Dynamic Machine Learning using Signal Processing and Tensor-Based Methods to Predict Clinical Outcomes

by

Olivia I. Pifer Alge

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2024

Doctoral Committee:

Professor Kayvan Najarian, Chair
Professor Harm Derksen, Northeastern University
Associate Professor Alla Karnovsky
Professor Jun Li
Associate Professor J. Scott VanEpps
Associate Professor Joshua Welch

Olivia Pifer Alge
oialge@umich.edu
ORCID iD: 0000-0002-1029-6664

DEDICATION

For my parents, Myrna and Ernesto, for my husband, Kaleb, for my best friends, Catherine and the Kathryns, and lastly, in memory of my grandmother, Oralia, and uncle, Sergio.

ACKNOWLEDGEMENTS

I am very grateful to those who made this work possible, be it through mentorship, support, friendship, or love. First, I thank Dr. Kayvan Najarian for giving me the opportunity to pursue a Ph.D. as part of his lab.

I thank my other committee members for their advice and support throughout my studies: Dr. J. Scott VanEpps, Dr. Alla Karnovsky, Dr. Harm Derksen, Dr. Jun Li, and Dr. Joshua Welch. I would like to particularly thank Dr. Karnovsky, first for acting as my mentor during my Master's studies, and second for introducing me to Dr. Najarian. I would not have pursued further study in this field without you. I am also grateful to Dr. VanEpps, who greatly influenced the design of my study; thank you for always being willing to have long and in-depth discussions about the clinical relevance and implications of the work that I pursue. Also a thank you to Dr. Gil Omenn, for providing additional advice as I pursued my sepsis research.

I would like to acknowledge the roles of the BCIL lab managers Dr. Jonathan Gryak and Emily Wittrup, who always showed me support, and helped shape me into a stronger and more conscientious researcher.

I acknowledge the support I received from the Department of Computational Medicine and Bioinformatics administrative staff and faculty, including Dr. Maureen Sartor, Dr. Margit Burmeister, Helen Severino, Julia Eussen, and also from Ivana Tullett.

I'm very thankful for all of the students, researchers, and faculty I have met through DCMB, my lab, and my studies here at the University of Michigan.

I acknowledge the funding I have received from University of Michigan NIH NIGMS Bioinformatics Training Grant of the National Institutes of Health under award number T32GM07044 and NSF grant no. 1837985. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Though much time has passed, I am also grateful to Dr. Joel Swanson, Dr. Jenna Wiens, and Mrs. Tina Larson for inspiring me and supporting me in pursuing computational medicine.

Lastly, I give great thanks to my friends and family. Thank you, Catherine Barnier, for being my absolute best friend in the world. Thank you, Kaleb Pifer Alge, for being

my partner, for giving me your love and support and strength, and for being a wonderful husband. Thank you to my parents, for helping me to pursue my education here at University of Michigan, raising me in a warm and loving home, and being there for me even when times were very dark. Finally, I want to remember my grandmother and my uncle Sergio. I had so wished you could see me defend, but for today, I hope that wherever you may be, I've made you proud.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ABBREVIATIONS

3FCV Three-Fold Cross-Validation

ABP Arterial Blood Pressure

ALS Alternating Least Squares

AUPRC Area Under the Precision-Recall Curve

AUROC Area Under the Receiver Operating Characteristic Curve

BP Blood Pressure

BVP Blood Volume Pulse

CDSS Clinical Decision Support System

CP Canonical Polyadic / Parallel Factors

CP-ALS Canonical-Parallel Factors Decomposition using Alternating Least Squares

DTCWPT Dual-Tree Complex Wavelet Packet Transform

ECG Electrocardiogram

EDA Electrodermal Activity

EHR Electronic Health Record

FFT Fast Fourier Transform

GCS Glasgow Coma Scale

HF High Frequency

HOSVD Higher Order Singular Value Decomposition

HRV Heart Rate Variability

ICU Intensive Care Unit

LF Low Frequency

LUCCK Learning Using Concave and Convex Kernels

LUPI Learning Using Privileged Information

ML Machine Learning

PC Principal Component

PCA Principal Component Analysis

PI Privileged Information

PPG Pulse Plethysmography

qSOFA Quick Sequential Organ Failure Assessment

RF Random Forest

RMS Root Mean Square

SD Standard Deviation

SOFA Sequential Organ Failure Assessment

SVM Support Vector Machine

TS Taut String

UTI Urinary Tract Infection

VLF Very Low Frequency

ABSTRACT

Machine learning and artificial intelligence can be used in improving patient care by providing important insights from the data generated throughout the duration of a patient's stay. Routinely collected data, such as electrocardiogram and electronic health record data, are two such examples of data that are frequently recorded in hospital settings. Electrocardiogram, specifically, is a noninvasive and continuously updated measure of a patient's cardiac electric activity, and as such, has the potential to provide a real-time view of a patient's current status.

This research is composed of four projects. In the first, we propose a system of signal processing for both heart rate variability and electrodermal activity to detect poor sleep quality of people with fibromyalgia, using a wearable device. In the second, we introduce a framework of processing signals using both Taut String and tensor decomposition to (1) extract meaningful features from input signals and (2) reduce the feature space to only the most pertinent information, while maintaining structural information from the input signals. This framework is applied to three cohorts of patients from Michigan Medicine, with each cohort increasing in heterogeneity. The physiological signals and electronic health record information collected from the patients in each cohort were used to predict adverse outcomes post-surgery. This study serves as validation to previous work on post-cardiac surgery, as well as generalizing the methodology outwards to other types of surgery. The third project uses the framework developed in the second for patients in the intensive care unit at risk to develop sepsis. The goal is, using continuous electrocardiogram, arterial line, and/or electronic health record data, to predict which patients are at more risk to develop poor outcomes related to sepsis. The fourth project expands upon the third by incorporating the learning using privileged information paradigm into the same sepsis prognosis design.

Together, the four projects discussed in this thesis contribute to dynamic trajectory prediction using signal processing in different health contexts. These studies demonstrate that further study of electrocardiogram data's utility in clinical decision support systems is warranted.

CHAPTER I

Introduction

1.1 Motivation

This work is motivated by a desire to incorporate the continuous nature of physiological signals into the prediction of health-related outcomes. A patient's status is dynamic during their care, and as such, prognostic methods should take this dynamism into account. The research efforts of this thesis have applications in signal processing and predicting the onset of adverse outcomes, with the ultimate goal of improving patient care. This includes the development of methods that leverage structural information included in tensor representations of physiological signals.

Signals like Electrocardiogram (ECG), which are both noninvasive and routinely collected in hospital settings, provide an opportunity for improving disease prognosis through the development of Clinical Decision Support System (CDSS)s that can use these continuous signals to create constantly updating views of a patient's current health status and risk for adverse outcomes. Used alongside Electronic Health Record (EHR) data and expert clinical knowledge, these CDSSs are expected to enhance the quality of care in the emergency room and in the Intensive Care Unit (ICU), where healthcare professionals often experience burnout [52].

1.2 Background

1.2.1 Electronic Health Record Data

Information from the EHR has previously been used to detect, monitor, or predict health outcomes [65, 79, 47, 45, 70]. EHR data can include static variables such as a patient's sex, age or comorbidities, or dynamic variables such as vital signs or lab values. EHR data is invaluable for determining a patient's status, but it is limited by time.

Discrete variables such as lab values require time for collection and processing. Similarly, variables that are updated over time such as fluid output or vital signs may be updated sporadically, or at inconsistent time intervals. In contrast, recordings generated from electrocardiography, blood pressure monitoring, or pulse oximetry, are collected continuously in real time with electronic monitoring devices.

1.2.2 Taut String

Taut String (TS) is a method of creating a piecewise linear approximation of a given signal using an input parameter, ϵ. Several chapters in this thesis make use of TS for signal processing.

Given a discrete signal $f = (f_1, f_2, \ldots, f_n)$, we define the first-order finite difference as

$$\text{diff}(f) = (f_2 - f_1, \ldots, f_n - f_{n-1}). \tag{1.1}$$

We use the parameter $\epsilon > 0$ to create the piecewise linear function g, the TS estimate of f, such that the max norm of $(f - g)$ is $\leq \epsilon$ and the Euclidean norm of diff(g) is minimal, defined below as

$$\|f - g\|_\infty = \max_i \{\|f_i - g_i\|\} \leq \epsilon \tag{1.2}$$

and

$$\|\text{diff}(g)\|_2 = \sqrt{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}, \tag{1.3}$$

respectively.

Next, diff* is defined as

$$\text{diff}^*(y_1, y_2, \ldots, y_{n-1}) = (-y_1, y_1 - y_2, \ldots, y_{n-2} - y_{n-1}, y_{n-1}). \tag{1.4}$$

Here, diff$^* : \mathbb{R}^{n-1} \to \mathbb{R}^n$ is dual to diff $: \mathbb{R}^n \to \mathbb{R}^{n-1}$. The function g also minimizes

$$\|\text{diff}^*\text{diff}\|_1 = \|g_2 - g_1\| + \sum_{i=2}^{n-1} \|g_{i-1} - 2g_i + g_{i+1}\| + \|g_n - g_{n-1}\|. \tag{1.5}$$

One can visualize the function g as a string pulled tightly between $f + \epsilon$ and $f - \epsilon$. Generally, the resulting piecewise linear function is a smoother line than f. Figure 1.1 shows a sample TS estimate of ECG with $\epsilon = 0.3050$.

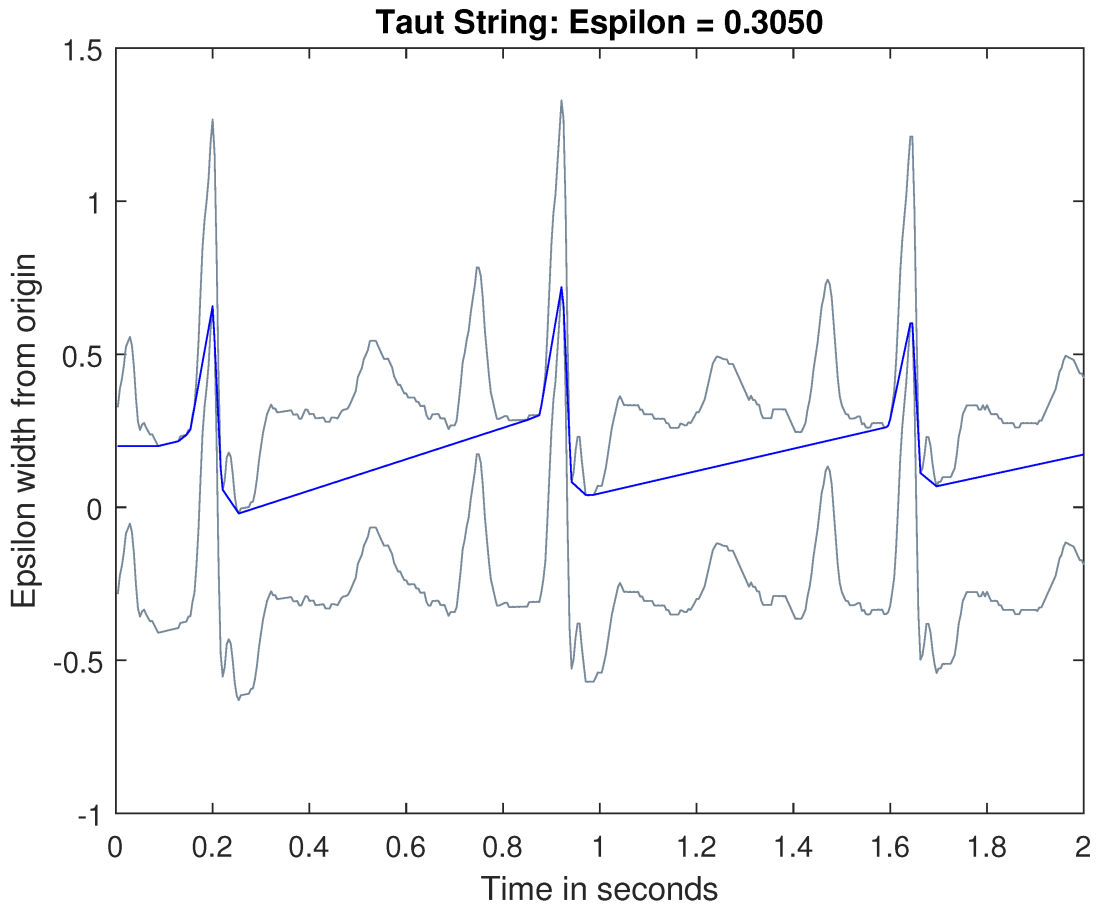


Figure 1.1: TS Approximation of ECG Signal.

The grey waveforms are the margins for TS, $f+\epsilon$ and $f-\epsilon$, the blue line is the TS estimation.

1.2.3 Tensor Reduction With Canonical Polyadic Decomposition

Because Machine Learning (ML) methods can be influenced by the number of features presented to them, we sometimes opt to use feature reduction methods to reduce the feature space of a dataset. If the data is formatted as a tensor, Canonical Polyadic / Parallel Factors (CP) decomposition is one method of feature reduction [35]. This allows the underlying tensor structure to be preserved in the reduced feature space.

In this work, we use CP decomposition to obtain the factor matrices of an input tensor, T. A CP decomposition breaks the initial tensor down into a sum of rank-1 tensors, so it can be considered an extension of singular value decomposition to a higher order. In the chapters using CP decomposition, T has four modes, and thus the CP decomposition results in four factor matrices: (A, B, C, D). Therefore, given a tensor $T = \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$ and rank r, the CP decomposition produces the tensor

$$\hat{T} = \sum_{i=1}^{r} a_i \otimes b_i \otimes c_i \otimes d_i \tag{1.6}$$

where $\otimes$ denotes the Kronecker product, and $\|T - \hat{T}\|$ is minimized, where the Frobenius norm of a tensor $\|\cdot\|$ is defined as

$$\|T\| = \sqrt{\sum_{i_1,\ldots,i_m} (t_{i_1 \ldots i_m})^2}, \tag{1.7}$$

where $t_{i_1 \ldots i_m}$ is the $(i_1, \ldots, i_m)$ entry of T.

The multiplication of vectors $a_i, \ldots d_i$ yields a component rank-1 tensor. The vectors $a_i, \ldots, a_r \in \mathbb{R}^{n_1}$, and so on, can be combined to form factor matrices, such as $A = [a_1, \ldots, a_r] \in \mathbb{R}^{n_1 \times r}$, and similarly for B, C, D. In this manner, each mode of the original tensor T can be approximated by the product of these factor matrices, such as:

$$T_{(1)} \approx A \left(D \odot C \odot B\right)^{\top} \tag{1.8}$$

where $\odot$ denotes the Khatri-Rao product, and $T_{(1)}$ is the matricization of T along its first mode.

We use the Alternating Least Squares (ALS) heuristic method to solve the CP decomposition, as solving for the CP decomposition is known to be NP-hard [27]. Canonical-Parallel Factors Decomposition using Alternating Least Squares (CP-ALS) is an iterative algorithm to find the best approximation of T for a given rank r using randomly generated starting factors [35]. Rank is selected as appropriate for the application in each chapter. After solving for $\hat{T}$, we reserve the necessary factor matrices for feature extraction. For example, if the

factor matrices $A - D$ represent the modes $\epsilon, \text{feature}, \text{time}, \text{patient}$, we reserve A and C to compute the feature matrix for a given patient.

When performing CP-ALS we use a fit score defined as

$$\text{fit} = 1 - \frac{\|T - \hat{T}\|}{\|T\|} \tag{1.9}$$

to determine how well the reduced tensor approximated the original. This process is repeated as appropriate, with the selected reduction being the one with the highest fit after a specified number of iterations, or the first reduction with fit score equal to one, whichever occurred first.

In the relevant chapters that use CP-ALS for feature reduction, CP-ALS is first applied to the training data, and the appropriate factor matrices retained to be able to solve for a patient's feature matrix.

With this process completed, for any given individual's third-order tensor T, a reduced set of features was extracted using the factor matrices computed from the training data. The feature vectors $b_{T,1}, \ldots, b_{T,r}$ were computed via a least squares problem, where

$$\|T - \sum_{i=1}^{r} a_i \otimes b_i \otimes c_{T,i}\| \tag{1.10}$$

was minimal. After computing the individual vectors, they were concatenated to create B_T, a feature matrix with a reduced set of features compared to matricization $T_{(2)}$ of the original tensor T along the feature mode.

1.3 Outline of Thesis

In this work, I developed and applied various signal processing methods to integrate ECG into the prediction of different health outcomes. This thesis is an interdisciplinary research effort that includes the domains of signal processing, ML, tensor methods, and healthcare.

In Chapter II, I present a signal processing method for analyzing ambulatory recordings of people with and without fibromyalgia. I propose a method of analyzing the electrodermal activity and heart rate variability recorded from a wrist sensor to compare against user-reported sleep quality, with the goal of using these processed signals to be used as a more objective measure of sleep quality or "restfulness".

In Chapter III, I describe a multimodal approach, using signal processing of ECG and other physiological signals, in addition to EHR data, collected from a post-surgery cohort to predict adverse outcomes. The methods of Taut String and Dual Tree Complex Wavelet

Packet Transform are used to generate tensors and extract signal features, and different tensor processing methods are introduced for feature reduction.

Chapter IV builds upon the Taut String and tensor methods of Chapter III, but applied now to a cohort of patients in the ICU at risk for developing poor outcomes related to sepsis. Chapter V uses learning using privileged information as a different method of predicting poor outcomes related to sepsis, as an extension of the analyses performed in Chapter IV.

Chapter VI provides a conclusion to the research performed in this thesis. Also addressed is the future direction of incorporating signal processing and tensor methods into CDSSs and healthcare settings.

CHAPTER II

Processing Signals from a Wrist Sensor to Predict Poor Sleep Quality

2.1 Introduction

Fibromyalgia is a medical condition characterized by chronic, widespread pain. The mean prevalence of fibromyalgia globally is 2.7%, and when broken down by sex, is 4.2% in women and 1.4% in men, giving a female-to-male ratio of 3:1 [56]. Besides pain, fibromyalgia is also associated with cognitive dysfunction [36], fatigue, and disturbed sleep [57]. More than 85% of people with fibromyalgia can experience symptoms of fatigue or disturbed sleep [57], which negatively impacts quality of life [72, 71].

One way to measure sleep quality is for patients to report it themselves. One such self-reported measure is the Patient-Reported Outcomes Measurement Information System (PROMIS) [10]. Self-reported scores can give insight to an individual's condition, but may not generalize across individuals, or even necessarily correspond to objective measures of sleep, such as polysomnography, or sleep studies. Specifically in fibromyalgia, unrefreshing sleep can be caused by autonomic nervous system imbalance, which can be recorded unobtrusively through ambulatory measures of Heart Rate Variability (HRV) and Electrodermal Activity (EDA) [73, 60], creating another method of more objectively modeling quality of sleep.

The Empatica E4 wristband (Empatica, Milano, Italy) is a sensor worn on the wrist that provides raw Blood Volume Pulse (BVP), EDA, temperature, heart rate, and 3-axis accelerometer signals [17]. Shaped like a watch, the E4 provides an unobtrusive and direct way to obtain these continuous signals over time. Other studies have previously investigated the utility of wrist sensors in assessing sleep quality. Sabeti et al. [60] used the E4 to assess different levels of fatigue and feeling refreshed using features from the BVP, EDA, temperature, and accelerometer signals, and Sano et al. [61] compared EDA data collected from a wrist sensor across different stages of sleep.

Acknowledging the limitations of wearable technology, the goal of this research was to determine the instances in which sleep quality had been negatively affected in fibromyalgia using ambulatory recordings of HRV, which was estimated using BVP data. In addition to BVP, we also calculated features from EDA data and tested their utility in the assessment of sleep quality. This furthers the pursuit of developing an automated system of sleep quality assessment. To aid in achieving this goal, we used Learning Using Concave and Convex Kernels (LUCCK) [60] and Support Vector Machine (SVM) [14] as classification methods. This study was originally published in the 2020 proceedings of the IEEE Engineering in Medicine and Biology Society Conference [1].

2.2 Methods

2.2.1 Dataset

The University of Michigan Medical Institutional Review Board approved the protocol prior to the initiation of any study activities. Volunteers provided written informed consent prior to being enrolled. We used a dataset consisting of 74 observations, or nights of sleep, across 26 unique individuals, with each having 1-6 observations. Each participant reported the night's sleep quality on a scale of 0-10. The observations in the fibromyalgia group had a self-reported sleep score of ≤ 2, indicating poor sleep, and those in the control group had a self-reported sleep score of ≥ 8, indicating restful sleep. These observations included 13 individuals in the control group, and 13 individuals in the fibromyalgia group. The average age in the control group was 44.9 years, with a Standard Deviation (SD) of 16.2, and 41.9 years in the fibromyalgia group, with SD of 13.1. Each observation contained BVP, EDA, temperature, and accelerometer signals captured by the E4 wristband, but for this project, we focused only on BVP and EDA data for feature extraction.

2.2.2 Feature Extraction

Each individual in the study pressed a button on the E4 wristband to record daily sleep and wake times. Upon reviewing the data, some accelerometer samples showed movement after the marked sleep time, indicating the individual was still awake. For these samples, we marked a sleep time which occurred after the accelerometer signal showed reduced movement, indicating the individual had fallen asleep.

2.2.2.1 BVP Data

The BVP signal was passed through an eighth-order Butterworth bandpass filter, with cutoff frequencies 0.08 and 3 Hz, to remove noise, then separated into non-overlapping 5-minute windows. The formula for a Butterworth filter is:

$$|H(j\omega)| = \frac{1}{\sqrt{1 + \left(\frac{\omega}{\omega_c}\right)^{2m}}},$$

where $|H(j\omega)|$ is the transfer function at the angular frequency ω, ω_c is the cutoff frequency as an angular value, and m is the number of elements in the filter.

A total of 46 BVP features were extracted for each window of signal. Features included select Kubios features [69], the number of peaks, and measurements related to the identified peaks. The full list of features is included in Table A.1.

2.2.2.2 EDA Data

We created approximations of the phasic and tonic components of the EDA signal [68], and extracted features from each of these approximations. To approximate the tonic component, we applied a third-order low-pass Butterworth filter with cutoff frequency 0.04 Hz to the EDA signal. To approximate the phasic component, we generated a 3-level filter bank using a Haar wavelet, generating 8 output samples from the EDA signal. An illustration of the filter bank is provided in Figure 2.1, showing how the input signal was high- and low-pass filtered at each level to yield the output coefficients.

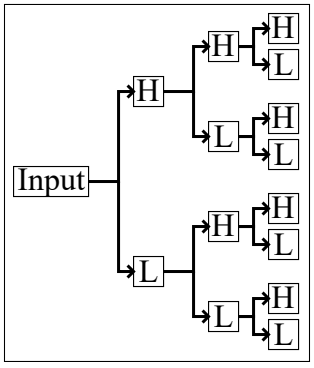


Figure 2.1: Illustration of Filter Bank Structure

Both the phasic and tonic components were separated into non-overlapping 5-minute windows. The features extracted are listed in Table A.1, where there are 58 EDA features for every 5 minutes of signal.

Feature extraction led to a total of 104 features for each window. For each observation, the 5-minute window features were separated into 3 blocks: the first 90 minutes of sleep, the last 90 minutes, and the remaining time in between. The mean, SD, kurtosis, and skewness were calculated for all features within the blocks, yielding a total of 1,248 features (552 BVP, 696 EDA). The formulae used for kurtosis and skewness are defined below, where n is the number of observations.

$$\text{Kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2\right)^2}$$

$$\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2}\right)^3}$$

2.2.3 Feature Reduction and Model Development

Before performing any feature reduction steps, the data were randomly divided into a training set of 20 individuals (10 fibromyalgia, 10 control) and a test set of 6 individuals (3 fibromyalgia, 3 control). Individuals did not overlap between the training and test sets.

We used a two-step process to reduce the number of features in the training set. First, we removed features with redundant information. We tested if the absolute value of the difference between two features was greater than 1e-12 for all observations, and if so, removed one of the features. Second, we used Principal Component Analysis (PCA). We trained the model using the number of Principal Component (PC)s required to reach a specific amount of variance explained. Upon introduction of validation or testing data to the model, we removed any features that were marked as duplicates in the training set, and extracted PCs using the coefficients that were calculated from the training set.

The classification step relied on the LUCCK method [60]. One advantage of LUCCK is that it allows for large deviations in features with a moderate penalty. It is possible that the signals in our dataset suffered measurement errors due to movement, loss of contact between the wristband and wrist, or other problems, and LUCCK would provide more flexibility in these cases. A broad overview of LUCCK is provided in Figure 2.2.

LUCCK functions as follows: suppose the feature space is comprised of real-valued vectors $\mathbf{x} \in \mathbb{R}^n$. The similarity function $Q : \mathbb{R}^n \to \mathbb{R}$ measures how close $\mathbf{x}$ lies to the origin and

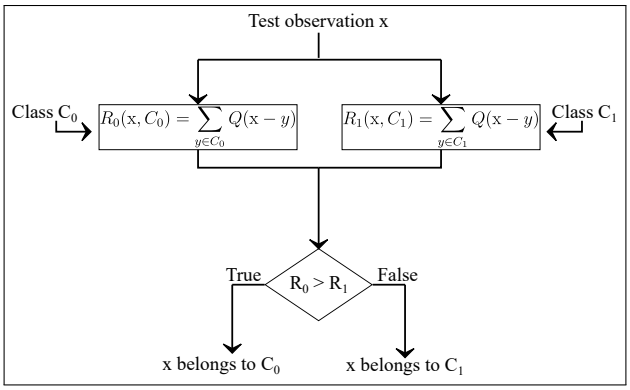


Figure 2.2: Basic Schematic of Learning Using Concave and Convex Kernels

satisfies the properties:

1. $Q(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathbb{R}^n$
2. $Q(\mathbf{x}) = Q(-\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$
3. $Q(\lambda \mathbf{x}) > Q(\mathbf{x})$ if $\mathbf{x} \in \mathbb{R}^n$ is nonzero and $|\lambda| < 1$.

The value $Q(\mathbf{x} - \mathbf{y})$ then provides a measure of the closeness between vectors $\mathbf{x}$ and $\mathbf{y}$, which can then be used for classification. The function

$$Q(\mathbf{x}) = \prod_{i=1}^{n} (1 + \lambda_i x_i^2)^{-\theta_i}$$

for some parameters $\lambda_i, \theta_i > 0$ is used [60], as it can behave similarly to a Cauchy distribution. This allows the function to be less sensitive to errors in individual features.

In this dataset, there are two classes: C_0, the control group, and C_1, the fibromyalgia group. The test vector $\mathbf{x}$ can be classified into class C_k, where k is chosen to maximize the function

$$R(\mathrm{x}, Y) = \sum_{\mathrm{y} \in Y} Q(\mathrm{x} - \mathrm{y}),$$

where Y is the set of training samples in C_k.

Three models were trained using the LUCCK method: one with EDA features, one with BVP features, and one with BVP features coupled with EDA features (BVP+EDA). To serve as a baseline for comparison to the LUCCK method, we also trained three SVM [14] models using the three datasets. SVM constructs the optimal separating hyperplane between

the two classes, which is formulated as:

$$\min_{\mathrm{w},b,\xi} \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{j}^{N} \xi_j$$

with the constraints

$$y_i(\mathrm{w} \cdot \psi(\mathrm{x}_j) + b) \geq 1 - \xi_j, \quad j = 1, ..., N$$
$$\xi \geq 0,$$

where b is the bias term, w is the weight vector, ξ_j functions as a slack variable, C is the penalty parameter, and $\psi(\mathrm{x})$ indicates a mapping of x to a different space [14]. These allow for soft-margin decision boundaries, as fibromyalgia and control observations are not linearly separable.

We implemented a voting system to create final predictions against the test set. The training data was split into 5 folds. When creating training folds, the underrepresented group was oversampled so that the number of observations from each group would be equal. Multiple observations from the same individual were kept in the same fold to prevent data leakage. In each iteration, 4 folds were used to train the model, the fifth fold was used to tune hyperparameters, and the model with selected hyperparameters was run against the test set, producing one set of votes per iteration. The votes from the 5 models were collected and the median used to determine final predictions.

The process of generating partitions into training/test sets, training 5 models, and creating final predictions was repeated 100 times, yielding means and standard deviations of the results.

2.3 Results

The results from the models trained with BVP, EDA, and BVP+EDA data are provided in Tables 2.1, 2.2, and 2.3, respectively. The mean is provided, with SD in parentheses. Results include F1 score, Sensitivity, Specificity, and Area Under the Receiver Operating Characteristic Curve (AUROC). Here, F1 score is defined as:

$$\mathrm{F1} = \left(\frac{2}{\mathrm{Precision}^{-1} + \mathrm{Recall}^{-1}}\right),$$

$$\text{Precision} = \frac{\text{true positive}}{\text{predicted positive}}$$

$$\text{Recall} = \frac{\text{true positive}}{\text{actual condition positive}}$$

Table 2.1: BVP Data Only

Variance Explained	60%	65%	70%
LUCCK			
AUROC	0.66 (0.12)	0.65 (0.12)	0.64 (0.12)
F1	0.62 (0.13)	0.62 (0.12)	0.60 (0.13)
Sensitivity	0.83 (0.16)	0.83 (0.16)	0.81 (0.18)
Specificity	0.49 (0.24)	0.48 (0.23)	0.46 (0.22)
SVM			
AUROC	0.65 (0.15)	0.64 (0.15)	0.62 (0.15)
F1	0.60 (0.18)	0.59 (0.17)	0.56 (0.18)
Sensitivity	0.73 (0.23)	0.70 (0.23)	0.67 (0.24)
Specificity	0.58 (0.26)	0.58 (0.26)	0.58 (0.25)

Table 2.2: EDA Data Only

Variance Explained	60%	65%	70%
LUCCK			
AUROC	0.52 (0.14)	0.51 (0.13)	0.53 (0.13)
F1	0.33 (0.21)	0.33 (0.20)	0.34 (0.21)
Sensitivity	0.31 (0.23)	0.30 (0.21)	0.30 (0.22)
Specificity	0.73 (0.20)	0.73 (0.20)	0.75 (0.20)
SVM			
AUROC	0.53 (0.16)	0.54 (0.15)	0.55 (0.15)
F1	0.43 (0.20)	0.44 (0.19)	0.46 (0.19)
Sensitivity	0.478 (0.25)	0.49 (0.25)	0.49 (0.23)
Specificity	0.58 (0.22)	0.59 (0.21)	0.61 (0.22)

2.4 Discussion

Aligning with the goals of our research, using BVP data with LUCCK and SVM achieved mean AUROC > 0.65 across different partitions of training/testing data. The performance of the LUCCK method are comparable with SVM, with LUCCK generally achieving lower SDs in the performance metrics than SVM on the BVP dataset.

Table 2.3: BVP and EDA Data

Variance Explained	60%	65%	70%
LUCCK			
AUROC	0.61 (0.15)	0.60 (0.15)	0.60 (0.14)
F1	0.51 (0.19)	0.50 (0.20)	0.50 (0.19)
Sensitivity	0.55 (0.26)	0.54 (0.26)	0.54 (0.26)
Specificity	0.67 (0.25)	0.67 (0.24)	0.67 (0.23)
SVM			
AUROC	0.60 (0.17)	0.61 (0.16)	0.61 (0.16)
F1	0.55 (0.18)	0.55 (0.17)	0.55 (0.18)
Sensitivity	0.64 (0.24)	0.65 (0.24)	0.64 (0.26)
Specificity	0.57 (0.27)	0.56 (0.26)	0.58 (0.26)

Features from EDA data alone are unable to achieve the same level of AUROC as BVP data, although with LUCCK, specificity is higher in the EDA dataset. Incorporating EDA data with BVP data does not improve performance compared to using only BVP data, indicating that in this study, EDA data is not as informative as BVP data for aligning the sleep quality score with physiological signals.

BVP data collected during sleep time may be informative due to the lack of noise and movement that are present during waking hours. In contrast, activity in the EDA signal during sleep time may reflect changes in body temperature and the resulting sweat if the wrist is heated under blankets. EDA signal may instead be more informative during waking hours, when the environment is not as warm and controlled, but this will require further study to confirm.

In future work, in addition to studying the time that the patient was known to be asleep, data could also be processed from the individual's sleep onset latency. This is the period of time between when the individual intends to go to sleep (button press) and when sleep is actually obtained (determined by accelerometer data), as greater sleep onset latency could lead to the individual reporting poorer sleep quality.

CHAPTER III

A Multimodal Approach to Predict Adverse Events After Surgery

3.1 Introduction

Monitoring devices in the ICU produce a wealth of data, but the vast amount of information produced, in addition to false alarms, alarm fatigue, and cognitive biases, can negatively affect patient care [3, 30, 3 1]. As such, there is a growing need for predictive models and effective CDSSs that can identify only the relevant data from multiple sources to give pertinent information to healthcare providers.

Previous work has incorporated EHR data into ML to create predictive models, as EHR data are both noninvasive and routinely collected. Some examples of previous studies include predicting coronary artery disease [78], computing survival risk scores [50], and predicting mortality or 30-day readmission [2, 25, 67].

Other research focuses on the utility of physiological signals. Several models [32, 63, 43] used features from HRV to predict mortality or vascular events. Belle et al. [6] showed that HRV and ECG features could better predict hemodynamic stability compared to standard biomarkers used in hemodynamic assessment. Others [26, 42, 40] took a multimodal approach, using ECG, HRV, Arterial Blood Pressure (ABP) waveform from an arterial line, Pulse Plethysmography (PPG) waveform from a pulse oximeter, and EHR data to predict hemodynamic decompensation. This study builds from Hernandez et al.'s work in [26], focusing on hemodynamic decompensation events post-surgery. This study focused on post-surgery patients in the ICU. Patients who undergo cardiovascular surgery are at risk of hemodynamic decompensation [48], which can include arrhythmia (e.g., atrial fibrillation), hypotension, or pulmonary edema. To build from the previous work, we extended this prediction model to other post-surgical cohorts of increasing heterogeneity. We used multimodal features from physiological signals and EHR data to create ML models that predict hemo-

dynamic decompensation after surgery in surgical ICU patients. This study was originally published in *Scientific Reports* [33].

3.2 Methods

3.2.1 Dataset

The dataset included features extracted from TS estimates of ECG, ABP, PPG, and HRV, Dual-Tree Complex Wavelet Packet Transform (DTCWPT) features extracted from the TS estimation of ECG waveform, and features extracted from EHR data from three postoperative cohorts of surgical patients. The TS and DTCWPT methods of feature extraction are described in further detail in the Signal Processing for Feature Extraction section, and the list of EHR data features is included in Section 3.2.3.

3.2.1.1 Patient Cohorts

All retrospective patient data was obtained from Michign Medicine. The three cohorts of surgical patients had increasing heterogeneity. Cohort 1 consisted of patients recovering from elective cardiac surgery, Cohort 2 major vascular surgeries, and Cohort 3 acute conditions that required urgent and/or major non-cardiac surgery. Examples of surgeries in Cohort 1 included coronary artery bypass grafting, cardiac valve repair/replacement, and thoracic aortic procedures; abdominal aortic aneurysm surgeries (both open repairs and endovascular stenting) and major vascular bypass procedures (e.g., aortofemoral bypass, axillofemoral bypass, and femoral-popliteal artery bypass) for Cohort 2; major abdominal surgeries (e.g., exploratory laparotomies), orthopedic surgeries (e.g., total hip replacements for hip fractures), and neurosurgeries (e.g., craniotomies and spinal fusions) for Cohort 3.

3.2.1.2 Adverse Events

Seven adverse events associated with hemodynamic decompensation were included in this study: low cardiac index, sustained low mean arterial pressure, epinephrine bolus, inotropic therapy initiated, inotropic therapy escalated by $\geq$100%, vasopressor therapy initiated, and vasopressor therapy escalated by $\geq$100%. We also included two additional adverse events that may not necessarily result from hemodynamic decompensation, but are clinically significant enough to warrant detection by an algorithm: re-intubation and mortality, as an algorithm missing these two events while detecting other decompensation events would be clinically less meaningful. The exact definitions, details, and rationale for inclusion of these events as agreed by our clinical team are available in Hernandez et al. [26].

3.2.2 Signal Processing for Feature Extraction

We divided 15-minute ECG, ABP, and PPG signals into five non-overlapping tumbling windows three minutes in length. We created four prediction windows of different lengths: 30 minutes, 1 hour, 2 hours, and 4 hours.

3.2.2.1 Data Preprocessing

To remove artifacts, each ECG tumbling window was preprocessed using a second order Butterworth bandpass filter with cutoff frequencies of 0.5 Hz and 40 Hz. Similarly, each ABP window was preprocessed with a third order Butterworth bandpass filter with cutoff frequencies 1.25 Hz and 25 Hz, and each PPG window with a third order Butterworth bandpass filter with cutoff frequencies 1.75 Hz and 10 Hz.

3.2.2.2 Heart Rate Variability

We identified peaks within the filtered ECG signal using the peak detection method defined in Hernandez et al.[26]. We computed the difference in time between subsequent peaks to produce HRV.

3.2.2.3 Taut String

Previous work has used TS to capture hemodynamic instability in ECG [26, 6], and we use these same TS features in our method with varying values for ϵ. The definition of TS is provided in Section 1.2.2.

We applied TS to each tumbling window of ECG, HRV, ABP, and PPG signals with different ϵ values. A summary of TS estimations performed and the number of features extracted are listed in Tables 3.1 and 3.2. The features extracted from PPG and ABP were: number of peaks, the minimum, maximum, mean, median, and SD of time between sequential systolic peaks, time between a systolic peak and its subsequent diastolic reading, relative amplitude between systolic peaks, and relative amplitude between a systolic peak and its subsequent diastolic reading. The features extracted from ECG and HRV were: number of line segments, number of inflection segments, total variation of noise, total variation of denoised signal, power of noise, and power of denoised signal.

3.2.2.4 Dual-Tree Complex Wavelet Packet Transform

DTCWPT [5] has previously been used in an ECG context [26]. More detail is available in Bayram et al. [5], but briefly: at each level k, DTCWPT uses a high- and low-pass perfect

Table 3.1: Epsilon Values for Each Feature Type

Feature Type	ϵ Values
ECG	0.0100, 0.1575, 0.3050, 0.4525, 0.6000
HRV	0.0010, 0.0258, 0.0505, 0.0753, 0.1000
DTCWPT	0.0100, 0.1575, 0.3050, 0.4525, 0.6000
ABP	0.1000, 0.7000, 1.3000, 1.9000, 2.5000
PPG	1.0000, 8.7500, 16.5000, 24.2500, 32.0000

Table 3.2: Features Extracted per Epsilon Value

Feature Type	Number of Features Extracted
ECG	6
HRV	6
DTCWPT	152
ABP	21
PPG	21

reconstruction wavelet filter bank to decompose the previous level's subbands. Increasing k yields increased frequency resolution, but at computational expense. We select $k = 2$ in this study.

The filter banks of k are selected such that the first filter bank's discrete Hilbert transform is the frequency response of each branch of the second filter bank. This allows for approximate shift-invariance. If Ψ is the wavelet for low-pass filter $h_0(n)$ and high-pass filter $h_1(n)$, and Ψ' or $\mathcal{H}\{\Psi\}$ is its Hilbert pair, then the z-transforms of the two filters, H_0 and H_1, are related by

$$H_1(e^{jw}) = -e^{jdw} H_0^*(e^{j(w-d)})$$

when the wavelet basis is orthonormal. H_1 and H_1' have the relationship

$$H_1' = -j \cdot \text{signum}(w) e^{j0.5w} H_1(e^{jw}),$$

Table 3.3: Tensors Formed for Each Feature Type

Feature Type	Tensor Dimensions
ECG	$5 \times 6 \times 5$
HRV	$5 \times 6 \times 5$
DTCWPT	$5 \times 152 \times 5$
ABP	$5 \times 21 \times 5$
PPG	$5 \times 21 \times 5$

where d is an odd integer and $\|w\| < \pi$.

Following this, if $H^{(k)}(e^{jw})$ is the equivalent response at level k, then:

$$H^{(k)}(e^{jw}) = H_1(e^{j2(k-1)w}) \prod_{m=0}^{k-2} H_0(e^{j2mw})$$

and the equivalent response of the second filter bank's corresponding branch is

$$H'(k)(e^{jw}) = -e^{j0.5w}\mathcal{H}\{H^{(k)}(e^{jw})\}$$

according to [5].

3.2.3 Electronic Health Records

EHR data was available for each subject in addition to physiological signals. It included static information, including age, race, and comorbidities, as well as temporal information, such as lab results and medications administered. The features that required different levels of representation through one-hot encoding are presented in Tables 3.4 and 3.5 and also further explained in Hernandez et al.[26]. For completeness, the temporal EHR information for each tumbling window was carried over from the most recent record. Lab values were represented as the encoded values "Low", "Normal", "High", or "Critical", defined in the reference range in Table 3.4. Cardiovascular Infusions were encoded as "None Given", "Normal Dosage", and "Elevated Dosage". Thresholds for "Normal" and "Elevated" are given in Table 3.5.

3.2.4 Tensor Formation and Reduction

Taking into account all values of ϵ from TS estimation, the filter banks of DTCWPT, and the five tumbling windows, a total of 5,150 features were extracted from ECG, HRV, ABP, and PPG signals. To reduce the feature space of this data and maintain structural information, we relied on tensor decomposition.

First, we formatted the data into a tensor. As seen in Table 3.3, each type of signal features had the structure of five tumbling windows and five values of ϵ, with varying numbers of features. We standardized the features within each tumbling window and ϵ value using mean and SD of each ϵ-window-feature entry from the training set. These training set values were later used to standardize the ϵ-window-feature entries from the test set.

Once the tensors of a feature type have been created for all subjects in the training set, they were stacked along a new fourth mode, generating a tensor of size ($\epsilon \times$ feature $\times$ window $\times N_{\text{train}}$), where N_{train} was the number of individuals in the training set. This

Lab Value	Reference Range	Critical	Unit
Creatinine	Female: 0.5-1.0 Male: 0.7-1.3	> 2.0	mg/dL
Glucose	70-180	< 40	mg/dL
Hematocrit	Female: 36-48 Male: 40-50	< 21	%
Hemoglobin	Female: 12-16 Male: 13.5-17	< 7	g/dl
International Normalized Ratio	0.9-1.2	> 2.0	
Lactate	Arterial: 0.5-1.6 Venous: 0.5-2.2	> 4.0	mmol/L
Platelet Count	150-400	< 50	10^9/L
Potassium	3.5-5.0	> 6.0	mmol/L
Sodium	136-146	> 155	mmol/L
White Blood Cell Count	4-10	> 20	10^9/L

Table 3.4: Lab Values and their Reference Ranges

Cardiovascular Infusion	Escalation Threshold
Dobutamine	2.0 μg/kg/min
Dopamine	2.5 μg/kg/min
Epinephrine	0.02 μg/kg/min
Isoproterenol	2.0 μg/kg/min
Milrinone	0.25 μg/kg/min
Norepinephrine	0.1 μg/kg/min
Vasopressin	2.0 μg/kg/min

Table 3.5: Cardiovascular Infusions and their Reference Ranges

was repeated for tensors of each feature type. We performed Higher Order Singular Value Decomposition (HOSVD) [74, 75, 16] using *Tensor Toolbox* [4] to reduce the DTCWPT, ABP, and PPG tensors to their core tensors, $G_{\text{DTCWPT}}, G_{\text{ABP}}, G_{\text{PPG}}$, and reserved their respective transformation matrices ($U_{\text{DTCWPT}}, U_{\text{ABP}}, U_{\text{PPG}}$).

Next, we stacked all fourth order tensors from the different feature types along the feature mode (second mode) to create a new tensor T. We then performed CP-ALS to reduce the feature space, defined in Section 1.2.3. We set rank $r = 4$ as in Hernandez et al.[26]. After solving for $\hat{T}$, we reserved factor matrices A and C for feature extraction.

Now having the transformation matrices from HOSVD and the factor matrices from CP-ALS, we could compute the core tensor of any DTCWPT, ABP, or PPG tensor in the test

set. We perform the following to compute features for each individual j in the test set: Using the transformation matrices from HOSVD ($U_{\text{DTCWPT}}, U_{\text{ABP}}, U_{\text{PPG}}$), core tensors were computed for DTCWPT, ABP, and PPG. Next, all of j's tensors were stacked along mode 2 to build the new tensor S_j. Lastly, the factor matrices A and C from T's CP decomposition were used to solve for B via least squares by minimizing

$$\|S_j - B(C \odot A)^\intercal\|.$$

The optimal solution was calculated as

$$B = S_{(2)}(C \odot A)(A^\intercal A * C^\intercal C)^\dagger,$$

where $S_{(2)}$ was the mode-2 slice of tensor S_j, $\odot$ was the Khatri-Rao product, and $\dagger$ was the Moore-Penrose pseudoinverse [35].

3.2.5 Machine Learning

The ML models were created to predict an occurrence of the nine adverse events detailed in Section 3.2.1.2 in prediction windows 0.5, 1, 2, and 4 hours before the event. The training set consisted of Cohort 1 while the test set was either Cohort 2 or 3 exclusively. Given that clinicians may want to detect more cases with adverse outcomes at the risk of somewhat increased false positives, we balanced the training set to have a ratio of 0.35 to 0.65 between positive and negative cases by undersampling negative cases. No such adjustment was performed for the test sets. Patients without complete signals in the analysis windows (e.g., ECG signals with missing R peaks or periods of no signal) were excluded. The respective sample size for each prediction window is available in Tables 3.6 - 3.8.

Model training consisted of 3-fold cross-validation (CV) of Cohort 1 data to select optimal hyperparameters for each model using a validation set. Models were then trained on all three folds with the selected hyperparameters in Cohort 1, and tested against Cohort 2 and Cohort 3 data. This process was repeated 101 times, with shuffling of data across CV folds, for each model to obtain the mean AUROC.

Naive Bayes (NB) models with normal distribution were trained with no hyperparameter tuning, serving as the simplest baseline models.

Random Forest (RF) models [9] were trained with varying numbers of trees (50, 75, or 100), minimum leaf size (1, 5, 10, 15, or 20), percentage of features to include for maximum number of splits (25%, 50%, 75%, 100%), split criterion (Gini impurity or cross entropy), and number of predictors to sample (10-100 in increments of 10) using grid search to select

hyperparameters.

SVMs [14] were trained with a linear kernel via sequential minimal optimization. Grid search was used to determine the optimal box constraint C and scaling parameter γ, where $C \in [10^{-7}, 10^{12}]$ and $\gamma \in [10^{-12}, 10^{12}]$ consisted of logarithmically-spaced values.

LUCCK, defined in 2.2.3, was trained with λ values in a range from 0.01-0.1 in increments of 0.01, and θ values in a range from 0.1-1.0 in increments of 0.1. We selected for λ and θ via grid search.

3.3 Results

The AUROCs of each ML model are presented in Tables 3.6 - 3.8. The tables present means and SD of AUROC for models trained on the Cohort 1 (Cardiac Surgery) and tested on Cohorts 2 (Vascular Surgery) and 3 (Acute Non-cardiac Surgery). The n in each row of Tables 3.6 through 3.8 represents the number of deterioration events (i.e., the sample size). The best performance in each prediction window is boldfaced.

3.3.1 Cohort 1 - Training with Cardiac Surgery Cohort

During training with Cohort 1, the RF models exhibited the highest AUROCs between 0.93 and 0.94 across all gaps, followed by SVM (0.90-0.91) and LUCCK (0.89-0.91). The NB models achieved the lowest performance (0.83-0.85).

3.3.2 Cohort 2 - Testing with Vascular Surgery Cohort

The RF models achieved the highest AUROCs on testing with Cohort 2. The LUCCK models performed marginally better on Cohort 2 compared to the training set, achieving AUROCs similar to those of RF. The NB and SVM models also achieved AUROCs comparable with those obtained on the training set, with the NB models on 2- and 4-hour windows actually achieving higher AUROCs than their respective training sets. The NB and SVM models demonstrated higher SDs compared to the LUCCK and RF models.

3.3.3 Cohort 3 - Testing with Acute Non-cardiac Surgery Cohort

With Cohort 3, the overall performance was lower across all models. The RF models achieved the highest performance, with the models on 0.5- and 1-hour windows achieving AUROCs greater than 0.8. LUCCK maintained mean AUROCs between 0.74 and 0.77 across models on all windows, while the SVM's AUROCs fluctuated to a larger extent, between 0.67 and

Table 3.6: Mean AUROC and SD of Cohort 1 - Cardiac Surgery Cohort (Training Set)

Prediction Window (hrs)	LUCCK	RF	NB	SVM
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
0.5 ($n = 423$)	0.90 (0.01)	**0.94** (0.01)	0.83 (0.02)	0.91 (0.01)
1 ($n = 466$)	0.89 (0.01)	**0.93** (0.01)	0.84 (0.02)	0.91 (0.01)
2 ($n = 426$)	0.89 (0.01)	**0.93** (0.01)	0.83 (0.02)	0.90 (0.01)
4 ($n = 414$)	0.91 (0.01)	**0.93** (0.01)	0.85 (0.02)	0.91 (0.01)

Table 3.7: Mean AUROC and SD of Cohort 2 - Vascular Surgery Cohort (Test Set A)

Prediction Window (hrs)	LUCCK	RF	NB	SVM
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
0.5 ($n = 66$)	**0.94** (0.02)	**0.94** (0.01)	0.82 (0.06)	0.91 (0.08)
1 ($n = 60$)	**0.94** (0.02)	**0.94** (0.01)	0.83 (0.07)	0.91 (0.06)
2 ($n = 64$)	0.92 (0.01)	**0.94** (0.02)	0.88 (0.04)	0.92 (0.04)
4 ($n = 63$)	0.90 (0.02)	**0.92** (0.02)	0.87 (0.03)	0.90 (0.06)

Table 3.8: Mean AUROC and SD of Cohort 3 - Acute Non-Cardiac Surgery Cohort (Test Set B)

Prediction Window (hrs)	LUCCK	RF	NB	SVM
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
0.5 ($n = 21$)	0.75 (0.13)	**0.82** (0.15)	0.42 (0.03)	0.80 (0.15)
1 ($n = 20$)	0.75 (0.12)	**0.82** (0.10)	0.51 (0.09)	0.67 (0.16)
2 ($n = 16$)	0.74 (0.08)	0.72 (0.16)	0.79 (0.07)	**0.81** (0.16)
4 ($n = 12$)	**0.77** (0.09)	**0.77** (0.15)	0.63 (0.10)	0.74 (0.14)

0.81. The NB models declined to the greatest extent, exhibiting an AUROC as low as 0.42 on the 0.5-hour window.

3.4 Discussion

In this study of postoperative deterioration events among three successively heterogeneous cohorts of surgical patients, the incidence rates of a patient-level deterioration event were 16-19%, 19-28%, and 75-90%, respectively, although for Cohort 1, we adjusted the rate of positive cases to 35% during training. The best performing models for each cohort yielded AUROCs of 0.94, 0.94 and 0.82, respectively, all for 0.5-hour prediction windows. Our study serves as a proof-of-concept that EHR data and physiologic waveform data may be combined to improve the early detection of postoperative deterioration events, even when trained on one surgical cohort and applied to other cohorts.

The predictive performance is well maintained throughout all prediction windows during training with Cohort 1; the RF models achieved the highest performance. With Cohort 2, likely due to smaller sample size, the SDs of AUROCs of the models tended to be higher than those of the training set. Nonetheless, these models achieved performance comparable to that achieved on the training set, with some models (especially LUCCK) even exceeding the mean AUROCs achieved in training. Overall, the models trained on Cohort 1 generalized extremely well to the Cohort 2.

When the models were tested on Cohort 3, the acute non-cardiac surgery cohort, the performance of all models declined. This is likely due to the paucity of samples in Cohort 3; the SDs of the AUROCs are also much larger than their counterparts in Cohorts 1 and 2. The NB models suffered most from the performance decline, with models on 0.5- and 1-hour prediction windows performing near or worse than random guessing. Given the simplicity of the NB models, it is not surprising that they fail to generalize to a small dataset from a different surgical cohort.

On the other hand, the other three models maintained an acceptable performance for most prediction windows despite the paucity of data. The RF models achieved the highest AUROC with the 0.5- and 1-hour prediction windows at 0.82. The AUROC declined for the 2-hour prediction window to 0.72 but recovered to 0.77 at the 4-hour prediction window. The SVM model displayed even more fluctuation, with the highest AUROC of 0.81 on the 2-hour prediction window and lowest AUROC of 0.67 on the 1-hour prediction window. The LUCCK models maintained the most consistent range of performance, between 0.74 and 0.77 across all prediction windows, with generally smaller SDs than RF and SVM. LUCCK also did not experience a steep decline that the RF and SVM models showed in

the 1- and 2-hour prediction windows, respectively. It should also be noted that LUCCK's generalization with Cohort 2 was the best among the four models, showing the largest increase in performance from training when tested on Cohort 2. Such findings are consistent with previous reports indicating that the LUCCK models are capable of demonstrating more robust and generalizable performance, especially when the number of samples is small [60, 26], which is often the case with clinical data.

A few predictive scoring systems for assessing mortality risks in the ICU patients have been developed in clinical settings. The APACHE II scoring system, which utilizes 12 physiological measurements, age, and previous health conditions from 5,030 ICU patients in 13 hospitals to create a logistic regression model to predict hospital death, reported an AUROC of 0.863 [34].

The SAPS II scoring system, which utilizes 12 physiological variables, age, type of admission, and three underlying disease variables to create a logistic regression model predicting the hospital mortality based on 12,997 ICU patients from 12 countries, reported an AUROC of 0.88 on the developmental dataset and 0.86 on the validation dataset [38].

The evaluation of Sequential Organ Failure Assessment (SOFA) scoring system, [21] which utilizes physiological variables from various organ systems to assign a score between 0-24, demonstrated the highest AUROC of 0.90 based on 352 patients when using the highest SOFA score taken during the entire ICU stay.

Our study differs from these studies in two major aspects: 1) We have included features derived from several digital signal processing techniques performed on the physiological waveform data, and 2) our models are trained to predict other clinical deterioration events besides mortality, given that the ultimate form of these models is intended for real-time clinical monitoring in a CDSS. Despite being trained on just a few hundred cases, our best-performing models can achieve AUROCs between 0.90 and 0.94 for all prediction windows on Cohort 2. Although Cohort 3's best performing models achieve a maximum AUROC of 0.82, we expect that such performance can be improved upon including additional quality physiological waveform data, given the largest sample size in this Cohort was 21.

Monitoring patients in the ICU for cardiovascular adverse events and complications is a crucial component of postoperative critical care. Early warning systems for patients predicted to be at risk of major cardiovascular complications in advance of clinicians' attention can enable potentially life-saving interventions to be pursued before deterioration onset and significantly improve clinical outcomes. However, because different surgical procedures imply different underlying patient pathologies, along with varying levels of invasiveness and resultant derangements to organ systems, a CDSS for predicting postoperative adverse outcomes may be challenged by inadequate generalizability across a variety of surgical populations. In

this study, we test the hypothesis that models trained on one surgical cohort may be successfully applied to other cohorts that underwent different surgical procedures, by training them on the cardiac surgery cohort and testing them on vascular and acute non-cardiac surgery cohorts. The results obtained in this work suggest such generalizability and serve as a prototype for a CDSS where the ML models trained in one surgical cohort can be successfully applied to others.

Additionally, one should consider ensembles of different ML models in implementation of a CDSS, since no model definitively stands out as the best model in all cases. For example, in Cohort 3, RF has the highest performance of 0.82 in the 0.5- and 1-hour prediction windows, SVM of 0.81 in the 2-hour window, and LUCCK of 0.77 (which is equal to that of RF but with a smaller SD) in the 4-hour window.

As for limitations, sufficient data for training were only available for Cohort 1, allowing us only to train the models on Cohort 1 and test on Cohorts 2 and 3, but not vice versa. Adjusting the positive/negative case ratio to increase sensitivity may result in more false positives. Also, all cases in this study were from a single quaternary care center. The care processes and patient populations at other facilities can be substantially different. Another limitation is the few negative cases in Cohort 3, resulting in very high incidence rates of adverse events that are unlikely to be encountered in actual clinical settings. In future work, more non-cardiac surgical cases from other surgical care facilities would be needed to further increase the generalizability and robustness of these models. Also, our results must be interpreted with caution given that they were based on all surgical cohorts and may not necessarily extrapolate to non-surgical cohorts.

3.5 Conclusion

In this study, we utilized digital signal processing techniques such as TS estimation and DTCWPT to generate features from physiological waveforms. A novel tensor dimension reduction algorithm successfully reduced the feature space while still demonstrating translatable performance across different surgical cohorts. Four different types of ML models were trained on a cardiac surgery cohort and tested on different surgery cohorts. The RF models perform the best in terms of the AUROC, but LUCCK models maintain the most consistent range of performance, especially when the sample sizes are small. Our study suggests that ML models trained on a combination of waveform and EHR data of one group of surgical cases to predict life-threatening cardiovascular complications have potential to be successfully applied to other types of surgical cases, opening doors to the clinical decision support system enabling early detection of such events and timely interventions to improve clinical

outcomes on a wide variety of surgical cases.

CHAPTER IV

A Multimodal Approach to Predict Trajectory of Sepsis in the ICU

4.1 Introduction

Sepsis is a syndrome induced by an existing infection in the body that produces life-threatening organ dysfunction in a chain reaction. The clinical criteria for sepsis include suspected or documented infection and an increase in two or more Sequential Organ Failure Assessment (SOFA) points. Septic shock, a more severe subset, consists of substantially increased abnormalities [65] and higher risk of mortality [51]. It is imperative to risk-stratify patients early during treatment in order to appropriately direct critical, but potentially limited, resources and therapies.

Sepsis' heterogeneity complicates its diagnosis and prognosis. Its current definition, based on SOFA score, requires measurement or collection of variables which may not be immediately available. The Quick Sequential Organ Failure Assessment (qSOFA) is a screening tool that can be performed at the bedside. It consists of three criteria - Glasgow Coma Scale (GCS) of < 15 (indicating mental status change), respiratory rate ≥ 22 breaths per minute, and systolic Blood Pressure (BP) ≤ 100 mmHg - where two of the three must be met [65]. It includes the poorly characterized variable mental status change, but it is a better predictor of organ dysfunction than systemic inflammatory response syndrome (SIRS), which is less sensitive [8, 64]. SIRS is the body's response to a stressor such as inflammation, trauma, surgery, or infection, while sepsis is specifically a response to infection; many septic patients have SIRS, but not all patients who meet SIRS criteria have an infection or experience septic organ failure. In comparison to qSOFA, SIRS has four criteria, three of which must be met to positively identify SIRS. These are: respiratory rate > 20 breaths per minute or partial pressure of $CO_2 < 32$ mmHg; heart rate > 90 beats per minute; white blood cell count $> 12{,}000$/microliter or $< 4{,}000$/microliter or bands $> 10\%$; and temperature $>38°$C

or $< 36°$C [12]. For each of these scoring systems, factors such as comorbidities, medication, and age may confound the phenotype in different patient groups.

A sepsis detection system that is too strict or time-consuming can delay necessary care to patients, and criteria that are too broad can lead to over-treatment or inappropriate use of limited resources. For example, false positive sepsis prognoses can lead to patients receiving unnecessary care and antibiotics, which contribute to antibiotic resistance and emergence of "superbugs" [76, 54, 13]. Similarly, qSOFA is not recommended as a single screening tool for diagnosis of sepsis [18], but it can be used as a method of predicting prolonged ICU stay or in-hospital mortality [64]. Predicting the trajectory of a patient with suspected infection may be a more efficient use of resources than detecting existing sepsis, and therefore trajectory prediction is the focus of this study.

Many models for detecting, monitoring, or predicting outcomes related to sepsis depend on EHR data, including SOFA score [65], EPIC's sepsis model [79], and others [47, 45, 70]. While useful for determining a patient's status, EHR data are limited by time. Lab values require time for collection and processing, and continuous variables may be updated less than hourly or at irregular intervals. In contrast, physiological readings, such as those generated from electrocardiography, blood pressure monitoring, or pulse oximetry, are collected continuously and at regular intervals. Our study examines the use of continuous physiological signals, namely ECG and ABP, in outcome prediction related to sepsis.

Previous works have used ECG signal information in the study of risk for sepsis and sepsis progression [7, 44, 46]. The advantage that continuous monitoring devices like ECG offer over EHR data is real-time, continuous assessment of a patient's status. In addition, ECG is routinely collected in the ICU, and is minimally invasive. In our analysis, we also include ABP, as both SOFA and qSOFA use BP to assess the status of a patient's cardiovascular system status [65].

Given the complexity and heterogeneity of sepsis, it is necessary to incorporate multiple variables into a trajectory prediction method. Modeling data as a tensor provides the ability to observe changes in different variables with respect to time and to one another. The prognosis and severity assessment of sepsis rely on a large amount of heterogeneous data, including body temperature, arterial blood pressure, blood culture tests, and molecular assays. Treatment of sepsis does not rely on any individual variable, but on all of these measurements, which vary as a function of time. Because no individual feature is sufficient, integrating data across time and incorporating structure is necessary for improved sepsis prognosis, and therefore can better inform care decisions.

In this study, we use ECG and ABP signals to predict an increase in an individual's qSOFA score, where a qSOFA of ≥ 2 indicates poor outcomes related to sepsis. The results

of signal-trained models are then compared to models trained using both signals and EHR data. This is to (1) predict which individuals are at risk to decompensate to septic shock, experience future organ failure, or other complications related to sepsis, rather than focusing on a sepsis diagnosis, and (2) assess the usefulness of continuous physiological signals in the event that EHR data are unavailable.

What differentiates this study from previous work is three-fold: first, that the models developed in this work are focused on risk to decompensate rather than on a diagnosis of sepsis; second, that physiological signals such as ECG and ABP are continuous and therefore can provide near real-time views of a patient's status; third, that the signals are further processed with tensor methods. The results of this research have been submitted to *Scientific Reports*.

4.2 Methods

A schematic of the methods used in this paper is presented in Fig.4.1.

4.2.1 Dataset

The retrospective dataset consisted of 1,803 unique individuals age $\geq$ 18 years with 3,516 unique encounters between 2013-2018 at Michigan Medicine. Individuals' characteristics are presented in Appendix Section B. The inclusion criteria selected for inpatient encounters with: ECG lead II waveforms at least 15 minutes in length and ICD 9/10 codes for pneumonia, cellulitis, or Urinary Tract Infection (UTI), excluding UTIs associated with catheters. Exclusion criteria included positive HIV status, solid organ or bone marrow transplant, and ongoing chemotherapy. These criteria created a dataset that did not specifically select for sepsis diagnosis, but instead focused on patients with an infection who were at risk to develop sepsis and septic shock. This dataset was selected from a Michigan Medicine biobank, whose data collection was approved by the institutional review board of University of Michigan. Informed consent was waived, as this was a retrospective study of previously collected and de-identified data, without direct involvement of human subjects and therefore no chance of physical harm or discomfort to the individuals being studied.

This larger dataset was reduced by selecting for individuals who had EHR, ECG, and ABP data available. Because poor signal quality can result in false alarms [22], the ECG signal was reviewed automatically using Pan-Tompkins to identify QRS complexes [49, 62]. Upon collecting 10-minute signals for feature extraction, signals determined to be 50% or more noise were discarded.

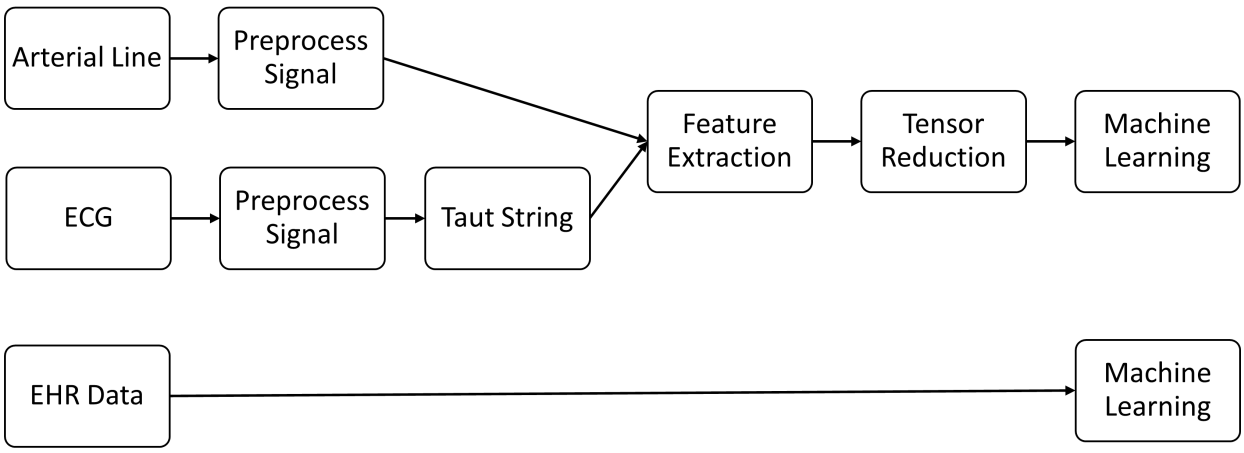


Figure 4.1: Schematic

Change in qSOFA score was used to assign positive and negative classes for machine learning. Given an individual who meets one of the criteria for qSOFA, the model predicts whether the score will increase to ≥ 2, which Sepsis-3 deems as "likely to have poor outcomes" [65]. This increase in qSOFA is considered the positive outcome in a learning context, because the patient meets at least 2 qSOFA criteria as defined by Sepsis-3 after the prediction gap. Thus, the negative outcome is qSOFA < 2 after the prediction gap.

We tested prediction gaps of six and twelve hours. For a six-hour gap, there were 199 negative and 59 positive cases. For a twelve-hour gap, there were 189 negative and 37 positive cases.

4.2.2 Signal Processing

For every sample, we collected the 10 minutes of signal occurring directly before the prediction gap for processing. This 10-minute signal was divided into 2 5-minute windows, and

then preprocessed according the relevant sections below.

4.2.2.1 Arterial Line Data

ABP signals were sampled at 120 Hz. We applied a third order Butterworth bandpass filter with cutoff frequencies 1.25 and 25 Hz to remove artifacts. The *BP_Annotate* software package [37] annotated the signal. Following previous methodology, [41, 26, 33], we extracted the same 21 features from the annotated signal as in Section 3.2.2.3: number of peaks, as well as the minimum, maximum, mean, median, and standard deviation (SD) of time between sequential systolic peaks, time between a systolic peak and its subsequent diastolic reading, relative amplitude between systolic peaks, and relative amplitude between a systolic peak and its subsequent diastolic reading.

4.2.2.2 Electrocardiogram Data

ECG data consisted of four leads and signals were sampled at 240 Hz. We used lead II of the ECG, following previously established methods [6]. A second order Butterworth bandpass filter with the cutoff frequencies 0.5 and 40 Hz removed noise and artifacts.

4.2.2.3 Taut String

We calculated peak-based and statistical features from the TS estimation [15] of the ECG waveform. Such features have previously been used to detect hemodynamic instability [6] and predict hemodynamic decompensation [26, 33]. More detail on the TS algorithm is provided in Section 1.2.2.

As in Chapter 3.2, TS estimation was applied to the filtered ECG signal using the previously selected five values of ϵ, and the same six features were computed from each TS estimate of a 5-minute window and value of ϵ, leading to the creation of a tensor of size $2 \times 5 \times 6$ for each signal, where the modes of the tensor were window, ϵ, feature.

4.2.3 Electronic Health Record Data

Rather than use a one-hot encoding as in Section 3.2.3, we assigned an ordinal encoding to labs and cardiovascular infusions ranging from 0-4 or 0-3, respectively. A score of 1 indicates less severity and a score of 3 or 4, more severity. If a lab value had been recorded before the time of interest, this value was carried forward. We assigned a score of 0 to represent a missing value with no previous recordings. Tables 3.4 and 3.5 provide the reference ranges for lab values and cardiovascular infusions, respectively. Vital signs and urine output were

included, but not given an ordinal encoding. If vital signs or urine output were not reported in the time of interest, we carried forward the most recent known value.

We added a retrospective component for lab values, cardiovascular infusions, and vital signs where, in addition to the 10 minutes occurring before the prediction gap, we include four look-back periods. For the prediction gap of 6 hours, these look-back periods were increments of 4 hours; for the prediction gap of 12 hours, they are increments of 8 hours.

4.2.4 Feature Reduction with Tensor Methods

For each 10-minute ECG signal, 60 features were computed and arranged as a tensor of size $2 \times 5 \times 6$. For each 10-minute ABP signal, 42 features were arranged as a tensor of size $2 \times 1 \times 21$, where the second mode, TS parameter ϵ, was inflated to create a uniform presentation to the tensor reduction algorithms. Rather than treating this information as 60 or 42 feature vectors, we preserved the underlying tensor structure by using a tensor-based dimensionality reduction method, inspired by previous work [26, 33] and described in further detail in Section 1.2.3.

First, each tensor's underlying structure was determined. All $2 \times 5 \times 6$ ECG-feature tensors in the training set were stacked along the fourth mode, generating a new tensor of size $2 \times 5 \times 6 \times N$, where N was the number of observations in the training set. Similarly, all $2 \times 1 \times 21$ ABP-feature tensors were stacked along the fourth mode to generate a new tensor of size $2 \times 1 \times 21 \times N$. *Tensor Toolbox's* [4] CP-ALS [35] was used to obtain the underlying structure of the tensors.

The dataset was divided into a 75/25 split 100 times, and tensor reduction was performed on each of those splits. CP-ALS was run using rank values of 1-4, and the fitting process was repeated 15 times.

After applying CP-ALS to the training data, the resulting factor matrices A and B were retained, which related to the modes of the original tensor that were not the feature mode (C) or the patient encounter mode (D).

With this process completed, for any given individual's third-order tensor T, a reduced set of features was extracted using the factor matrices computed from the training data. The feature vectors $c_{T,1}, \ldots, c_{T,r}$ were computed via a least squares problem, where

$$\|T - \sum_{i=1}^{r} a_i \otimes b_i \otimes c_{T,i}\|$$

is minimal. After computing the individual vectors, they were concatenated to create C_T, a feature matrix with a reduced set of features compared to matricization $T_{(3)}$ of the original

tensor T along the third mode.

4.2.5 Machine Learning

When constructing training and test datasets, 75/25 splits were created based on individuals so that no individual would overlap between the training and test sets.

After extracting features, the three types of learning models used for training were linear SVMs [14], RF [9], and LUCCK [60]. We selected a linear kernel for SVM in this experiment because linear kernels tend to be less susceptible to overfitting when many features are present [24] (such as in the case when no tensor reduction is used), and a linear kernel is both faster to train and more easily interpretable than a nonlinear kernel [29]. Additionally, datasets with many features can become linearly separable, making the linear kernel a good option both in terms of its transparency as well as its faster training time [11]. We opted not to test deep learning models because we wanted to offer transparency to the end user of the model and to patients who would receive care, as deep learning models are known for operating as a "black box"; a patient would trust a clinician who understands the "explainable" machine learning method that they use to assist in their decision-making (referred to as the AI-user dyad) [20].

For all methods, the training phase consisted of Three-Fold Cross-Validation (3FCV) on a 75/25 split of the data, where the test set was held and not used for training. The test set was presented to the three models generated from 3FCV to produce three sets of prediction scores. We computed the final prediction scores for the test set by taking the median of the three prediction scores, thus creating a voting system. This process was repeated 100 times to obtain mean and SD of model performance.

A grid search selected optimal hyperparameters for each model using the validation fold in 3FCV. For RF, these hyperparameters included: number of trees, minimum leaf size, fraction of maximum number of splits, and number of predictors to sample. For SVM, grid search selected the best box constraint C. Sequential minimal optimization [19] was used for the optimization routine. For LUCCK, grid search selected optimal Λ and Θ parameters. All grid searches used F1 score as the value to optimize.

Different signal feature based models were tested using tensor reduction. The first, using only ECG data and presented in Figure 4.2, was the most restricted model, assuming that both EHR and ABP data were unavailable. This would apply to patients recently admitted, who would not have lab values or other EHR data available, and is also minimally invasive compared to having an arterial line in place. Next was a model trained on both ECG and ABP features, presented in Figure 4.3, which was tested to determine if the more invasive

Table 4.1: ECG-Only Models, 6-hour gap

Model	Rank	F1 Score	Recall	Specificity	AUROC
LUCCK	1	0.47 (0.08)	0.63 (0.12)	0.69 (0.10)	0.61 (0.09)
	2	0.48 (0.07)	0.68 (0.11)	0.66 (0.09)	0.64 (0.08)
	3	0.48 (0.07)	0.67 (0.11)	0.67 (0.08)	0.64 (0.08)
	4	0.48 (0.07)	0.69 (0.12)	0.67 (0.09)	0.65 (0.07)
	None	0.43 (0.06)	0.66 (0.13)	0.60 (0.11)	0.60 (0.07)
RF	1	0.46 (0.06)	0.66 (0.12)	0.64 (0.11)	0.64 (0.06)
	2	0.47 (0.05)	0.71 (0.09)	0.63 (0.10)	0.66 (0.06)
	3	0.47 (0.06)	0.69 (0.11)	0.64 (0.09)	0.66 (0.06)
	4	0.48 (0.06)	0.72 (0.10)	0.64 (0.09)	0.67 (0.06)
	None	0.41 (0.06)	0.64 (0.12)	0.57 (0.11)	0.57 (0.08)
SVM	1	0.37 (0.08)	0.50 (0.15)	0.66 (0.15)	0.49 (0.10)
	2	0.38 (0.08)	0.52 (0.14)	0.64 (0.15)	0.50 (0.10)
	3	0.38 (0.08)	0.52 (0.14)	0.66 (0.11)	0.50 (0.10)
	4	0.38 (0.09)	0.53 (0.16)	0.64 (0.12)	0.50 (0.11)
	None	0.44 (0.07)	0.64 (0.12)	0.63 (0.14)	0.62 (0.09)

arteial line improved performance compared to only using ECG data. Lastly, a model trained on signal features alongside EHR data was built, presented in Figure 4.4.

4.3 Results

RF, LUCCK, and SVM were trained on tensor-reduced ECG features, presented in Table 4.2. We compared these models to those trained on tensor-reduced ECG features and ABP features, presented in Table 4.3. These figures display the mean F1 Score and AUROC over 100 iterations, with error bars indicating one SD. The x-axis indicates the rank selected for CP-ALS, with the rightmost columns, separated with a dashed line, representing the case where no tensor decomposition was applied.

Figure 4.4 shows the results of models trained on both the tensor-reduced signal features and EHR data.

4.4 Discussion

RF and LUCCK models performed similarly across different experiments, both performing better than SVM when tensor reduction was applied to the dataset. RF's strong performance across different levels of feature reduction could be due to its bagging and bootstrapping procedures, which work to prevent overfitting and ignore noise [55, 9]. In its introductory paper, LUCCK was shown to perform well even when trained with few samples of signal

Table 4.2: ECG-Only Models, 12-hour gap

Model	Rank	F1 Score	Recall	Specificity	AUROC
LUCCK	1	0.47 (0.10)	0.69 (0.12)	0.74 (0.11)	0.69 (0.10)
	2	0.47 (0.09)	0.69 (0.12)	0.75 (0.11)	0.68 (0.10)
	3	0.49 (0.08)	0.72 (0.12)	0.76 (0.09)	0.71 (0.09)
	4	0.49 (0.09)	0.73 (0.12)	0.75 (0.10)	0.72 (0.09)
	None	0.41 (0.08)	0.67 (0.13)	0.69 (0.10)	0.65 (0.09)
RF	1	0.43 (0.08)	0.69 (0.12)	0.70 (0.11)	0.68 (0.09)
	2	0.44 (0.08)	0.70 (0.12)	0.71 (0.10)	0.69 (0.08)
	3	0.47 (0.08)	0.73 (0.11)	0.72 (0.09)	0.72 (0.08)
	4	0.48 (0.09)	0.73 (0.12)	0.73 (0.09)	0.73 (0.08)
	None	0.38 (0.07)	0.67 (0.12)	0.63 (0.12)	0.62 (0.09)
SVM	1	0.31 (0.10)	0.47 (0.16)	0.68 (0.15)	0.46 (0.12)
	2	0.33 (0.09)	0.52 (0.16)	0.66 (0.16)	0.48 (0.13)
	3	0.32 (0.09)	0.51 (0.16)	0.67 (0.14)	0.48 (0.11)
	4	0.35 (0.10)	0.54 (0.17)	0.70 (0.12)	0.52 (0.13)
	None	0.40 (0.08)	0.70 (0.14)	0.63 (0.13)	0.64 (0.10)

Table 4.3: Models Trained on ECG and Art Line, 6-hour gap

Model	Rank	F1 Score	Recall	Specificity	AUROC
LUCCK	1	0.50 (0.07)	0.74 (0.11)	0.65 (0.12)	0.69 (0.08)
	2	0.52 (0.07)	0.72 (0.10)	0.69 (0.10)	0.71 (0.07)
	3	0.51 (0.06)	0.69 (0.10)	0.70 (0.08)	0.70 (0.06)
	4	0.51 (0.06)	0.72 (0.10)	0.68 (0.09)	0.70 (0.06)
	None	0.50 (0.06)	0.73 (0.11)	0.65 (0.10)	0.69 (0.07)
RF	1	0.49 (0.06)	0.72 (0.12)	0.65 (0.11)	0.68 (0.07)
	2	0.51 (0.06)	0.74 (0.10)	0.66 (0.09)	0.70 (0.06)
	3	0.51 (0.06)	0.73 (0.11)	0.67 (0.08)	0.70 (0.05)
	4	0.52 (0.06)	0.75 (0.10)	0.67 (0.09)	0.71 (0.07)
	None	0.49 (0.06)	0.71 (0.10)	0.66 (0.09)	0.69 (0.07)
SVM	1	0.42 (0.08)	0.61 (0.15)	0.63 (0.13)	0.58 (0.09)
	2	0.41 (0.07)	0.61 (0.14)	0.61 (0.12)	0.57 (0.09)
	3	0.43 (0.07)	0.64 (0.14)	0.61 (0.12)	0.58 (0.09)
	4	0.42 (0.08)	0.64 (0.14)	0.60 (0.12)	0.58 (0.10)
	None	0.47 (0.07)	0.70 (0.13)	0.64 (0.12)	0.63 (0.08)

Table 4.4: Models Trained on ECG and Art Line, 12-hour gap

Model	Rank	F1 Score	Recall	Specificity	AUROC
LUCCK	1	0.45 (0.08)	0.73 (0.14)	0.69 (0.14)	0.72 (0.08)
	2	0.45 (0.07)	0.72 (0.13)	0.71 (0.11)	0.72 (0.07)
	3	0.47 (0.07)	0.75 (0.11)	0.71 (0.10)	0.73 (0.07)
	4	0.46 (0.08)	0.73 (0.13)	0.71 (0.11)	0.71 (0.08)
	None	0.44 (0.06)	0.77 (0.12)	0.67 (0.10)	0.72 (0.06)
RF	1	0.42 (0.08)	0.70 (0.13)	0.68 (0.12)	0.68 (0.09)
	2	0.44 (0.07)	0.71 (0.12)	0.69 (0.10)	0.70 (0.08)
	3	0.47 (0.07)	0.74 (0.12)	0.72 (0.10)	0.72 (0.08)
	4	0.48 (0.08)	0.75 (0.10)	0.72 (0.10)	0.74 (0.07)
	None	0.44 (0.08)	0.76 (0.12)	0.67 (0.10)	0.72 (0.07)
SVM	1	0.32 (0.07)	0.57 (0.17)	0.61 (0.15)	0.49 (0.10)
	2	0.32 (0.08)	0.57 (0.18)	0.62 (0.15)	0.51 (0.12)
	3	0.33 (0.08)	0.57 (0.16)	0.61 (0.16)	0.51 (0.11)
	4	0.33 (0.07)	0.54 (0.15)	0.61 (0.15)	0.51 (0.12)
	None	0.36 (0.07)	0.65 (0.16)	0.61 (0.15)	0.56 (0.12)

Table 4.5: Models Trained on ECG, Art Line, and EHR Data, 6-hour gap

Model	Rank	F1 Score	Recall	Specificity	AUROC
LUCCK	1	0.53 (0.06)	0.74 (0.11)	0.70 (0.09)	0.74 (0.07)
	2	0.53 (0.06)	0.74 (0.10)	0.69 (0.09)	0.74 (0.07)
	3	0.53 (0.07)	0.74 (0.10)	0.70 (0.09)	0.73 (0.07)
	4	0.53 (0.07)	0.74 (0.11)	0.70 (0.08)	0.73 (0.07)
	None	0.52 (0.06)	0.73 (0.10)	0.69 (0.08)	0.72 (0.07)
RF	1	0.57 (0.07)	0.76 (0.09)	0.73 (0.09)	0.76 (0.06)
	2	0.57 (0.06)	0.77 (0.08)	0.72 (0.08)	0.77 (0.06)
	3	0.58 (0.07)	0.77 (0.09)	0.74 (0.08)	0.77 (0.06)
	4	0.58 (0.07)	0.78 (0.09)	0.73 (0.09)	0.77 (0.06)
	None	0.53 (0.06)	0.76 (0.09)	0.69 (0.08)	0.74 (0.06)
SVM	1	0.49 (0.07)	0.70 (0.11)	0.67 (0.10)	0.67 (0.08)
	2	0.45 (0.08)	0.66 (0.13)	0.63 (0.12)	0.62 (0.10)
	3	0.44 (0.07)	0.65 (0.12)	0.63 (0.12)	0.60 (0.10)
	4	0.43 (0.07)	0.63 (0.12)	0.63 (0.11)	0.58 (0.09)
	None	0.52 (0.06)	0.74 (0.10)	0.69 (0.08)	0.72 (0.06)

Table 4.6: Models Trained on ECG, Art Line, and EHR Data, 12-hour gap

Model	Rank	F1 Score	Recall	Specificity	AUROC
LUCCK	1	0.57 (0.09)	0.82 (0.09)	0.78 (0.09)	0.82 (0.06)
	2	0.56 (0.09)	0.80 (0.09)	0.79 (0.08)	0.82 (0.06)
	3	0.55 (0.09)	0.81 (0.08)	0.78 (0.08)	0.82 (0.06)
	4	0.56 (0.09)	0.83 (0.09)	0.78 (0.08)	0.83 (0.06)
	None	0.55 (0.08)	0.81 (0.10)	0.78 (0.08)	0.80 (0.06)
RF	1	0.58 (0.10)	0.82 (0.10)	0.79 (0.08)	0.84 (0.05)
	2	0.57 (0.08)	0.80 (0.09)	0.80 (0.07)	0.84 (0.06)
	3	0.56 (0.08)	0.83 (0.10)	0.78 (0.07)	0.84 (0.05)
	4	0.57 (0.09)	0.83 (0.09)	0.78 (0.09)	0.84 (0.06)
	None	0.57 (0.09)	0.85 (0.09)	0.78 (0.09)	0.83 (0.06)
SVM	1	0.43 (0.09)	0.71 (0.14)	0.68 (0.09)	0.67 (0.10)
	2	0.39 (0.10)	0.67 (0.15)	0.65 (0.13)	0.61 (0.13)
	3	0.37 (0.09)	0.64 (0.16)	0.64 (0.12)	0.58 (0.13)
	4	0.35 (0.09)	0.60 (0.15)	0.63 (0.15)	0.53 (0.11)
	None	0.53 (0.09)	0.80 (0.10)	0.75 (0.09)	0.79 (0.07)

data, in part due to its similarity function, which allows for noise or large deviations in some features to not overwhelm the model [60]. Although SVM is known to perform well when few training samples are available [23], there are also cases where if the data is feature-dense, linear SVM will perform as well as SVM trained with a nonlinear kernel [28], as a large number of features can make a dataset linearly separable [11]. This may be why the non-tensor-reduced datasets tended to have stronger performance than datasets with tensor reduction for SVM.

For RF and LUCCK, both F1 Score and AUROC tended to increase when moving from no tensor reduction to tensor reduction when using only ECG signal data. For example, for LUCCK in the 6-hour dataset, mean F1 score increased from 0.43 to 0.48 with SD remaining similar (0.06 to 0.07), while RF's F1 score increased from 0.41 to 0.48 without a change in SD. We observed a similar increase in mean AUROC for LUCCK (0.60 ± 0.07 to 0.65 ± 0.07) and RF (0.57 ± 0.08 to 0.67 ± 0.06) going from using no tensor reduction to using CP-ALS with rank 4. SVM does not follow this trend, however, and tends to increase in performance as more information is added to the model, with no tensor reduction performing the best. We see a similar trend in the 12-hour dataset.

For 6-hour data, including the ABP features improved both mean F1 Score and mean AUROC across different CP-ALS ranks, as can be seen comparing Figures 4.2 and 4.3. For 12-hour data, RF and LUCCK results are mixed across the different ranks, but including both ABP and ECG data decreased SVM's performance when no tensor reduction took

place. When CP-ALS was used with ranks 1-3 to reduce the feature space for SVM, there is an increase in performance in the ECG + ABP scenario; this suggests that SVM may not be a reliable model for these scenarios.

Adding EHR data to the signal features, presented in Figure 4.4, further improves performance for both the 6- and 12-hour datasets, across all three model types.

While the results of models trained on tensor-reduced signal features show consistent mean AUROC ≥ 0.65 for both LUCCK and RF, it is noted that these experiments were trained on data from only one hospital, the availability of signals led to a small sample pool, and the datasets used do not feature strong racial and ethnic diversity. To ensure the reproducibility and generalizability of these results, it will be necessary to perform similar experiments on a larger and more diverse dataset in future iterations.

4.5 Conclusion

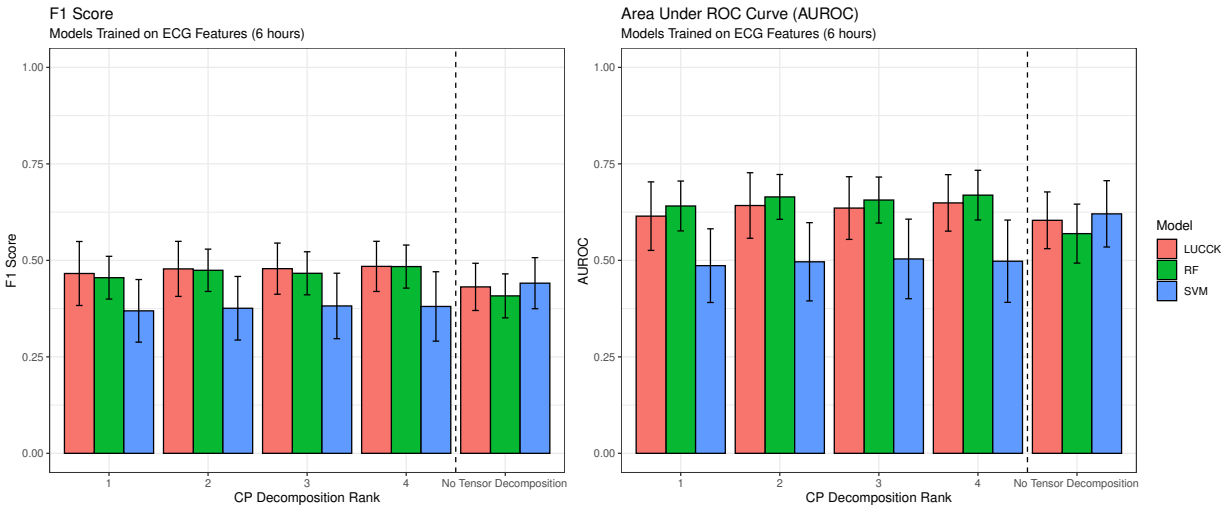
In this study, predictions of increase in qSOFA score were created using tensor-reduced signal features and EHR data. It is possible to make a prediction of increase in qSOFA score using ECG data alone (for RF, AUROC 0.67 $\pm$ 0.06; for LUCCK, 0.65 $\pm$ 0.07), and results can be improved if tensor-reduced ABP features are added (for RF, AUROC 0.71 $\pm$ 0.07; for LUCCK, 0.71 $\pm$ 0.07), but results are mixed when signal features are directly added without tensor reduction (for RF, AUROC 0.69 $\pm$ 0.07; for LUCCK, 0.69 $\pm$ 0.07). This may be because the models are overwhelmed with information, whereas tensor reduction improves performance because only pertinent information is given and noise is removed.

The previous experiments simulate the scenario when EHR data are completely unavailable. When EHR data are available and CP-ALS is used to reduce the feature space of the signal data, results can be further improved (for RF, AUROC 0.77 $\pm$ 0.06; for LUCCK, 0.73 $\pm$ 0.07). This indicates that ECG signal features, ABP signal features, and EHR data features can all contribute to sepsis prognosis.

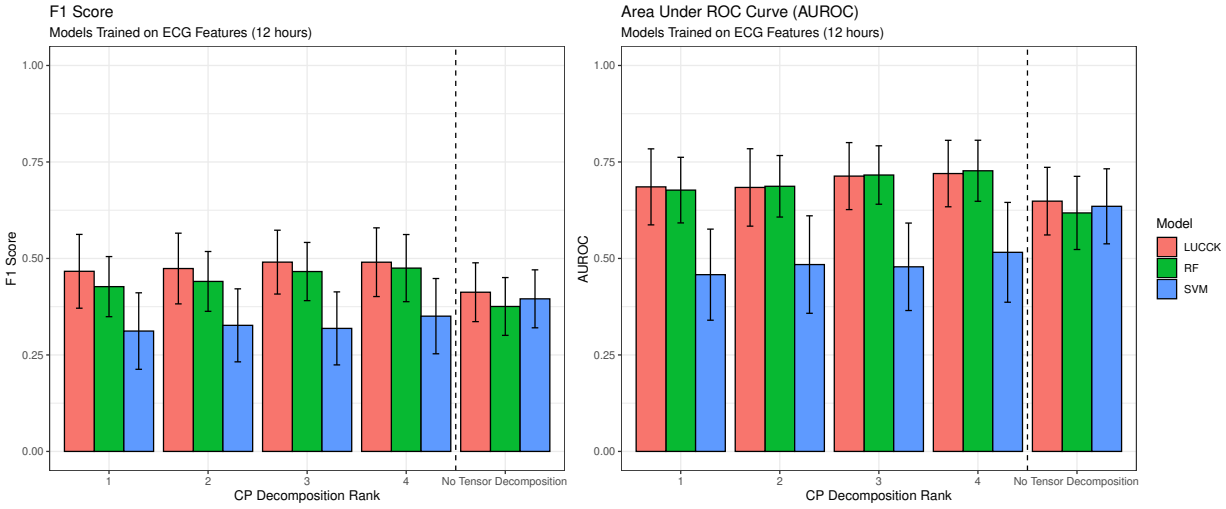
That said, we wish to draw attention to the first scenario, with signals information alone used for model training. The advantage of a signal features-based model is that predictions can be made in the ICU on a continuous basis in real-time; this model would not be limited by the wait times or availability of EHR data variables. From a clinical standpoint, further developing an ECG-only model would be advantageous as, (1) it is minimally invasive compared to an ABP, and (2) it is possible to monitor ECG remotely outside the hospital. Devices such as Holter monitors and Zio patches could be used so that a patient with initially low qSOFA could be monitored at home, with a 6-hour window to predict an increased risk for poor outcomes. Six hours would be adequate time for warning and arrival to the

emergency department to seek appropriate treatment.

We stress that, while it may not achieve F1 or AUROC scores as high as the model including EHR data, our signal features-only model offers an advantage in that it is not prone to issues such as availability or inaccuracies of EHR data. Furthermore, it is continuously collected allowing for real-time evaluation and assessment. For future work, we recommend (1) the combination of EHR, tensor-reduced ECG, and tensor-reduced ABP for use in the hospital or ICU and (2) tensor-reduced ECG only for use in home monitoring.

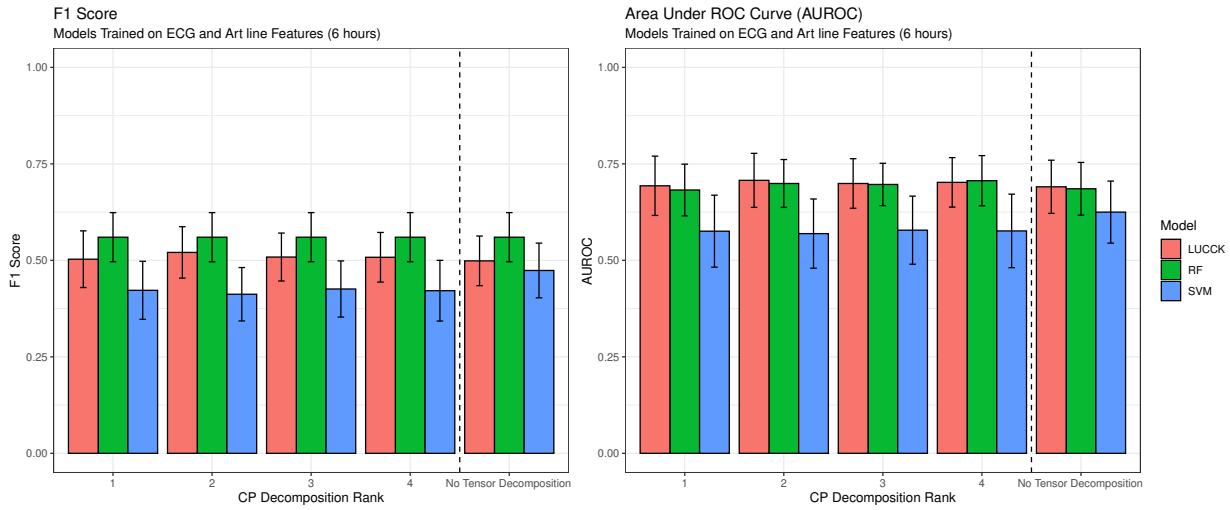


(a) 6-hour data

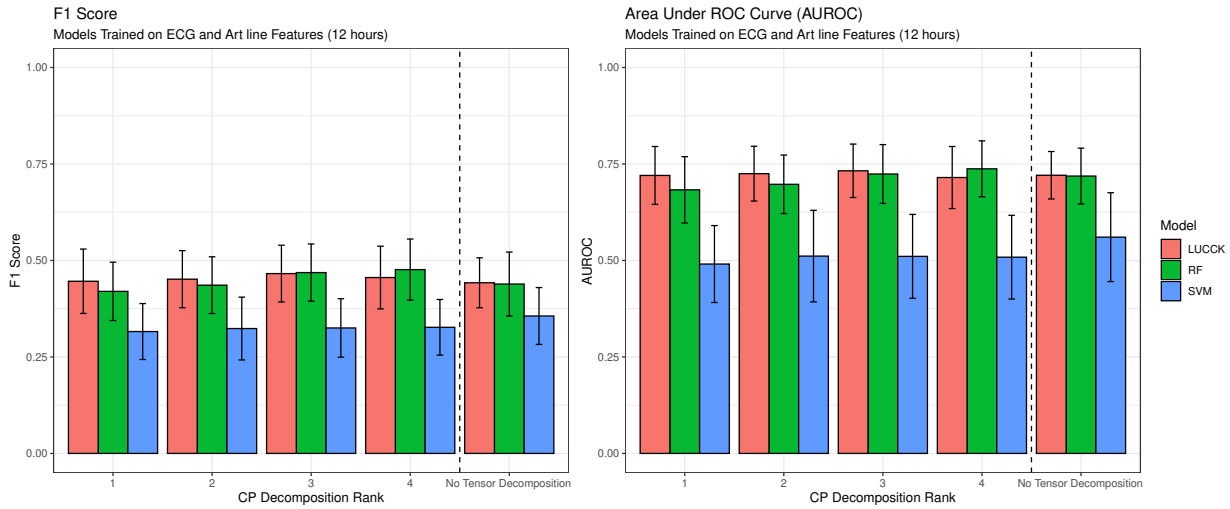


(b) 12-hour data

Figure 4.2: Models Trained with ECG

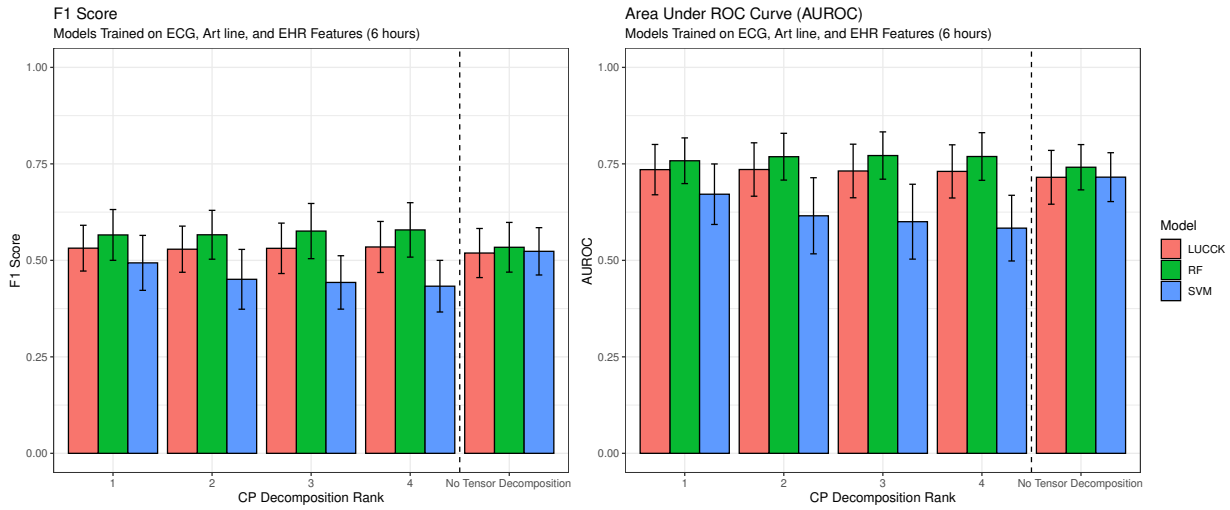


(a) 6-hour data

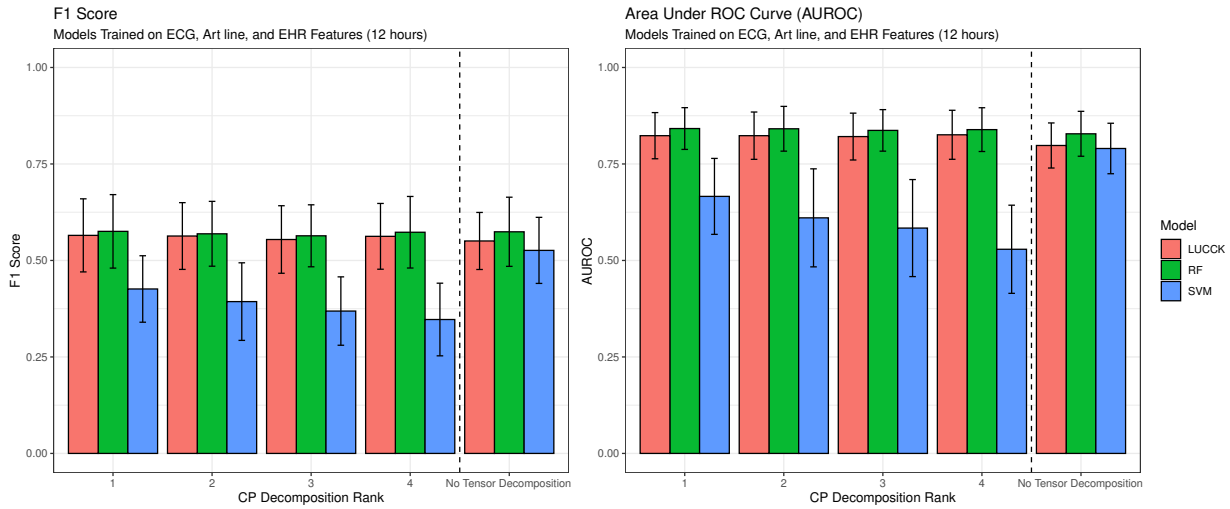


(b) 12-hour data

Figure 4.3: Models Trained with ABP and ECG



(a) 6-hour data



(b) 12-hour data

Figure 4.4: Models Trained with ABP, ECG, and EHR Data

CHAPTER V

Predicting Sepsis Trajectory with Privileged Information and Continuous Physiological Signals

5.1 Introduction

Following from the previous Chapter, the work presented in this Chapter also focuses on identifying patients at risk to develop poor outcomes related to sepsis, using a qSOFA score of ≥ 2 to represent increased risk for poor outcomes. This chapter also builds upon the continuous nature of ECG by including Privileged Information (PI), which, in a machine learning context, means data that is available at the training stage but not at the validation or testing stages. Because the dataset used for model-training is retrospective, we can let our model view future events for the training cases, in order to improve predictions on the test set. Learning using PI is described further in Section 5.2.1.2. The results of this research have been submitted to *MDPI Diagnostics*.

5.2 Methods

5.2.1 Machine Learning

The basic model used for ML was SVM [14]. For all models trained, we used a Gaussian kernel with the sequential minimal optimization [19] solver. A grid search selected a box constraint and kernel scale that resulted in the greatest AUROC value in the validation set. The process of model training was repeated 100 times. In each iteration, the dataset was divided into a training, validation, and test set based on patient, such that no patient in one set (training, test, or validation) could appear in another. The test set was withheld from model training. We recorded the mean and SD of F1 score, sensitivity, specificity, AUROC,

and Area Under the Precision-Recall Curve (AUPRC) over the 100 iterations and report them in Section 5.3.

5.2.1.1 Support Vector Machine

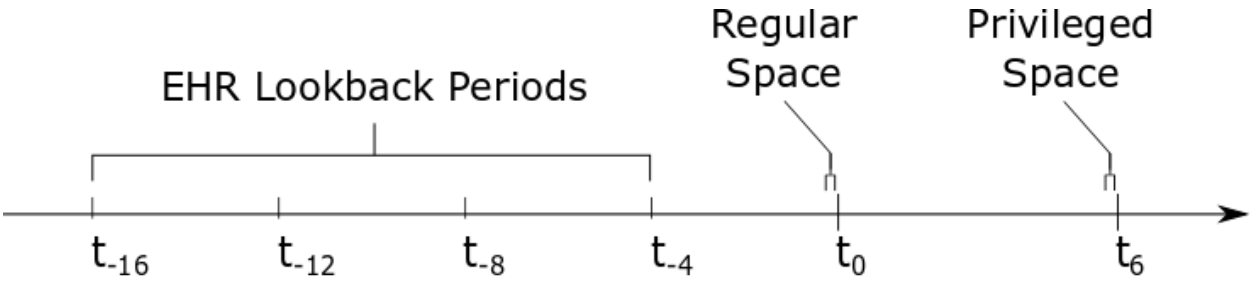
The model that we used to benchmark performance is the SVM with a Gaussian kernel. To learn the decision rule $y = f(x)$, it maps vectors of $x \in X$ into vectors $z \in Z$ and constructs the optimal separating hyperplane between the two classes. The optimal separating hyperplane between the two classes is constructed by learning the decision rule $f(z) = \mathrm{w}z + b$, where w and b are parameters of the hyperplane (weight and bias, respectively), and SVM's objective function is:

$$\min_{\mathrm{w},b,\xi} \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i}^{n} \xi_i$$

with the constraints

$$y_i(\mathrm{w} \cdot z_i + b) \geq 1 - \xi_i, \quad i = 1, ..., n$$
$$\xi \geq 0, \quad C > 0$$

where (x_i, y_i) are a sample's input and label pair, ξ_i functions as a slack variable and C is the penalty parameter [14]. These allow for soft-margin decision boundaries when classes are not linearly separable.



Here, t_0 is the point where qSOFA is 1, and t_6 is six hours later, where qSOFA either increases to 2 or 3 (positive) or not (negative). The times t_{-4} to t_{-16} are lookback periods included in the EHR data in the regular space. The brackets at time t_0 show the ECG signal collected in the regular space, x, and the brackets at time t_6 show the ECG signal collected in the privileged space, x^*.

Figure 5.1: Illustration of Timeline

5.2.1.2 Learning Using Privileged Information

In this work, the regular space was information available at or before time t_0, which was when qSOFA was recorded as being equal to 1. Anything occurring after t_0 was considered the privileged space. An illustration is provided in Figure 5.1 to show where the regular and privileged space for this particular experiment appeared on a timeline. Our method of learning using PI in a medical context was built upon previous work [59, 39, 58]. The features that we included as PI are described in Section 5.2.3.2. The method of SVM+ used in [39] was a modified version of the SVM+ algorithm developed in [77], which in turn was an extension of SVM [14].

As defined by Vapnik and Vashist [77], Learning Using Privileged Information (LUPI) is a paradigm where, in the training stage, the teacher presents both training example x as well as additional information x^* to the learner:

$$x_1, \ldots, x_n \in X \quad \text{and} \quad x_1^*, \ldots, x_n^* \in X^*,$$

where n is the number of samples in the training set, and X and X^* are different spaces. Privileged information is not included in the test or validation sets. Vapnik and Vashist go on to define the paradigm as: when given a set of triplets

$$(x_1, x_1^*, y_1), \ldots, (x_n, x_n^*, y_n)$$

where

$$y \in -1, 1$$

is the classification created according to unknown probability measure $P(x, x^*, y)$, find the function

$$y = f(x, \alpha*), \alpha \in \Lambda$$

that guarantees the smallest probability of incorrect classification.

Building on how SVM maps $x \in X$ to $z \in Z$, SVM+ maps privileged information $x^* \in X^*$ to $z^* \in Z^*$. The objective function of SVM+ is:

$$\min_{w^*, b^*, w, b} \frac{1}{2} \left(\|w\|^2 + \gamma \|w^*\|^2 \right) + C \sum_i^n \xi \left(w^*, b^*, z_i^* \right)$$

such that

$$\begin{aligned} y_i \left(w \cdot z_i + b \right) &\geq 1 - \xi \left(w^*, b^*, z_i^* \right), \\ \xi \left(w^*, b^*, z_i^* \right) &\geq 0, \\ \gamma &> 0 \end{aligned}$$

where $\xi \left(w^*, b^*, z_i^* \right) = w^* \cdot z_i^* + b^*$ is the slack function for the privileged space, replacing the slack variables ξ_i, and γ is a hyperparameter. From this, the hyperplane of SVM+ can be tuned by PI, as privileged training samples x_i^* can be used to regularize the loss from training samples x_i.

We used an implementation of SVM+ created by Li et al., which expanded upon SVM+'s implementation, producing an efficient sequential minimal optimization algorithm to solve the SVM+ problem [39].

5.2.2 Dataset

The data used in this study were from the same retrospective dataset created by the University of Michigan in Chapter IV, and used the same qSOFA score criteria to determine positive and negative classes in a six-hour prediction window.

5.2.2.1 Cohort 1

To create this cohort from the full dataset, we selected for individuals with EHR, ECG, and ABP data available 10 minutes before and up to t_0, as well as 10 minutes before and up to t_6. In this study, EHR data included labs, medications, hourly fluid output, and vital signs. Upon collecting 10-minute signals for feature extraction, signals determined to be 50% or more noise were discarded.

With these conditions in place, the final dataset consisted of 106 instances of 105 patients, with 59 positive cases and 47 negative cases. Due to the small size of the cohort, we opted to use repeated train/test splits rather than 3-fold cross-validation as in Chapter IV. The train/test split was 80/20, with a further 20% of the training set being reserved as a validation set for the grid search.

5.2.2.2 Cohort 2

Due to the small size of cohort 1, we created cohort 2 with more relaxed criteria. Namely, we only selected for individuals with EHR and ECG available in both the regular and privileged space, and omitted the requirement for ABP data. This cohort 2 consisted of 453 instances

of 434 unique individuals, with 144 positive cases and 309 negative cases. We used a similar train/test split as in cohort 1. We created this second, larger cohort for two reasons: (1) to see if ECG- and EHR-related results were consistent across both cohorts, and (2) as ABP is typically only used for critically ill patients [53], we wanted to validate our findings on a greater variety of patients with different statuses.

5.2.3 Signal Processing

For every sample, we collected the 10 minutes of ECG signal occurring directly before the prediction gap for processing. This 10-minute signal was divided into 2 5-minute windows. This constitutes the signal collected in the regular space. For signals collected in the privileged space, we used the 10 minutes of ECG signal directly at the end of the prediction gap, that is, a 10-minute period that ends at the event time, t_6. As in Chapters III and IV, we collected 10 minutes of ECG lead II for the analysis, filtered it with a Butterworth bandpass filter, and divided into two five-minute windows.

5.2.3.1 Feature Extraction in the Regular Space

We calculated peak-based and statistical features from the TS approximation [15] of each window of the 10-minute signal captured six hours before the increase of qSOFA. These TS features have previously been used in previous work in healthcare contexts [6, 26, 33]. For consistency with the study in the previous chapter, the same five ϵ values were used for TS estimation and the same six features were extracted.

In addition to signal features, EHR data features were collected from both the 10 minutes before t_0 as well as four additional lookback periods at $t_{-4}, t_{-8}, t_{-12}, \text{and} t_{-16}$, in accordance with those used in Chapter IV.

5.2.3.2 Feature Extraction in the Privileged Space

To generate features in the privileged space, ten minutes of ECG signal were extracted starting from ten minutes before the event of interest up to the event ($t_6 - 10$ minutes to t_6). The signal underwent the same Butterworth bandpass filter as the regular space ECG data. Two different sets of features were computed from this period of time in the ECG signal. First was a set of statistical summary features: mean, median, variance, kurtosis, skewness, Shannon entropy, and the mean absolute value of the Fast Fourier Transform (FFT). These were adapted from a set of features computed in [26]. This first set of features is referred to as the set of SF-ECG privileged features, with "SF" standing for "statistical features".

To create the second set of features, we applied TS to this 10-minute signal from the privileged space. Using the same ϵ values as from the regular space, we computed the number of line segments, number of inflection segments, total variation of noise, total variation of denoised signal, power of denoised signal, and power of noise over the 10-minute segment. This second set of features is called the TS-ECG privileged features, with "TS" standing for "Taut String".

Lastly, one more set of features was computed from the privileged space: EHR data features. These features were the same as those computed in the regular space, but did not include the four sets of lookback features.

5.3 Results

The tables included here show results of SVM models and SVM+ models trained with different types of privileged information. In each table, the **PI Type** "none" indicates a basic SVM model with a Gaussian kernel. All other PI types use SVM+ to incorporate the privileged information. The row with the greatest AUROC is shown in bold.

PI Type	F1 Score	Sensitivity	Specificity	AUROC	AUPRC
None	0.71 (0.10)	0.71 (0.16)	0.65 (0.17)	0.65 (0.13)	0.66 (0.10)
TS-ECG	0.70 (0.11)	0.69 (0.16)	0.69 (0.14)	**0.68 (0.12)**	0.68 (0.10)
SF-ECG	0.70 (0.10)	0.68 (0.15)	0.68 (0.15)	0.65 (0.12)	0.67 (0.12)
EHR	0.70 (0.12)	0.70 (0.18)	0.66 (0.15)	0.65 (0.13)	0.66 (0.11)

Table 5.1: Taut String ECG in the Regular Space with Different Types of Privileged Information Available in Cohort 1

PI Type	F1 Score	Sensitivity	Specificity	AUROC	AUPRC
None	0.51 (0.06)	0.62 (0.10)	0.63 (0.10)	**0.62 (0.07)**	0.42 (0.07)
TS-ECG	0.48 (0.06)	0.60 (0.11)	0.59 (0.10)	0.58 (0.07)	0.39 (0.07)
SF-ECG	0.50 (0.06)	0.62 (0.10)	0.58 (0.10)	0.59 (0.07)	0.39 (0.06)
EHR	0.51 (0.07)	0.64 (0.10)	0.59 (0.09)	0.61 (0.08)	0.40 (0.08)

Table 5.2: Taut String ECG in the Regular Space with Different Types of Privileged Information Available in Cohort 2

For models trained on cohort 1 with Taut String ECG data, shown in Table 5.1, using SVM+ with additional Taut String ECG privileged information increased average AUROC by 0.03 and average AUPRC by 0.02, with standard deviation remaining similar, compared to the base SVM model. Cohort 2 does not show this increase with PI, but rather, yields the

highest AUROC and F1 Score when no PI is added. Although cohort 2's average F1 Score, AUROC and AUPRC are lower than cohort 1's, the standard deviation for each is smaller, shown in Table 5.2.

PI Type	F1 Score	Sensitivity	Specificity	AUROC	AUPRC
None	0.69 (0.09)	0.66 (0.13)	0.71 (0.14)	**0.65 (0.11)**	0.66 (0.10)
TS-ECG	0.69 (0.10)	0.67 (0.15)	0.70 (0.14)	0.64 (0.12)	0.65 (0.09)
SF-ECG	0.68 (0.10)	0.67 (0.15)	0.67 (0.14)	0.62 (0.12)	0.65 (0.10)
EHR	0.68 (0.11)	0.66 (0.16)	0.69 (0.16)	0.63 (0.13)	0.65 (0.10)

Table 5.3: Results of ECG and EHR in the Regular Space for Cohort 1

PI Type	F1 Score	Sensitivity	Specificity	AUROC	AUPRC
None	0.60 (0.06)	0.70 (0.09)	0.69 (0.09)	**0.72 (0.05)**	0.54 (0.08)
TS-ECG	0.58 (0.05)	0.70 (0.09)	0.67 (0.08)	0.70 (0.06)	0.50 (0.08)
SF-ECG	0.58 (0.05)	0.69 (0.08)	0.68 (0.08)	0.71 (0.05)	0.51 (0.07)
EHR	0.59 (0.05)	0.70 (0.09)	0.69 (0.08)	0.72 (0.05)	0.53 (0.06)

Table 5.4: Results of ECG and EHR in the Regular Space for Cohort 2

Models trained on both TS and EHR are shown in Table 5.3 for cohort 1 and Table 5.4 for cohort 2. Neither model shows improvement upon adding PI. Cohort 1's results show a greater F1 Score and AUPRC compared to cohort 2's results, but cohort 2 has an increased AUROC, with smaller standard deviations across all values.

PI Type	F1 Score	Sensitivity	Specificity	AUROC	AUPRC
None	0.59 (0.15)	0.59 (0.21)	0.61 (0.18)	0.51 (0.13)	0.55 (0.10)
TS-ECG	0.59 (0.13)	0.59 (0.19)	0.58 (0.17)	0.49 (0.12)	0.54 (0.09)
SF-ECG	0.62 (0.11)	0.61 (0.17)	0.60 (0.17)	0.51 (0.12)	0.56 (0.09)
EHR	0.62 (0.12)	0.59 (0.17)	0.64 (0.17)	**0.54 (0.12)**	0.58 (0.09)

Table 5.5: Results of EHR in the Regular Space for Cohort 1

For models trained on EHR data in cohort 1, shown in Table 5.5, adding privileged EHR data increased mean F1 Score, AUROC, and AUPRC by 0.03 with standard deviation decreasing in all cases. In cohort 2, adding PI did not improve performance. However, AUROC is higher and with a smaller standard deviation compared to cohort 1, as shown in Table 5.6.

PI Type	F1 Score	Sensitivity	Specificity	AUROC	AUPRC
None	0.59 (0.06)	0.68 (0.09)	0.71 (0.10)	**0.71 (0.06)**	0.55 (0.09)
TS-ECG	0.55 (0.06)	0.67 (0.10)	0.65 (0.09)	0.67 (0.07)	0.47 (0.07)
SF-ECG	0.56 (0.06)	0.68 (0.09)	0.66 (0.07)	0.68 (0.06)	0.49 (0.08)
EHR	0.57 (0.06)	0.66 (0.09)	0.68 (0.10)	0.68 (0.06)	0.51 (0.08)

Table 5.6: Results of EHR in the Regular Space for Cohort 2

5.4 Discussion and Conclusion

For the two cohorts in the previous sections, we found differing effects of adding PI to a SVM model. In cohort 1, the smaller cohort which selected for patients more likely to be critically ill, adding taut string privileged information was slightly beneficial when ECG alone was being used as the regular space (AUROC 0.68 ± 0.12 compared to 0.65 ± 0.13). In cohort 2, the larger and broader cohort, PI was not as informative to the models in any of the presented scenarios.

For cohort 1, the TS-ECG SVM+ model with ECG as the regular space outperformed EHR in the regular space and ECG and EHR in the regular space across F1 Score, AUROC, and AUPRC. Cohort 2 had more positive influence from EHR data, where the models including both ECG and EHR data in the regular space outperformed any variation of ECG or EHR data alone in the regular space, regardless of adding PI. In both cohorts, ECG information is strongly contributing to the model.

It is possible that the EHR data is more informative in the broader cohort as the patients are more diverse; critically ill patients would be receiving similar antibiotic, vasopressor, and other therapies, and therefore EHR data would be similar across all patients, whereas a broader patient cohort may have different treatments being given to them, making EHR data more distinctive between the more and less severe cases.

Cohort 1 was initially selected with the goal of also including arterial line features as both regular space and privileged information features, however, neither of these features significantly improved performance compared to the models only trained on ECG data.

It is also noted that in addition to the dataset being somewhat small, when constraints based on signal availability are created, the dataset also loses racial and ethnic diversity, with the vast majority of the cohort being made of white individuals, although distribution of sex was roughly equal. Studies of sepsis prognosis using LUPI should be replicated on both larger and more diverse cohorts outside of this one particular hospital, to ensure that results are generalizable to a greater patient population.

Future trials could investigate different lengths of signal or windowing parameters, as well

as different designs of PI collection. For example, Sabeti et al. have used LUPI where PI is only available for certain samples, using a "learning using partially available privileged information" paradigm [58, 59]. Additionally, different outcome variables, such as start of mechanical ventilation, vasopressor administration, change of antibiotic dose, or others which are clinically relevant, could be studied with a LUPI approach.

CHAPTER VI

Concluding Remarks and Future Direction

Ongoing research into machine learning and artificial intelligence have been changing the medical landscape, from clinical decision support, advances in radiology and image processing, to novel drug design and re-purposing. These different fields all work toward the goal of improving the quality of care for patients, and alleviating burden for physicians and other caretakers.

In this thesis, I proposed different scenarios for using continuously updating signals for use in different medical applications, either in an ambulatory (Chapter II) or hospital setting (Chapters III-V). This research also includes the use of different signal processing methods for ECG, HRV, and ABP, such as using TS, DTCWPT, filter banks, and using a multimodal approach to also incorporate EHR data. These different methods were tested and validated on data obtained from the University of Michigan.

In Chapter II, signal processing of EDA and HRV were used to predict poor sleep quality in fibromyalgia patients compared to non-fibromyalgia controls. In Chapter III, signal processing of ECG, ABP, other signals and EHR data were used in a multimodal approach to predict adverse events post-surgery. Chapters IV and V use signal processing of ECG as well as ABP and/or EHR data to predict increased risk of poor outcomes related to sepsis.

Much of the research performed in these analyses was limited by the availability of data. For example, both groups in Chapter II and cohort 3 of Chapter III were limited by the small number of individuals in each cohort.

In addition to small datasets, the availability of signal data across different individuals was a limitation. For instance, in Chapters IV and V, increase in qSOFA score was chosen as the outcome of interest, but other outcomes were investigated before selecting this choice. For example, we also explored the placing of the patient on mechanical ventilation, the patient beginning a course of antibiotics, or being given a vasopressor as potential outcomes. However, very few patients had ECG or other signal data available before the first event of interest; rather, in most cases, signals were only available once patients were already being given these treatments.

In addition to the limitation of signal availability, we also lost patient diversity across the cohorts. As can be seen in Appendices B and C, selecting for patients who have both (1) signal data available at the desired time and (2) signal data deemed high enough quality or noise-free enough for analysis, reduces the already limited racial and ethnic diversity of the full dataset. It is worth investigating if the limited availability of these physiological signals for non-white patient groups is due to unintended biases from equipment, such as how PPG readings may be less accurate for Black patients [66], or potential biases from providers in determining which patients should have these signals collected and monitored.

In future work, I would like to explore signals collected earlier in the patients' treatment, to be able to better view the changes in their physiological signals before starting treatments which can affect the signals themselves. Similar to how Chapter III focuses on generalizing the methodology used to predict adverse outcomes post cardiac surgery to more heterogeneous cohorts, replicating these studies on other cohorts outside of the University of Michigan hospital system would provide increased insight into the generalizability of the methods proposed and increase patient diversity in the development of these models, especially for conditions as heterogeneous as sepsis. The limited availability of longitudinal signal data, both in the retrospective biobank used in Chapters IV and V as well as public datasets, is currently a limiting factor. Ambulatory devices, such as the wristband in Chapter II or Holter monitors, provide opportunities for studying longitudinal signal data outside of the hospital setting, and could be less expensive than initiating telemetry on many more patients.

Further study of the utility of routinely collected, continuously updating, and noninvasive ECG signals in the hospital setting should be pursued. Over all, this thesis is part of a greater effort for developing clinical decision support systems for the ICU and other settings where continuous status updates on patients are warranted. Using computationally efficient methods such as taut string, and structure-preserving methods such as tensor decomposition, serve to further this goal. The methods developed and tested in this thesis contribute to improving prognosis for patients in an ambulatory setting for Chapter II, and in hospital settings for the remaining chapters. Our long-term goal is to develop automated systems for processing these continuous signals in real-time to aid caretakers in their decision-making processes. The continuous nature of these signals makes them especially useful in providing near real-time updates of a patient's status. In addition to using continuous signals, their mapping to individual patients can further be used in personalized care.

One future application of these methods is to move from the more population-focused models to models tailored to the individual. For example, reinforcement learning can be used to optimize these models on the patient level, with the individual's EHR data and

ECG signals continuously updating the model to better inform the model's predictions.

APPENDIX A

Features Extracted from BVP and EDA

This is the list of features extracted from BVP and EDA. The abbreviations used are Root Mean Square (RMS), Very Low Frequency (VLF), Low Frequency (LF), and High Frequency (HF).

Table A.1: Table of Features for Each Five-Minute Window

Signal	Feature List
BVP (Kubios Features)	Mean and SD of RR intervals, Mean and SD of heart rate, RMS of successive RR interval differences, Count of successive RR interval differences > 50 milliseconds, Area, Height, and Baseline width of RR histogram, VLF, LF, and HF band peak frequencies, Absolute VLF, LF, and HF power, Relative VLF, LF, and HF power, Normalized LF and HF power, LF/HF power ratio, Approximate entropy, Sample entropy, Poincaré plot short- and long-term variability
BVP (Additional Features)	Number of peaks, Mean, Median, SD, Maximum, Minimum of: Time between consecutive peaks, Time between first and secondary peak in cycle, Relative amplitude between consecutive primary peaks, Relative amplitude between primary and secondary peaks
EDA	From tonic signal: Mean, Variance From filter bank coefficients: Mean, Variance, Maximum, Kurtosis, Skewness, Area, Shannon entropy

APPENDIX B

Demographics of Michigan Medicine Retrospective Dataset

A table of patient demographics from the Michigan Medicine retrospective dataset. The first column lists characteristics, and the second column give the counts of each characteristic in the full dataset. The third through sixth give counts of each characteristic for the positive and negative outcomes in the 6- and 12-hour gap datasets. Note that the 6- and 12-hour datasets are subsets of the total cohort of 1,803 patients who met the inclusion/exclusion criteria.

Characteristic	Full cohort (N = 1803)	6 hours, positive (N = 59)	6 hours, negative (N = 199)	12 hours, positive (N = 37)	12 hours, negative (N = 189)
Age, Mean (SD)	58.9 (17.9)	60.0 (16.8)	58.7 (16.9)	59.7 (17.0)	58.5 (16.9)
Race and Ethnicity					
Asian	20	1	2	0	2
Black or African American	198	9	14	4	14
Hispanic or Latine	28	0	2	0	2
White	1520	48	175	32	166
Other	65	1	8	1	7
Sex					
Female	866	26	91	16	84
Male	937	33	108	21	105

Table B.1: Characteristics of Patients from the Michigan Medicine Dataset

APPENDIX C

Demographics of Second Michigan Medicine Retrospective Dataset

Table C.1: Characteristics of Cohorts

Characteristic	Full Cohort (N = 1803)	Cohort 1 (N = 105)	Cohort 2 (N = 434)
Age, Mean (SD)	58.9 (17.9)	56.6 (17.2)	56.2 (18.9)
Sex, Female/Male	866/937	48/57	221/213
Race and Ethnicity			
Asian	20	1	9
Black or African-American	198	13	54
Hispanic or Latine	28	0	6
White	1520	88	352
Other	65	3	19

The first column lists patient characteristics, and the second gives counts of each characteristic in the full dataset. The third and fourth columns give counts of each characteristic for cohorts 1 and 2. Note that cohort 1 and cohort 2 are both subsets of the total cohort, and cohort 1 is a subset of cohort 2.

BIBLIOGRAPHY

[1] Olivia Alge, S. M. Reza Soroushmehr, Jonathan Gryak, Anna Kratz, and Kayvan Najarian. Predicting Poor Sleep Quality in Fibromyalgia with Wrist Sensors. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4290–4293, Montreal, QC, Canada, July 2020. IEEE.

[2] Ruben Amarasingham, Ferdinand Velasco, Bin Xie, Christopher Clark, Ying Ma, Song Zhang, Deepa Bhat, Brian Lucena, Marco Huesch, and Ethan A. Halm. Electronic medical record-based multicondition models to predict the risk of 30 day readmission or death among adult medicine patients: validation and comparison to existing models. *BMC Medical Informatics and Decision Making*, 15(1):39, December 2015.

[3] Anthony C. Antonacci, Samuel P. Dechario, Caroline Antonacci, Gregg Husk, Vihas Patel, Jeffrey Nicastro, Gene Coppa, and Mark Jarrett. Cognitive Bias Impact on Management of Postoperative Complications, Medical Error, and Standard of Care. *Journal of Surgical Research*, 258:47–53, February 2021.

[4] Brett W. Bader, Tamara G. Kolda, and others. Matlab Tensor Toolbox, August 2017.

[5] I. Bayram and I.W. Selesnick. On the Dual-Tree Complex Wavelet Packet and M-Band Transforms. *IEEE Transactions on Signal Processing*, 56(6):2298–2310, June 2008.

[6] Ashwin Belle, Sardar Ansari, Maxwell Spadafore, Victor A. Convertino, Kevin R. Ward, Harm Derksen, and Kayvan Najarian. A Signal Processing Approach for Detection of Hemodynamic Instability before Decompensation. *PloS One*, 11(2):e0148544, 2016.

[7] Tony Berger, Jeffrey Green, Timothy Horeczko, Yolanda Hagar, Nidhi Garg, Alison Suarez, Edward Panacek, and Nathan Shapiro. Shock Index and Early Recognition of Sepsis in the Emergency Department: Pilot Study. *Western Journal of Emergency Medicine*, 14(2):168–174, March 2013.

[8] Roger C. Bone, Robert A. Balk, Frank B. Cerra, R. Phillip Dellinger, Alan M. Fein, William A. Knaus, Roland M.H. Schein, and William J. Sibbald. Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis. *Chest*, 101(6):1644–1655, June 1992.

[9] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[10] David Cella, William Riley, Arthur Stone, Nan Rothrock, Bryce Reeve, Susan Yount, Dagmar Amtmann, Rita Bode, Daniel Buysse, Seung Choi, Karon Cook, Robert Devellis, Darren DeWalt, James F. Fries, Richard Gershon, Elizabeth A. Hahn, Jin-Shei Lai, Paul Pilkonis, Dennis Revicki, Matthias Rose, Kevin Weinfurt, Ron Hays, and PROMIS Cooperative Group. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*, 63(11):1179–1194, November 2010.

[11] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408:189–215, September 2020.

[12] Rebanta K. Chakraborty and Bracken Burns. Systemic Inflammatory Response Syndrome. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2022.

[13] Aastha Chokshi, David Cennimo, Helen Horng, and Ziad Sifri. Global contributors to antibiotic resistance. *Journal of Global Infectious Diseases*, 11(1):36, 2019.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach Learn*, 20(3):273–297, September 1995.

[15] P. L. Davies and A. Kovac. Local Extremes, Runs, Strings and Multiresolution. *The Annals of Statistics*, 29(1):1 – 65, 2001.

[16] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A Multilinear Singular Value Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, January 2000.

[17] Empatica. E4 wristband user's manual, 2018.

[18] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M. Coopersmith, Craig French, Flávia R. Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C. Prescott, Christa Schorr, Steven Simpson, W. Joost Wiersinga, Fayez Alshamsi, Derek C. Angus, Yaseen Arabi, Luciano Azevedo, Richard Beale, Gregory Beilman, Emilie Belley-Cote, Lisa Burry, Maurizio Cecconi, John Centofanti, Angel Coz Yataco, Jan De Waele, R. Phillip Dellinger, Kent Doi, Bin Du, Elisa Estenssoro, Ricard Ferrer, Charles Gomersall, Carol Hodgson, Morten Hylander Møller, Theodore Iwashyna, Shevin Jacob, Ruth Kleinpell, Michael Klompas, Younsuck Koh, Anand Kumar, Arthur Kwizera, Suzana Lobo, Henry Masur, Steven McGloughlin, Sangeeta Mehta, Yatin Mehta, Mervyn Mer, Mark Nunnally, Simon Oczkowski, Tiffany Osborn, Elizabeth Papathanassoglou, Anders Perner, Michael Puskarich, Jason Roberts, William Schweickert, Maureen Seckel, Jonathan Sevransky, Charles L. Sprung, Tobias Welte, Janice Zimmerman, and Mitchell Levy. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021. *Critical Care Medicine*, 49(11):e1063–e1143, November 2021.

[19] Rong-En Fan, Pai-Hsuen Chen, and Chih-Jen Lin. Working set selection using second order information for training support vector machines. *Journal of Machine Learning Research*, 6(63):1889–1918, 2005.

[20] Andrea Ferrario and Michele Loi. How Explainability Contributes to Trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1457–1466, Seoul Republic of Korea, June 2022. ACM.

[21] Flavio Lopes Ferreira, Daliana Peres Bota, Annette Bross, Christian Mélot, and Jean-Louis Vincent. Serial evaluation of the sofa score to predict outcome in critically ill patients. *Jama*, 286(14):1754–1758, 2001.

[22] Nicolò Gambarotta, Federico Aletti, Giuseppe Baselli, and Manuela Ferrario. A review of methods for the signal quality assessment to improve reliability of heart rate and blood pressures derived parameters. *Medical & Biological Engineering & Computing*, 54(7):1025–1035, July 2016.

[23] Raoof Gholami and Nikoo Fakhari. Support Vector Machine: Principles, Parameters, and Applications. In *Handbook of Neural Computation*, pages 515–535. Elsevier, 2017.

[24] Henry Han and Xiaoqian Jiang. Overcome Support Vector Machine Diagnosis Overfitting. *Cancer Informatics*, 13s1:CIN.S13875, January 2014.

[25] Shiying Hao, Bo Jin, Andrew Young Shin, Yifan Zhao, Chunqing Zhu, Zhen Li, Zhongkai Hu, Changlin Fu, Jun Ji, Yong Wang, Yingzhen Zhao, Dorothy Dai, Devore S. Culver, Shaun T. Alfreds, Todd Rogow, Frank Stearns, Karl G. Sylvester, Eric Widen, and Xuefeng B. Ling. Risk Prediction of Emergency Department Revisit 30 Days Post Discharge: A Prospective Study. *PLoS ONE*, 9(11):e112944, November 2014.

[26] Larry Hernandez, Renaid Kim, Neriman Tokcan, Harm Derksen, Ben E. Biesterveld, Alfred Croteau, Aaron M. Williams, Michael Mathis, Kayvan Najarian, and Jonathan Gryak. Multimodal tensor-based method for integrative and continuous patient monitoring during postoperative cardiac care. *Artificial Intelligence in Medicine*, 113:102032, March 2021.

[27] Christopher J. Hillar and Lek-Heng Lim. Most Tensor Problems Are NP-Hard. *Journal of the ACM*, 60(6):1–39, November 2013.

[28] Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear SVM. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 408–415, Helsinki, Finland, 2008. ACM Press.

[29] Shujun Huang, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics*, 15(1):41–51, 2018.

[30] Michael Imhoff and Silvia Kuhls. Alarm Algorithms in Critical Care Monitoring:. *Anesthesia & Analgesia*, 102(5):1525–1537, May 2006.

[31] Kendall R. Johnson, James I. Hagadorn, and David W. Sink. Alarm Safety and Alarm Fatigue. *Clinics in Perinatology*, 44(3):713–728, September 2017.

[32] Mohammad Karimi Moridani, Seyed Kamaledin Setarehdan, Ali Motie Nasrabadi, and Esmaeil Hajinasrollah. Non-linear feature extraction from HRV signal for mortality prediction of ICU cardiovascular patient. *Journal of Medical Engineering & Technology*, 40(3):87–98, April 2016.

[33] Renaid B. Kim, Olivia P. Alge, Gang Liu, Ben E. Biesterveld, Glenn Wakam, Aaron M. Williams, Michael R. Mathis, Kayvan Najarian, and Jonathan Gryak. Prediction of postoperative cardiac events in multiple surgical cohorts using a multimodal and integrative decision support system. *Scientific Reports*, 12(1):11347, December 2022.

[34] William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

[35] Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, August 2009.

[36] Anna L. Kratz, Daniel Whibley, Samsuk Kim, Martin Sliwinski, Daniel Clauw, and David A. Williams. Fibrofog in daily life: An examination of ambulatory subjective and objective cognitive function in fibromyalgia. *Arthritis Care Res (Hoboken)*, October 2019.

[37] Alexandre Laurin. BP_annotate, April 2017.

[38] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

[39] Wen Li, Dengxin Dai, Mingkui Tan, Dong Xu, and Luc Van Gool. Fast Algorithms for Linear and Kernel SVM+. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2258–2266, Las Vegas, NV, 2016. IEEE.

[40] Gang Liu, Renaid Kim, Shuyang Cheng, Harm Derksen, B.E. Biesterveld, A.M. Williams, Michael Mathis, Kayvan Najarian, and Jonathan Gryak. Prediction of hemodynamic decompensation in patients recovering from major vascular surgeries using a multimodal tensor - based approach. In *Military Health System Research Symposium (MHSRS)*, August 2021.

[41] Yurong Luo. *The severity of stages estimation during hemorrhage using error correcting output codes method*. PhD thesis, VCU Libraries, 2012. Publication Title: VCU Theses and Dissertations.

[42] Michael R. Mathis, Milo Engoren, Sachin Kheterpal, Kyle Gunnerson, Aaron Williams, Benjamin Biesterveld, Alfred Croteau, Kevin Ward, Hasan Alam, Harm Derksen, Gang Liu, Renaid Kim, Neriman Tokcan, Kayvan Najarian, and Jonathan Gryak. Early detection of postoperative deterioration in cardiac surgery patients using electronic health

record and waveform data: A machine learning approach. *International Anesthesia Research Society*, 132:999–1003, May 2021.

[43] Paolo Melillo, Raffaele Izzo, Ada Orrico, Paolo Scala, Marcella Attanasio, Marco Mirra, Nicola De Luca, and Leandro Pecchia. Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis. *PLOS ONE*, 10(3):e0118504, March 2015.

[44] J. Randall Moorman, John B. Delos, Abigail A. Flower, Hanqing Cao, Boris P. Kovatchev, Joshua S. Richman, and Douglas E. Lake. Cardiovascular oscillations at the bedside: early diagnosis of neonatal sepsis using heart rate characteristics monitoring. *Physiological Measurement*, 32(11):1821–1832, October 2011.

[45] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, and T. Lyons. The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. In *2019 Computing in Cardiology (CinC)*, pages Page 1–Page 4, 2019.

[46] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D. Stanley, Gari D. Clifford, and Timothy G. Buchman. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU:. *Critical Care Medicine*, 46(4):547–553, April 2018.

[47] Naimahmed Nesaragi, Shivnarayan Patidar, and Veerakumar Thangaraj. A correlation matrix-based tensor decomposition method for early prediction of sepsis from clinical data. *Biocybernetics and Biomedical Engineering*, 41(3):1013–1024, July 2021.

[48] Sean M. O'Brien, Liqi Feng, Xia He, Ying Xian, Jeffrey P. Jacobs, Vinay Badhwar, Paul A. Kurlansky, Anthony P. Furnary, Joseph C. Cleveland, Kevin W. Lobdell, Christina Vassileva, Moritz C. Wyler von Ballmoos, Vinod H. Thourani, J. Scott Rankin, James R. Edgerton, Richard S. D'Agostino, Nimesh D. Desai, Fred H. Edwards, and David M. Shahian. The society of thoracic surgeons 2018 adult cardiac surgery risk models: part 2—statistical methods and results. *The Annals of Thoracic Surgery*, 105(5):1419–1428, 2018.

[49] Jiapu Pan and Willis J. Tompkins. A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, March 1985.

[50] Maryam Panahiazar, Vahid Taslimitehrani, Naveen Pereira, and Jyotishman Pathak. Using EHRs and Machine Learning for Heart Failure Survival Analysis. *Studies in Health Technology and Informatics*, 216:40–44, 2015.

[51] Carly J. Paoli, Mark A. Reynolds, Meenal Sinha, Matthew Gitlin, and Elliott Crouser. Epidemiology and Costs of Sepsis in the United States-An Analysis Based on Timing of Diagnosis and Severity Level. *Critical Care Medicine*, 46(12):1889–1897, 2018.

[52] Roberta Petrino, Luis Garcia-Castrillo Riesgo, and Basak Yilmaz. Burnout in emergency medicine professionals after 2 years of the COVID-19 pandemic: a threat to the healthcare system? *European Journal of Emergency Medicine*, 29(4):279–284, August 2022.

[53] Louisdon Pierre, Divij Pasrija, and Michael Keenaghan. Arterial Lines. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023.

[54] Francesca Prestinaci, Patrizio Pezzotti, and Annalisa Pantosti. Antimicrobial resistance: a global multifaceted phenomenon. *Pathogens and Global Health*, 109(7):309–318, October 2015.

[55] Yanjun Qi. Random Forest for Bioinformatics. In Cha Zhang and Yunqian Ma, editors, *Ensemble Machine Learning*, pages 307–323. Springer New York, New York, NY, 2012.

[56] Luiz Paulo Queiroz. Worldwide Epidemiology of Fibromyalgia. *Curr Pain Headache Rep*, 17(8):356, August 2013.

[57] Srinivas G Rao and Robert M Bennett. Pharmacological therapies in fibromyalgia. *Best Practice & Research Clinical Rheumatology*, 17(4):611–627, August 2003.

[58] Elyas Sabeti, Joshua Drews, Narathip Reamaroon, Jonathan Gryak, Michael Sjoding, and Kayvan Najarian. Detection of acute respiratory distress syndrome by incorporation of label uncertainty and partially available privileged information. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1717–1720, 2019.

[59] Elyas Sabeti, Joshua Drews, Narathip Reamaroon, Elisa Warner, Michael W. Sjoding, Jonathan Gryak, and Kayvan Najarian. Learning Using Partially Available Privileged Information and Label Uncertainty: Application in Detection of Acute Respiratory Distress Syndrome. *IEEE Journal of Biomedical and Health Informatics*, 25(3):784–796, 2021.

[60] Elyas Sabeti, Jonathan Gryak, Harm Derksen, Craig Biwer, Sardar Ansari, Howard Isenstein, Anna Kratz, and Kayvan Najarian. Learning Using Concave and Convex Kernels: Applications in Predicting Quality of Sleep and Level of Fatigue in Fibromyalgia. *Entropy*, 21(5):442, April 2019.

[61] Akane Sano, Rosalind W. Picard, and Robert Stickgold. Quantitative analysis of wrist electrodermal activity during sleep. *International Journal of Psychophysiology*, 94(3):382–389, December 2014.

[62] Hooman Sedghamiz. Matlab Implementation of Pan Tompkins ECG QRS detector., March 2014.

[63] Francesco Sessa, Valenzano Anna, Giovanni Messina, Giuseppe Cibelli, Vincenzo Monda, Gabriella Marsala, Maria Ruberto, Antonio Biondi, Orazio Cascio, Giuseppe Bertozzi, Daniela Pisanelli, Francesca Maglietta, Antonietta Messina, Maria P. Mollica, and Monica Salerno. Heart rate variability as predictive factor for sudden cardiac death. *Aging*, 10(2):166–177, February 2018.

[64] Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus.

Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):762, February 2016.

[65] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810, February 2016.

[66] Michael W. Sjoding, Robert P. Dickson, Theodore J. Iwashyna, Steven E. Gay, and Thomas S. Valley. Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine*, 383(25):2477–2478, December 2020.

[67] Miroslava Sladojevic, Milenko Cankovic, Snezana Cemerlic, Bojan Mihajlovic, Filip Adjic, and Milana Jarakovic. Data mining approach for in-hospital treatment outcome in patients with acute coronary syndrome. *Medicinski pregled*, 68(5-6):157–161, 2015.

[68] Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. Publication recommendations for electrodermal measurements: Publication standards for EDA. *Psychophysiology*, 49(8):1017–1034, August 2012.

[69] Mika P. Tarvainen, Juha-Pekka Niskanen, Jukka A. Lipponen, Perttu O. Ranta-Aho, and Pasi A. Karjalainen. Kubios HRV–heart rate variability analysis software. *Comput Methods Programs Biomed*, 113(1):210–220, 2014.

[70] R. Andrew Taylor, Joseph R. Pare, Arjun K. Venkatesh, Hani Mowafi, Edward R. Melnick, William Fleischman, and M. Kennedy Hall. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Academic Emergency Medicine*, 23(3):269–278, March 2016.

[71] Alice Theadom and Mark Cropley. 'This constant being woken up is the worst thing' – experiences of sleep in fibromyalgia syndrome. *Disability and Rehabilitation*, 32(23):1939–1947, January 2010.

[72] Alice Theadom, Mark Cropley, and Kirsty-Louise Humphrey. Exploring the role of sleep and coping in quality of life in fibromyalgia. *Journal of Psychosomatic Research*, 62(2):145–151, February 2007.

[73] John Trinder, Jan Kleiman, Melinda Carrington, Simon Smith, Sibilah Breen, Nellie Tan, and Young Kim. Autonomic activity during human sleep as a function of time and sleep stage. *J Sleep Res*, 10(4):253–264, December 2001.

[74] Ledyard R Tucker. Implications of factor analysis of three-way matrices for measurement of change. *Problems in measuring change*, 15:122–137, 1963.

[75] Ledyard R Tucker. The extension of factor analysis to three-dimensional matrices. *Contributions to mathematical psychology*, 110119, 1964.

[76] J. Scott VanEpps. Reducing exposure to broad-spectrum antibiotics for bloodstream infection. *Journal of Laboratory and Precision Medicine*, 3:100–100, December 2018.

[77] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

[78] Luxmi Verma, Sangeet Srivastava, and P. C. Negi. A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. *Journal of Medical Systems*, 40(7):178, July 2016.

[79] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*, 181(8):1065, August 2021.