by integrating laboratory test results with monitor alarms [92]. We then generated SuperAlarm sequences by detecting consecutively emerging SuperAlarm patterns (termed "SuperAlarm triggers") when monitoring data from patients. To be able to thoroughly evaluate this developed framework, we first used data from the SuperAlarm study database II as we reported in Chapter 4 that includes mapped monitor alarms and laboratory test results extracted from a total

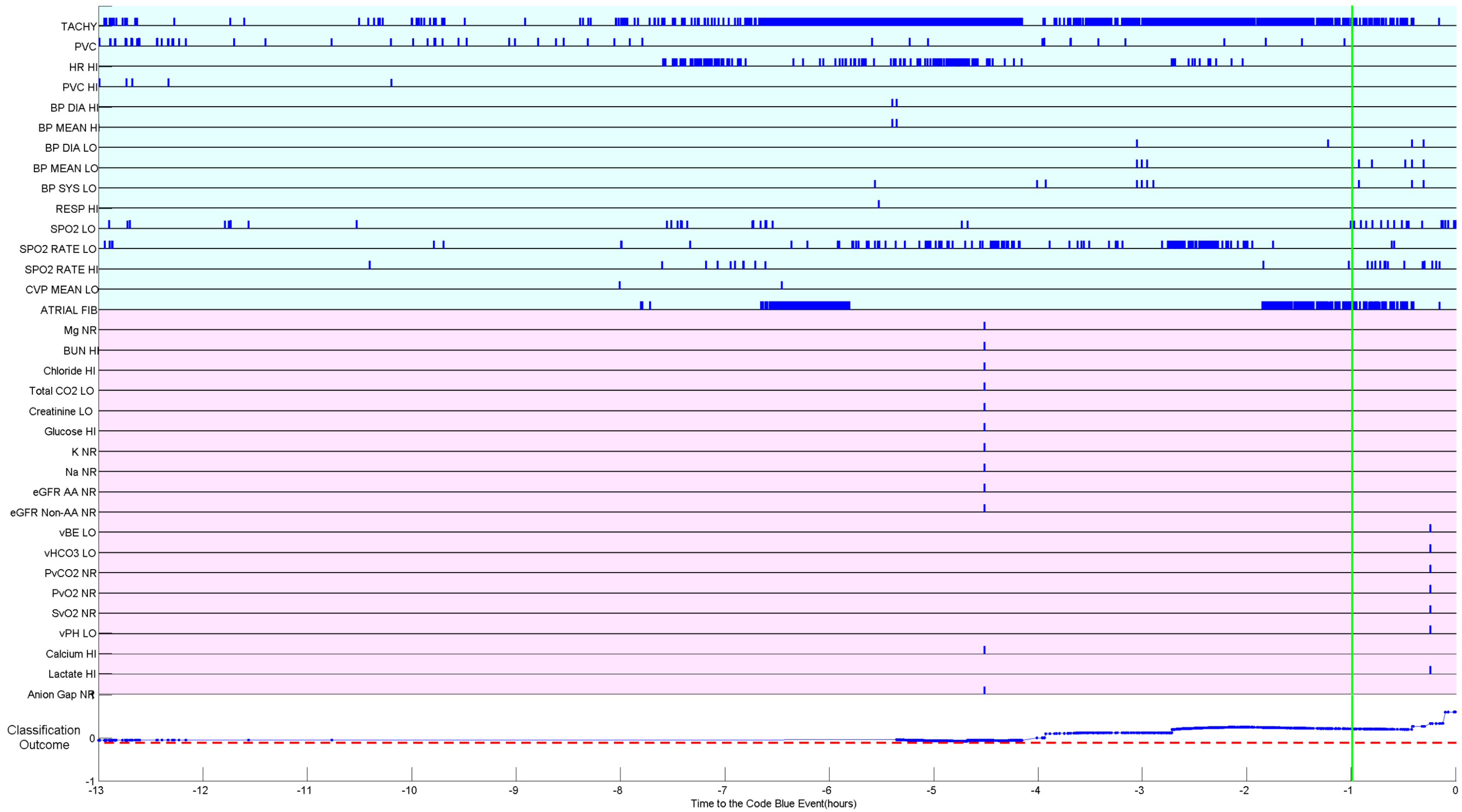


Figure 5.6: An example of a true positive (TP) case's data profile of monitor alarms (top panel), laboratory test results (middle panel), and continuous-valued classification outcomes that are obtained using the SVM classifier based on the proposed TWSSR method (bottom panel). The red dash line represents the optimal threshold of SVM classifier. The green vertical line represents the 1-hour lead time. Zero point on x-axis represents the time point when code blue event occur.

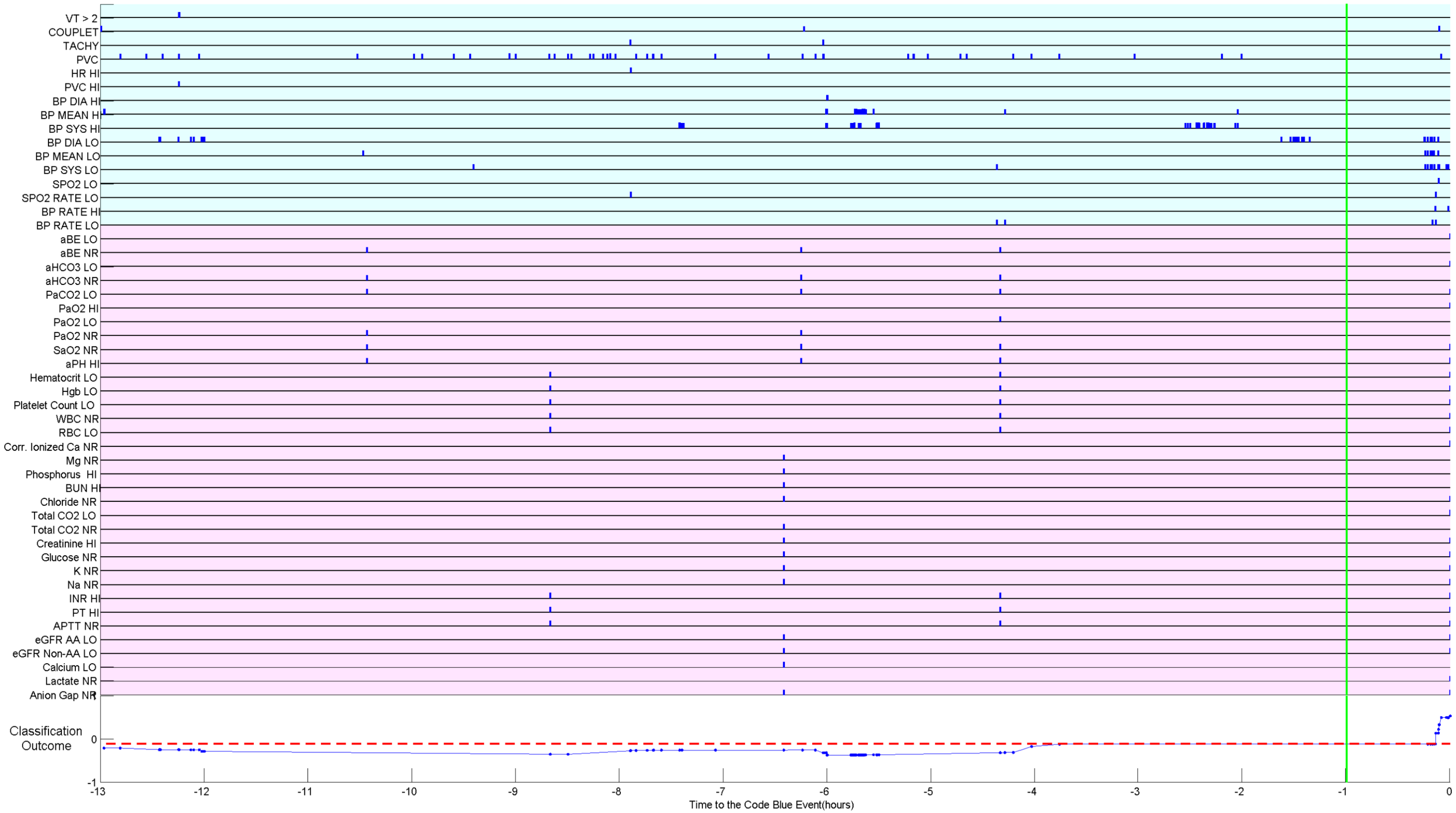


Figure 5.7: An example of a false negative (FN) case's data profile that is displayed in the same way as Figure 5.6.

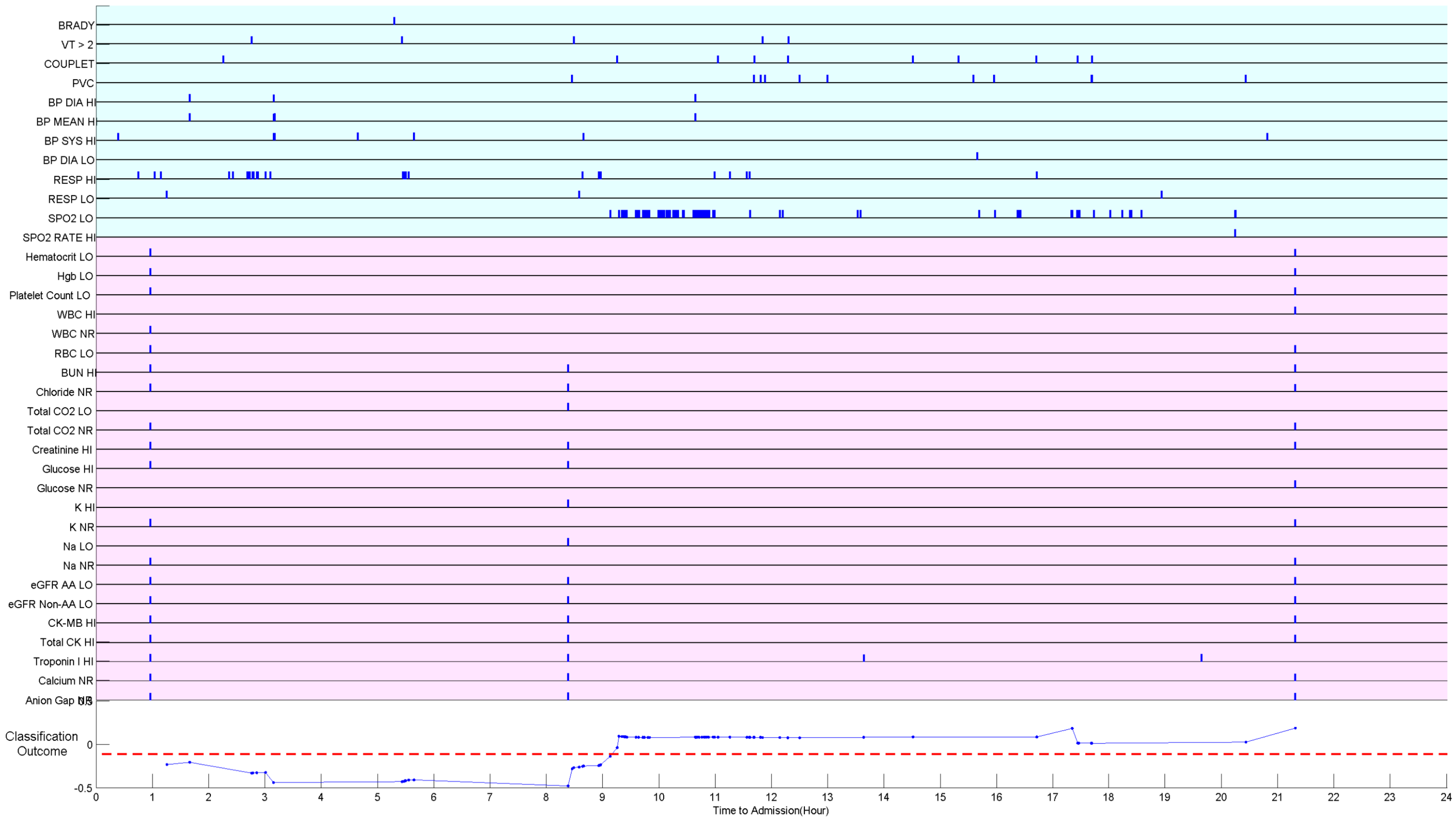


Figure 5.8: An example of a false positive (FP) case's data profile that is displayed in the same way as Figure 5.6. Zero point on x-axis represents time point of the patient's admission.

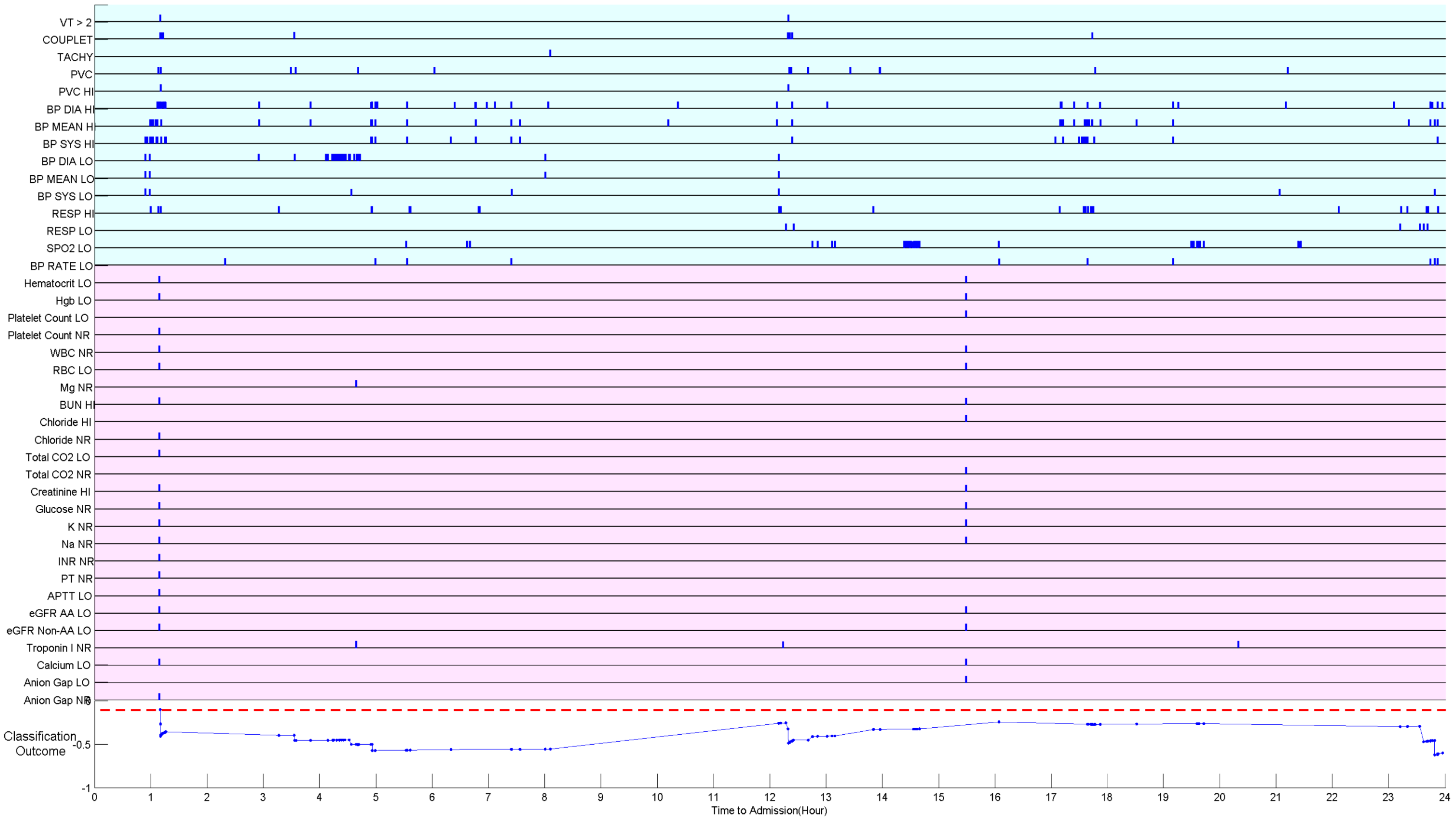


Figure 5.9: An example of a true negative (TN) case's data profile that is displayed in the same way as Figure 5.6. Zero point on x-axis represents time point of the patient's admission.

of 555 coded patients and 5782 control patients admitted to the ICUs in the UCLA and UCSF Medical Centers for identification of SuperAlarm patterns. We then focused on studying a new approach of representing SuperAlarm sequences as fixed-dimensional vectors, i.e., the TWSSR method. We adopted the SVM-RFE algorithm to carry out classification in conjunction with feature selection. As an embedded method, the RFE takes feature dependencies into consideration and iteratively eliminates features with least effect on performance. The results in Table 5 show that for the 8-hour long SuperAlarm sequences, the TWSSR method can yield the highest Sen^L@1 of 83.95% under the optimal algorithm parameters, with negative predictive value (NPV) of 98.90%, specificity of 61.16%, positive predictive value (PPV) of 18.01%, AFRR of 88.21%, and WDR of 5.55.

Figure 5.5 clearly illustrates the superior performance of the newly proposed TWSSR sequence representation approach because a higher sensitivity with 1-hour lead time is obtained under a desirable range of high AFRR or low WDR, compared to the TFIDF representation method. While TFIDF only considers the relative frequencies of different SuperAlarm triggers, the TWSSR approach incorporates two additional important factors to weight the relative frequencies. The first factor is related to how close a SuperAlarm trigger is to the current time such that a higher weight is associated with triggers in a sequence that are closer to the current time. The second factor is related to the importance of a SuperAlarm pattern that a trigger represents. The importance of a SuperAlarm pattern is calculated based on the training data so that those patterns with more frequent triggers among coded patients and less frequent triggers among control patients will have a higher weight. These two heuristics make sense and have been demonstrated to achieve better performance. Further work is needed to study their influences on the performance separately [125].

Being the first study using the SuperAlarm study database II, it is difficult to compare the performance obtained in this study to what was reported in our previous studies. Our recent study [93] developed an algorithm, called weighted accumulated occurrence representation (WAOR) to represent SuperAlarm sequences, which used the temporal closeness of triggers to the current time to weight the frequency of SuperAlarm triggers in a sequence. This was a similar idea to that used in the newly proposed TWSSR. Furthermore, a L1-logistic regression, a classifier model that is different to what used in the present work, was used to perform classification and achieved Sne^L@2 of 90.0%, AFRR of 88.5%, WDR of 4.75. Another study [132] compared the performance of using the sequence representation method (TFIDF) on SuperAlarm sequence and monitor alarm sequences. It was found that SuperAlarm sequence was clearly superior to monitor alarm sequences as an input to predict code blue events. In that study, we used the SVM classifier but adopted information gain (IG) for feature selection, and yielded Sen^L@2 of 93.33% AFRR of 87.28% and WDR of 3.01. The performance reported in the present study is inferior to that obtained in those prior studies. However, a much smaller database with data from one institution was used in those two studies. In terms of the methodology, the choice of the methods adopted in this study was informed by these previous efforts. For example, TWSSR was developed to improve the WAOR sequence representation because this approach additionally considers the distribution of individual SuperAlarm patterns between coded patients and control patients, i.e., the predictive performance of SuperAlarm patterns. Compared to the IG feature selection approach used in [132], the SVM-RFE is more appealing because it considers the interactions among different features as shown in other studies [133–135].

Although different classifiers were used among these three studies, it is unlikely that they would have caused the performance difference seen here. We argue that the most likely reason may be that the previous studies were developed and evaluated based on a dataset including only 254 coded patients and 2213 control patients from a single institution. The larger database in this study contains data from two different medical centers and hence may introduce confounding factors that the proposed SuperAlarm model has failed to capture. One possible confounding factor is the difference between the alarm data collection methods that were used in these two medical centers. Since message-level alarms are captured at UCSF Medical Center, the number of alarms for UCSF patients is much larger. Therefore, it is more likely to have more SuperAlarm triggers for these patients. If the false detections among these patients outweigh true detections, the overall performance may suffer. To verify this hypothesis, we further investigated all the 12-hour windows that were randomly selected for calculating WDR, and we found that 63.61% of the 12-hour windows from UCSF control patients were predicted incorrectly (i.e., false positive rate), compared with 36.90% of the 12-hour windows from UCLA control patients. The higher false positive rate resulting from UCSF control patients than that from UCLA control patients implies that SuperAlarm triggers occur more frequently among UCSF control patients due to a larger number of alarms captured at UCSF Medical Center.

The limitations of the present study include retrospective design, lack of clinical verification and assessment of each identified SuperAlarm pattern and use of laboratory test results without numeric values. Future studies can also be directed towards incorporation of metrics derived from ECG signals such as PR interval, QRS duration, QT interval and so on [119] so as to enrich SuperAlarm patterns. Future studies also need to be conducted in a real-time and prospective manner to evaluate feasibility of streaming data analytics, true predictive power of the proposed SuperAlarm sequence classifier approach.

5.5 Conclusion

We have reported prediction of code blue events and reduction of monitor alarm burden by recognizing temporal patterns in the SuperAlarm sequences. We first extracted the mapped monitor alarms and laboratory test results from the SuperAlarm study database II for the coded patients and control patients admitted to the ICUs at the UCLA and UCSF Medical Centers. We then utilize the previously developed framework to identify SuperAlarm patterns and further construct SuperAlarm sequences.

We have proposed a novel sequence representation method, called time weighted supervised sequence representation (TWSSR) to convert the SuperAlarm sequences into fixed-dimensional vectors in order for recognition of temporal patterns in SuperAlarm sequences. We adopt SVM-RFE algorithm to perform classification in conjunction with feature selection. Compared with the TFIDF representation scheme, the results demonstrate that TWSSR method improves the performance in prediction of code blue events as it leads to higher sensitivity under a desirable range of high alarm frequency reduction rate or low work-up to detection ratio. Therefore, the approach proposed in this study can potentially assist clinicians in the prediction of patient deterioration and reduce alarm burden so as to improve patient monitoring as a complementary tool.

CHAPTER 6

Conclusion and future work

Current physiologic monitors remain imprecise in terms of anticipate patient deterioration as evidenced by the widespread alarm fatigue problem in the intensive care units (ICUs). In this dissertation we have developed a data fusion system to predict code blue events and reduce alarm burden by identifying SuperAlarm patterns and then recognizing temporal patterns in the SuperAlarm sequences. The system accommodates monitor alarms available from physiologic monitors and laboratory test results available in the electronic heath record system, and creates a super set of these physiological and clinical events by exploring relationships among them, and hence a term “SuperAlarm”. The SuperAlarm system is capable of recognizing patient deterioration without causing alarm fatigue to potentially enable early therapeutic interventions and treatments. Therefore, the SuperAlarm system proposed in this dissertation may have potentials of providing a new paradigm for patient monitoring and enhancing the quality of care.

In this final Chapter, we first review the primary contributions of the present dissertation in Section 6.1. In Section 6.2, we then discuss about future research directions.

6.1 Contribution of the dissertation

6.1.1 Development of a data fusion framework for identification of SuperAlarm patterns

In Chapter 2, a data fusion framework for identification of SuperAlarm patterns to predict code blue events and reduce alarm frequency has been developed and evaluated.

The SuperAlarm patterns are defined as multivariate combinations of monitor alarms and laboratory test results that co-occur sufficiently often in a time window preceding code blue events but rarely among control patients. In particular, we proposed two approaches to integrate patient data streams of monitor alarms and abnormality flags of laboratory test results. The abnormality flags were obtained by comparing the numeric values of laboratory test results against their corresponding reference ranges. We then exploited a maximal frequent itemset algorithm to mine the combinations, which were further filtered out if they also occurred frequently among control patients. The results showed that the use of SuperAlarm patterns can achieve high sensitivity in prediction of code blue events and significantly reduce alarm frequency.

6.1.2 Recognition of temporal patterns in SuperAlarm sequences

In Chapter 3, we deployed the SuperAlarm patterns to monitor patients and detected the emerging SuperAlarm patterns which were termed SuperAlarm triggers. The consecutive SuperAlarm triggers over the monitoring time forms SuperAlarm sequences. We developed a sequence classifier to recognize temporal patterns in SuperAlarm sequences. We proposed a novel method to sample subsequences from the compete sequences and employed term frequency inverse document frequency (TFIDF) method to represent the sequences as fixed-dimension numerical-value vectors. Results have demonstrated that classifier using SuperAlarm sequences can achieve higher performance in prediction of code blue events and reduction of alarm burden than that using monitor alarm sequences.

It is of tremendous importance to exploit vectorization methods for representing SuperAlarm sequences so that machine learning algorithms can be readily used to recognize temporal patterns encoded by these sequences. As mentioned in Chapter 3, TFIDF representation has several limitations for representing SuperAlarm sequences. Therefore, in Chapter 5 we proposed a novel representation method to convert SuperAlarm sequences into fixed-dimensional vectors. This representation method, called time weighted supervised sequence representation (TWSSR) is not only a supervised weighting scheme that takes into account the distribution of sequences between coded patients and control patient, it also considers the timing of each SuperAlarm trigger when calculating the weights for them. We employed support vector machine based recursive feature elimination (SVM-RFE) algorithm to perform classification in conjunction with feature selection. The results demonstrate that the performance of the sequence classifier based on TWSSR method is higher than that based on TFIDF method.

6.1.3 Development of SuperAlarm study database

In Chapter 4, we have reported a large-scale and comprehensive patient database that was used for the development and evaluation of the latest SuperAlarm algorithm (and future ones) described in Chapter 5. The SuperAlarm study database II consolidates and aggregates a large volume of temporal physiologic and clinical data, including patient demographics, monitor alarms, laboratory test results, physiologic waveforms and vital signs from coded patients and control patients admitted to UCLA and UCSF Medical Centers. The protected health information was de-identified. We also designed two naming codebooks for automated mapping of monitor alarms and laboratory tests. We further developed a software application to extract waveforms and vital signs, and save them into binary files so that they can be readily available for researchers for further analysis.

6.2 Future work

6.2.1 Use of numeric values of laboratory test results instead of abnormality flags to identify SuperAlarm patterns

In Chapter 2, we proposed a framework to identify SuperAlarm patterns using the data streams of monitor alarms and abnormality flags of laboratory test results. One possible extension of the framework is to consider numeric values of laboratory test results rather than the abnormality flags. One likely approach is to discretize the numeric values into several intervals as what we did for vital sign parameter alarms. For example, “HEMOGLOBIN 10.1 g/dL” and “HEMOGLOBIN 7.8 g/dL” can be represented as a laboratory “alarm” if their values fall into the same interval after discretization (e.g., “$5.5 \leq$ HEMOGLOBIN LO ≤ 11.5 ”). Other potential method for utilization of laboratory values is to analyze the time series of laboratory test results. Unlike physiologic waveforms with regular sampling rate (e.g., 240 Hz), laboratory tests are often ordered at different rate depending on the physiologic process (i.e., irregular sampling rate). However, trend-based approaches such as qualitative shape analysis (QSA) [44] can also be applicable to the irregularly sampled laboratory time series.

6.2.2 Enrichment of SuperAlarm patterns by accommodating non-monitored physiological variables derived from electrocardiographic (ECG) signals

In this dissertation, we integrated laboratory test results with monitor alarms to identify SuperAlarm patterns. However, continuous ECG signals play a crucial role in clinical diagnosis, and the morphological characteristics of these signals timely capture clinical events and reflect the changes in a patient’s condition. Studies have shown that some non-monitored ECG variables derived by secondary analysis of ECG signals can be potential predictors of adverse events such as bradyasystolic cardiac arrest [119, 136]. These non-monitored ECG variables that are not measured by existing bedside monitors can include, for example, PR interval, P-wave duration, QRS duration, RR interval, QT interval, ST segment levels for each available ECG leads and heart rate variability. Therefore, the extension of the framework can also be achieved by accommodating these non-monitored variables, treating them as virtual alarms and integrating them with existing types of data input to the SuperAlarm algorithms.

6.2.3 Development of probabilistic models to recognize temporal patterns in SuperAlarm sequences

In this dissertation, we recognized temporal patterns in SuperAlarm sequences by developing methods to represent the sequences as fixed-dimensional vectors. The sequential nature of SuperAlarm sequences allow developing probabilistic models to explore the underlying mechanism of sequence generation and to recognize temporal patterns in SuperAlarm sequences. One of such models is Hidden Markov Model (HMM) [137], which has been widely used in biomedical domain to model sequential data [138–140]. In addition, it is known that some physiologic variables may be the causal factors of others. For example, decrease in systolic blood pressure may be a concomitant of narrowing of the pulse pressure, potassium may be highly associated with ECG arrhythmia alarms. Hence, different SuperAlarm triggers in the sequences may also have such causality. Therefore, the development of models to capture the relations and also have capability of predicting patient deterioration is needed. One of potential methods may be exploration of Dynamic Bayesian Network (DBN) based models for the sequences of SuperAlarm triggers. DBN [141], which includes HMM as a special case, is a generalized version of the Bayesian network (BN) with an extension to temporal dimension. DBN allows incorporating the representation of causes and effects and the temporal nature among SuperAlarm triggers in sequences. Many studies have been conducted to build prediction model based on DBN in the biomedical domain, such as early detection of sepsis [142], prognosis of carcinoid patients [143], and prediction organ failures [144].

6.2.4 Development of a prototype to evaluate the SuperAlarm system in real time

Further efforts are also needed to develop a prototype to evaluate feasibility of streaming data analytics, true predictive power of the SuperAlarm system in a real-time and prospective manner. The proposed SuperAlarm algorithms that were developed based on a retrospective database should be validated by running them online to track the patient's status. The validation process requires near real-time data streams obtained from bedside physiologic monitors and EHR systems. Hence, the development of a SuperAlarm prototype can provide interfaces for receiving these data streams asynchronously from different data source. The data streams can be continuously buffered and once they are sufficient to be processed, the SuperAlarm prototype will activate a deployed SuperAlarm algorithm, and the validation process can be performed.

References

[1] B. J. Drew, P. Harris, J. K. Zégre-Hemsey, T. Mammone, D. Schindler, R. Salas-Boni, Y. Bai, A. Tinoco, Q. Ding, and X. Hu, "Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients," *PloS one*, vol. 9, no. 10, p. e110274, 2014.

[2] M. A. DeVita, G. B. Smith, S. K. Adam, I. Adams-Pizarro, M. Buist, R. Bellomo, R. Bonello, E. Cerchiari, B. Farlow, D. Goldsmith *et al.*, "Identifying the hospitalised patient in crisis – a consensus conference on the afferent limb of rapid response systems," *Resuscitation*, vol. 81, no. 4, pp. 375–382, 2010.

[3] R. Schein, N. Hazday, M. Pena, B. H. Ruben, and C. L. Sprung, "Clinical antecedents to in-hospital cardiopulmonary arrest." *Chest Journal*, vol. 98, no. 6, pp. 1388–1392, 1990.

[4] A. F. Smith and J. Wood, "Can some in-hospital cardio-respiratory arrests be prevented? a prospective survey," *Resuscitation*, vol. 37, no. 3, pp. 133–137, 1998.

[5] M. D. Buist, E. Jarmolowski, P. R. Burton, S. A. Bernard, B. P. Waxman, and J. Anderson, "Recognising clinical instability in hospital patients before cardiac arrest or unplanned admission to intensive care. a pilot study in a tertiary-care hospital." *The Medical Journal of Australia*, vol. 171, no. 1, pp. 22–25, 1999.

[6] K. Hillman, P. Bristow, T. Chey, K. Daffurn, T. Jacques, S. Norman, G. Bishop, and G. Simmons, "Antecedents to hospital deaths," *Internal medicine journal*, vol. 31, no. 6, pp. 343–348, 2001.

[7] M. Buist, S. Bernard, T. V. Nguyen, G. Moore, and J. Anderson, "Association between clinically abnormal observations and subsequent in-hospital mortality: a prospective study," *Resuscitation*, vol. 62, no. 2, pp. 137–141, 2004.

[8] T. Jacques, G. A. Harrison, M.-L. McLaws, and G. Kilborn, "Signs of critical conditions and emergency responses (soccer): a model for predicting adverse events in the inpatient setting," *Resuscitation*, vol. 69, no. 2, pp. 175–183, 2006.

[9] P. K. Gazarian, E. A. Henneman, and G. E. Chandler, "Nurse decision making in the prearrest period," *Clinical Nursing Research*, 2009.

[10] A. H. Taenzer, J. B. Pyke, and S. P. McGrath, "A review of current and emerging approaches to address failure-to-rescue," *The Journal of the American Society of Anesthesiologists*, vol. 115, no. 2, pp. 421–431, 2011.

[11] M. Imhoff and S. Kuhls, "Alarm algorithms in critical care monitoring," *Anesthesia & Analgesia*, vol. 102, no. 5, pp. 1525–1537, 2006.

[12] B. J. Drew, R. M. Califf, M. Funk, E. S. Kaufman, M. W. Krucoff, M. M. Laks, P. W. Macfarlane, C. Sommargren, S. Swiryn, and G. F. Van Hare, "Practice standards for electrocardiographic monitoring in hospital settings an american heart association scientific statement from the councils on cardiovascular nursing, clinical cardiology, and cardiovascular disease in the young: Endorsed by the international society of computerized electrocardiology and the american association of critical-care nurses," *Circulation*, vol. 110, no. 17, pp. 2721–2746, 2004.

[13] M. Cvach, "Monitor alarm fatigue: an integrative review," *Biomedical Instrumentation & Technology*, vol. 46, no. 4, pp. 268–277, 2012.

[14] C. L. Tsien and J. C. Fackler, "Poor prognosis for existing monitors in the intensive care unit," *Critical care medicine*, vol. 25, no. 4, pp. 614–619, 1997.

[15] M.-C. Chambrin, P. Ravaux, D. Calvelo-Aros, A. Jaborska, C. Chopin, and B. Boniface, "Multicentric study of monitoring alarms in the adult intensive care unit (icu): a descriptive analysis," *Intensive care medicine*, vol. 25, no. 12, pp. 1360–1366, 1999.

[16] C. Atzema, M. J. Schull, B. Borgundvaag, G. R. Slaughter, and C. K. Lee, "Alarmed: adverse events in low-risk patients with chest pain receiving continuous electrocardiographic monitoring in the emergency department. a pilot study," *The American journal of emergency medicine*, vol. 24, no. 1, pp. 62–67, 2006.

[17] S. Siebig, S. Kuhls, M. Imhoff, U. Gather, J. Schölmerich, and C. E. Wrede, "Intensive care unit alarmshow many do we need?*," *Critical care medicine*, vol. 38, no. 2, pp. 451–456, 2010.

[18] M. Borowski, M. Görges, R. Fried, O. Such, C. Wrede, and M. Imhoff, "Medical device alarms," *Biomedizinische Technik/Biomedical Engineering*, vol. 56, no. 2, pp. 73–83, 2011.

[19] K. C. Graham and M. Cvach, "Monitor alarm fatigue: standardizing use of physiological monitors and decreasing nuisance alarms," *American Journal of Critical Care*, vol. 19, no. 1, pp. 28–34, 2010.

[20] S. Sendelbach, "Alarm fatigue," *Nursing Clinics of North America*, vol. 47, no. 3, pp. 375–382, 2012.

[21] "Impact of clinical alarms on patient safety," (Date last accessed March 17, 2016). [Online]. Available: http://thehtf.org/white%20paper.pdf

[22] J. Edworthy and E. Hellier, "The hazards of alarm overload, keeping excessive physiologic monitoring alarms from impeding care," *Health Devices*, vol. 36, no. 3, pp. 73–83, 2007.

[23] P. K. Gazarian, "Nurses response to frequency and types of electrocardiography alarms in a non-critical care setting: a descriptive study," *International journal of nursing studies*, vol. 51, no. 2, pp. 190–197, 2014.

[24] K. M. Weil, "Alarming monitor problems," *Nursing2015*, vol. 39, no. 9, p. 58, 2009.

[25] S. T. Lawless, "Crying wolf: false alarms in a pediatric intensive care unit." *Critical care medicine*, vol. 22, no. 6, pp. 981–985, 1994.

[26] M. Imhoff and R. Fried, "The crying wolf: still crying?" *Anesthesia & Analgesia*, vol. 108, no. 5, pp. 1382–1383, 2009.

[27] F. Schmid, M. S. Goepfert, D. Kuhnt, V. Eichhorn, S. Diedrichs, H. Reichenspurner, A. E. Goetz, and D. A. Reuter, "The wolf is crying in the operating room: patient monitor and anesthesia workstation alarming patterns during cardiac surgery," *Anesthesia & Analgesia*, vol. 112, no. 1, pp. 78–83, 2011.

[28] M. Funk, J. T. Clark, T. J. Bauld, J. C. Ott, and P. Coss, "Attitudes and practices related to clinical alarms," *American Journal of Critical Care*, vol. 23, no. 3, pp. e9–e18, 2014.

[29] ECRI, "2012 top 10 health technology hazards," *Health Devices*, vol. 40, no. 11, 2011.

[30] ECRI, "Top 10 health technology hazards for 2013," *Health Devices*, vol. 41, no. 11, 2012.

[31] ECRI, "Top 10 health technology hazards for 2014," *Health Devices*, vol. 42, no. 11, 2013.

[32] ECRI, "Top 10 health technology hazards for 2015," *Health Devices*, 2014.

[33] ECRI, "Top 10 health technology hazards for 2016," *Health Devices*, 2015.

[34] V. Chopra and L. F. McMahon, "Redesigning hospital alarms for patient safety: alarmed and potentially dangerous," *JAMA*, vol. 311, no. 12, pp. 1199–1200, 2014.

[35] F. A. Drews, "Patient monitors in critical care: Lessons for improvement," 2008.

[36] C. Force, "Impact of clinical alarms on patient safety: a report from the american college of clinical engineering healthcare technology foundation," *Journal of Clinical Engineering*, vol. 32, pp. 22–33, 2007.

[37] J. P. Bliss and M. C. Dunn, "Behavioural implications of alarm mistrust as a function of task workload," *Ergonomics*, vol. 43, no. 9, pp. 1283–1300, 2000.

[38] M. D. Buist, G. E. Moore, S. A. Bernard, B. P. Waxman, J. N. Anderson, and T. V. Nguyen, "Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: preliminary study," *Bmj*, vol. 324, no. 7334, pp. 387–390, 2002.

[39] J. M. Blum and K. K. Tremper, "Alarms in the intensive care unit: Too much of a good thing is dangerous: Is it time to add some intelligence to alarms?*," *Critical care medicine*, vol. 38, no. 2, pp. 702–703, 2010.

[40] J. Moore, M. Hravnak, and M. Pinsky, "Afferent limb of rapid response system activation," in *Annual Update in Intensive Care and Emergency Medicine 2012*. Springer, 2012, pp. 494–503.

[41] F. Schmid, M. S. Goepfert, and D. A. Reuter, "Patient monitoring alarms in the icu and in the operating room," *Crit Care*, vol. 17, no. 2, p. 216, 2013.

[42] M. E. Janusz and V. Venkatasubramanian, "Automatic generation of qualitative descriptions of process trends for fault detection and diagnosis," *Engineering Applications of Artificial Intelligence*, vol. 4, no. 5, pp. 329–339, 1991.

[43] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis: Part iii: Process history based methods," *Computers & chemical engineering*, vol. 27, no. 3, pp. 327–346, 2003.

[44] R. Rengaswamy, T. Hägglund, and V. Venkatasubramanian, "A qualitative shape analysis formalism for monitoring control loop performance," *Engineering Applications of Artificial Intelligence*, vol. 14, no. 1, pp. 23–33, 2001.

[45] S. Charbonnier and S. Gentil, "A trend-based alarm system to improve patient monitoring in intensive care units," *Control Engineering Practice*, vol. 15, no. 9, pp. 1039–1050, 2007.

[46] S. Charbonnier and S. Gentil, "On-line adaptive trend extraction of multiple physiological signals for alarm filtering in intensive care units," *International Journal of Adaptive Control and Signal Processing*, vol. 24, no. 5, pp. 382–408, 2010.

[47] W. Zong, G. Moody, and R. Mark, "Reduction of false arterial blood pressure alarms using signal quality assessement and relationships between the electrocardiogram and arterial blood pressure," *Medical and Biological Engineering and Computing*, vol. 42, no. 5, pp. 698–706, 2004.

[48] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter," *Physiological measurement*, vol. 29, no. 1, p. 15, 2007.

[49] J. Behar, J. Oster, Q. Li, and G. D. Clifford, "Ecg signal quality during arrhythmia and its application to false alarm reduction," *Biomedical Engineering, IEEE Transactions on*, vol. 60, no. 6, pp. 1660–1666, 2013.

[50] R. Salas-Boni, Y. Bai, P. R. E. Harris, B. J. Drew, and X. Hu, "False ventricular tachycardia alarm suppression in the icu based on the discrete wavelet transform in the ecg signal," *Journal of electrocardiology*, vol. 47, no. 6, pp. 775–780, 2014.

[51] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database," *Critical care medicine*, vol. 39, no. 5, p. 952, 2011.

[52] A. Aboukhalil, L. Nielsen, M. Saeed, R. G. Mark, and G. D. Clifford, "Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform," *Journal of biomedical informatics*, vol. 41, no. 3, pp. 442–451, 2008.

[53] Q. Li and G. D. Clifford, "Signal quality and data fusion for false alarm reduction in the intensive care unit," *Journal of electrocardiology*, vol. 45, no. 6, pp. 596–603, 2012.

[54] G. Borges and V. Brusamarello, "Sensor fusion methods for reducing false alarms in heart rate monitoring," *Journal of clinical monitoring and computing*, pp. 1–9, 2015.

[55] H. Gao, A. McDonnell, D. A. Harrison, T. Moore, S. Adam, K. Daly, L. Esmonde, D. R. Goldhill, G. J. Parry, A. Rashidian *et al.*, "Systematic review and evaluation of physiological track and trigger warning systems for identifying at-risk patients on the ward," *Intensive care medicine*, vol. 33, no. 4, pp. 667–679, 2007.

[56] G. B. Smith, D. R. Prytherch, P. E. Schmidt, and P. I. Featherstone, "Review and performance evaluation of aggregate weighted track and triggersystems," *Resuscitation*, vol. 77, no. 2, pp. 170–179, 2008.

[57] R. Morgan, F. Williams, and M. Wright, "An early warning scoring system for detecting developing critical illness," *Clin Intensive Care*, vol. 8, no. 2, p. 100, 1997.

[58] C. Stenhouse, S. Coates, M. Tivey, P. Allsop, and T. Parker, "Prospective evaluation of a modified early warning score to aid earlier detection of patients developing critical illness on a general surgical ward," *British Journal of Anaesthesia*, vol. 84, no. 5, pp. 663–663, 2000.

[59] C. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Validation of a modified early warning score in medical admissions," *Qjm*, vol. 94, no. 10, pp. 521–526, 2001.

[60] J. Gardner-Thorpe, N. Love, J. Wrightson, S. Walsh, and N. Keeling, "The value of modified early warning score (mews) in surgical in-patients: a prospective observational study," *The Annals of The Royal College of Surgeons of England*, vol. 88, no. 6, pp. 571–575, 2006.

[61] U. Kyriacos, J. Jelsma, and S. Jordan, “Monitoring vital signs using early warning scoring systems: a review of the literature,” *Journal of nursing management*, vol. 19, no. 3, pp. 311–330, 2011.

[62] C. Subbe, R. Davies, E. Williams, P. Rutherford, and L. Gemmell, “Effect of introducing the modified early warning score on clinical outcomes, cardio-pulmonary arrests and intensive care utilisation in acute medical admissions*,” *Anaesthesia*, vol. 58, no. 8, pp. 797–802, 2003.

[63] D. R. Prytherch, G. B. Smith, P. E. Schmidt, and P. I. Featherstone, “Viewstowards a national early warning score for detecting adult inpatient deterioration,” *Resuscitation*, vol. 81, no. 8, pp. 932–937, 2010.

[64] L. Tarassenko, A. Hann, A. Patterson, E. Braithwaite, K. Davidson, V. Barber, and D. Young, “Biosign: Multi-parameter monitoring for early warning of patient deterioration,” in *Medical Applications of Signal Processing, 2005. The 3rd IEE International Seminar on (Ref. No. 2005-1119).* IET, 2005, pp. 71–76.

[65] L. Tarassenko, A. Hann, and D. Young, “Integrated monitoring and analysis for early warning of patient deterioration,” *British journal of anaesthesia*, vol. 97, no. 1, pp. 64–68, 2006.

[66] M. J. Rothman, S. I. Rothman, and J. Beals, “Development and validation of a continuous measure of patient condition using the electronic medical record,” *Journal of biomedical informatics*, vol. 46, no. 5, pp. 837–848, 2013.

[67] L. A. Lynn and J. P. Curry, “Patterns of unexpected in-hospital deaths: a root cause analysis,” *Patient safety in surgery*, vol. 5, no. 1, p. 1, 2011.

[68] M.-C. Chambrin *et al.*, “Alarms in the intensive care unit: how can the number of false alarms be reduced?” *CRITICAL CARE-LONDON-*, vol. 5, no. 4, pp. 184–188, 2001.

[69] E. M. Koski, A. Mäkivirta, T. Sukuvaara, and A. Kari, “Clinicians’ opinions on alarm limits and urgency of therapeutic responses,” *International journal of clinical monitoring and computing*, vol. 12, no. 2, pp. 85–88, 1995.

[70] F. Scalzo, D. Liebeskind, and X. Hu, “Reducing false intracranial pressure alarms using morphological waveform features,” *Biomedical Engineering, IEEE Transactions on*, vol. 60, no. 1, pp. 235–239, 2013.

[71] F. Scalzo and X. Hu, “Semi-supervised detection of intracranial pressure alarms using waveform dynamics,” *Physiological measurement*, vol. 34, no. 4, p. 465, 2013.

[72] L. Clifton, D. A. Clifton, P. J. Watkinson, and L. Tarassenko, “Identification of patient deterioration in vital-sign data using one-class support vector machines,”

in *Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on.* IEEE, 2011, pp. 125–131.

[73] L. Tarassenko, D. A. Clifton, M. R. Pinsky, M. T. Hravnak, J. R. Woods, and P. J. Watkinson, "Centile-based early warning scores derived from statistical distributions of vital signs," *Resuscitation*, vol. 82, no. 8, pp. 1013–1018, 2011.

[74] X. Hu, M. Sapo, V. Nenov, T. Barry, S. Kim, D. H. Do, N. Boyle, and N. Martin, "Predictive combinations of monitor alarms preceding in-hospital code blue events," *Journal of biomedical informatics*, vol. 45, no. 5, pp. 913–921, 2012.

[75] C.-J. Tsai, C.-I. Lee, and W.-P. Yang, "A discretization algorithm based on class-attribute contingency coefficient," *Information Sciences*, vol. 178, no. 3, pp. 714–731, 2008.

[76] R. J. Bayardo Jr, "Efficiently mining long patterns from databases," in *ACM Sigmod Record*, vol. 27, no. 2. ACM, 1998, pp. 85–93.

[77] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 1994, pp. 487–499.

[78] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207–216, 1993.

[79] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 1–12.

[80] K. Gouda and M. Zaki, "Efficiently mining maximal frequent itemsets," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on.* IEEE, 2001, pp. 163–170.

[81] A. Rahman and P. Balasubramanie, "An efficient algorithm for mining maximal frequent item sets," *Journal of Computer Science*, vol. 4, no. 8, p. 638, 2008.

[82] D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu, "Mafia: A maximal frequent itemset algorithm," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 11, pp. 1490–1504, 2005.

[83] D.-I. Lin and Z. M. Kedem, "Pincer-search: A new algorithm for discovering the maximum frequent set," in *Advances in Database TechnologyEDBT'98.* Springer, 1998, pp. 103–119.

[84] G. J. Escobar, J. C. LaGuardia, B. J. Turk, A. Ragins, P. Kipnis, and D. Draper, "Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record," *Journal of hospital medicine*, vol. 7, no. 5, pp. 388–395, 2012.

[85] T. Heldt, B. Long, G. C. Verghese, P. Szolovits, and R. G. Mark, “Integrating data, models, and reasoning in critical care,” in *Engineering in Medicine and Biology Society, 2006. EMBS’06. 28th Annual International Conference of the IEEE.* IEEE, 2006, pp. 350–353.

[86] E. J. Huang, C. P. Bonafide, R. Keren, V. M. Nadkarni, and J. H. Holmes, “Medications associated with clinical deterioration in hospitalized children,” *Journal of Hospital Medicine*, vol. 8, no. 5, pp. 254–260, 2013.

[87] R. Kawamoto, A. Nazir, A. Kameyama, T. Ichinomiya, K. Yamamoto, S. Tamura, M. Yamamoto, S. Hayamizu, and Y. Kinosada, “Hidden markov model for analyzing time-series health checkup data.” in *MedInfo*, 2013, pp. 491–495.

[88] I. A. Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper, *The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks.* Springer, 1989.

[89] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, “Text classification using string kernels,” *The Journal of Machine Learning Research*, vol. 2, pp. 419–444, 2002.

[90] M. Borowski et al., “Reducing false alarms of intensive care online-monitoring systems: an evaluation of two signal extraction algorithms,” *Computational and mathematical methods in medicine*, vol. 2011, 2011.

[91] J. W. Dukes et al., “Ventricular ectopy as a predictor of heart failure and death,” *Journal of the American College of Cardiology*, vol. 66, no. 2, pp. 101–109, 2015.

[92] Y. Bai et al., “Integrating monitor alarms with laboratory test results to enhance patient deterioration prediction,” *Journal of biomedical informatics*, vol. 53, pp. 81–92, 2015.

[93] R. Salas-Boni, Y. Bai, and X. Hu, “Cumulative time series representation for code blue prediction in the intensive care unit.” *AMIA Summits on Translational Science Proceedings*, vol. 2015, p. 162, 2015.

[94] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[95] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[96] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[97] E. Osuna et al., “Support vector machines: Training and applications,” 1997.

[98] C. D. Manning et al., *Introduction to information retrieval.* Cambridge university press Cambridge, 2008, vol. 1.

[99] G. Salton et al., "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[100] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[101] J. Hua et al., "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.

[102] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, vol. 97, 1997, pp. 412–420.

[103] J. Li et al., "A gene-based information gain method for detecting gene–gene interactions in case–control studies," *European Journal of Human Genetics*, 2015.

[104] H. Kodaz et al., "Medical application of information gain based artificial immune recognition system (airs): Diagnosis of thyroid disease," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3086–3092, 2009.

[105] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[106] R. Akbani et al., "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004.* Springer, 2004, pp. 39–50.

[107] A. Ben-Hur and J. Weston, "A users guide to support vector machines," in *Data mining techniques for the life sciences.* Springer, 2010, pp. 223–239.

[108] C.-W. Hsu et al., "A practical guide to support vector classification," 2003.

[109] C. Elkan, "Deriving tf-idf as a fisher kernel," in *String Processing and Information Retrieval.* Springer, 2005, pp. 295–300.

[110] R.-E. Fan et al., "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[111] V. Chandola et al., "Anomaly detection for discrete sequences: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 823–839, 2012.

[112] K. Desai et al., "Hemodynamic-impact-based prioritization of ventricular tachycardia alarms," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE.* IEEE, 2014, pp. 3456–3459.

[113] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[114] F. J. Provost et al., "The case against accuracy estimation for comparing induction algorithms." in *ICML*, vol. 98, 1998, pp. 445–453.

[115] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine learning*, vol. 42, no. 3, pp. 203–231, 2001.

[116] C. J. McDonald, S. M. Huff, J. G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook *et al.*, "Loinc, a universal standard for identifying laboratory observations: a 5-year update," *Clinical chemistry*, vol. 49, no. 4, pp. 624–633, 2003.

[117] J. J. Frassica, "Frequency of laboratory test utilization in the intensive care unit and its implications for large-scale data collection efforts," *Journal of the American Medical Informatics Association*, vol. 12, no. 2, pp. 229–233, 2005.

[118] "Generating labchart for windows binary files," (AD Instrument). [Online]. Available: http://cdn.adinstruments.com/adi-web/manuals/translatebinary/LabChartBinaryFormat.pdf

[119] Q. Ding, Y. Bai, A. Tinoco, D. Mortara, D. Do, N. G. Boyle, M. M. Pelter, and X. Hu, "Developing new predictive alarms based on ecg metrics for bradyasystolic cardiac arrest," *Physiological measurement*, vol. 36, no. 12, p. 2405, 2015.

[120] J. Arroyo-Palacios, M. Rudz, R. Fidler, W. Smith, N. Ko, S. Park, Y. Bai, and X. Hu, "Characterization of shape differences among icp pulses predicts outcome of external ventricular drainage weaning trial," *Neurocritical Care*, pp. 1–10, 2016.

[121] J. Welch, "An evidence-based approach to reduce nuisance alarms and alarm fatigue," *Biomedical Instrumentation & Technology*, vol. 45, no. s1, pp. 46–52, 2011.

[122] F. E. Block, L. Nuutinen, and B. Ballast, "Optimization of alarms: A study on alarm limits, alarm sounds, and false alarms, intended to reduce annoyance," *Journal of Clinical Monitoring and Computing*, vol. 15, no. 2, pp. 75–83, 1999.

[123] M. Görges, B. A. Markewitz, and D. R. Westenskow, "Improving alarm performance in the medical intensive care unit using delays and clinical context," *Anesthesia & Analgesia*, vol. 108, no. 5, pp. 1546–1552, 2009.

[124] M. C. Mozer, "Neural net architectures for temporal sequence processing," in *A. Weigend, N. Gershenfeld (Eds.),Predicting the Future and Understanding the Past, Santa Fe Institute Studies in the Science of Complexity*, vol. 15. Addison-Wesley, Redwood City, 1993, pp. 243–243.

[125] M. Lan, C. L. Tan, J. Su, and Y. Lu, “Supervised and traditional term weighting methods for automatic text categorization,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 721–735, 2009.

[126] Z. Li, Z. Xiong, Y. Zhang, C. Liu, and K. Li, “Fast text categorization using concise semantic analysis,” *Pattern Recognition Letters*, vol. 32, no. 3, pp. 441–448, 2011.

[127] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, “Support vector machines and kernels for computational biology,” *PLoS Comput Biol*, vol. 4, no. 10, p. e1000173, 2008.

[128] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[129] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, “Feature selection for svms,” in *NIPS*, vol. 12. Citeseer, 2000, pp. 668–674.

[130] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[131] Y. LeCun, J. S. Denker, S. A. Solla, R. E. Howard, and L. D. Jackel, “Optimal brain damage.” in *NIPs*, vol. 89, 1989.

[132] Y. Bai, D. H. Do, Q. Ding, J. Arroyo-Palacios, Y. Shahriari, M. M. Pelter, N. Boyle, R. Fidler, and X. Hu, “Is the sequence of superalarm triggers more predictive than sequence of the currently utilized patient monitor alarms?” *Biomedical Engineering, IEEE Transactions on*, Submitted, 2016.

[133] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, “Robust biomarker identification for cancer diagnosis with ensemble feature selection methods,” *Bioinformatics*, vol. 26, no. 3, pp. 392–398, 2010.

[134] E. Hemphill, J. Lindsay, C. Lee, I. I. Măndoiu, and C. E. Nelson, “Feature selection and classifier performance on diverse biological datasets,” *BMC bioinformatics*, vol. 15, no. Suppl 13, p. S4, 2014.

[135] F. Zhang, H. L. Kaufman, Y. Deng, and R. Drabier, “Recursive svm biomarker selection for early detection of breast cancer in peripheral blood,” *BMC medical genomics*, vol. 6, no. 1, p. 1, 2013.

[136] X. Hu, D. Do, Y. Bai, and N. G. Boyle, “A case–control study of non-monitored ecg metrics preceding in-hospital bradyasystolic cardiac arrest: Implication for predictive monitor alarms,” *Journal of electrocardiology*, vol. 46, no. 6, pp. 608–615, 2013.

[137] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[138] K. Karplus, C. Barrett, and R. Hughey, "Hidden markov models for detecting remote protein homologies." *Bioinformatics*, vol. 14, no. 10, pp. 846–856, 1998.

[139] X. Qian and B.-J. Yoon, "Effective identification of conserved pathways in biological networks using hidden markov models," *PLoS One*, vol. 4, no. 12, p. e8070, 2009.

[140] N. A. Bykova, A. V. Favorov, and A. A. Mironov, "Hidden markov models for evolution and comparative genomics analysis," *PloS one*, vol. 8, no. 6, p. e65012, 2013.

[141] K. P. Murphy, "Dynamic bayesian networks," *Probabilistic Graphical Models, M. Jordan*, vol. 7, 2002.

[142] S. K. Nachimuthu and P. J. Haug, "Early detection of sepsis in the emergency department using dynamic bayesian networks," in *AMIA Annual Symposium Proceedings*, vol. 2012. American Medical Informatics Association, 2012, p. 653.

[143] M. A. Van Gerven, B. G. Taal, and P. J. Lucas, "Dynamic bayesian networks as prognostic models for clinical patient management," *Journal of biomedical informatics*, vol. 41, no. 4, pp. 515–529, 2008.

[144] M. Sandri, P. Berchialla, I. Baldi, D. Gregori, and R. A. De Blasi, "Dynamic bayesian networks to predict sequences of organ failures in patients admitted to icu," *Journal of biomedical informatics*, vol. 48, pp. 106–113, 2014.