A Framework for Evaluation and Identification of Time Series Models for Multi-Step Ahead Prediction of Physiological Signals

by

Hisham Mohammad Wahbi S. ElMoaqet

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mechanical Engineering)
in The University of Michigan
2015

Doctoral Committee:

Professor Dawn Tilbury, Chair
Professor Satinder Singh Baveja
Assistant Professor Eunshin Byon
Professor Jun Ni
Assistant Professor Satya Krishna Ramachandran

To my mother and father, who taught me the value of education.

ACKNOWLEDGEMENTS

First and foremost, I am grateful to Professor Dawn Tilbury for being the best PhD research advisor I could ask for. She has continuously provided me with her inspiration, encouragement, and unwavering confidence in my abilities. Her support and guidance have helped me learn how to do research and make contributions. She has been a great teacher, mentor, and a role model that I will follow in my future academic career. None of the work presented in this dissertation would have been possible without her inspiration and technical intuition.

I owe a tremendous debt of gratitude to Professor Satya Krishna Ramachandran, whose contributions to the work presented in this dissertation are immeasurable. He has given valuable directions to the research with his clinical insights and experience and also helped me gain access to the data sets that were essential for the development and validation of the methods presented in this dissertation.

Special thanks go to Professors Satinder Singh Baveja, Eunshin Byon, and Jun Ni for accepting to be on my dissertation committee. I appreciate their interest in my work and their suggestions that helped me to improve my dissertation. The graduate machine learning class that I took with Professor Baveja as well as his input and questions during our interactions were very inspiring to improve the modeling framework presented in this dissertation. Professor Byon has been exposed to my work since I took the time series modeling class with her and it is a pleasure for me that she agreed to stay on as part of my committee. I also very much appreciate to have Professor Ni serving as a member of my dissertation committee.

I am grateful to all the engineering faculty I interacted with at the University of Michigan, especially Professors Afred Hero, Huei Peng, Sandeep Pradhan, Kazuhiro Saitou, and Zeeshan Syed. I am also indebted to Professors Kurt Barbé, Rik Pintelon, and Johan Schoukens at Vrije Universiteit Brussel (VUB).

I would like to thank clinicians and professors at the University of Michigan Hospital whom I interacted with during my doctoral studies. Special thanks go to Professors Engorin Milo and Ronald Chervin for the helpful feedback and inspiring comments they provided during our frequent meetings. I would like also to acknowledge the help I got from the clinicians at the University of Michigan Sleep Disorders Center. Particulary, I am grateful to Judy Fetterolf who has been a great help in clarifying many clinical aspects about the data sets used in this research.

I owe much appreciation to my fellow colleagues, past and present, in Tilbury Research Group. Many thanks to Dhananjay Anand, John Broderick, and Justin Storms for stimulating discussions and great friendship. Additional thanks go to the research assistants that worked for me – Ge Bian, Anqi Sun, and Jason Teno.

Most importantly, I want to thank my family without whom I would not be where I am today. My parents have made many sacrifices to allow me to reach this point. The love and support they have always provided me is immeasurable. They have taught me to advance myself at every step of life, value integrity, work hard, and demonstrate a concern for others. I am also grateful to my sister Hanan and my brother Zakaria, for always supporting me in my life's endeavors. I owe a heart-felt gratitude for my amazing wife Eshrak, whose support and encouragement have been an immensely valuable asset. Also, my little son Mohammad has been a source of constant amazement and inspiration.

Above all, I thank God for the many blessings He has given me and my family. I submit myself to Him, and seek His mercy. I always pray that He protects me and my loved ones and increases my knowledge so that I can better serve mankind.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

A Framework for Evaluation and Identification of Time Series Models for Multi-Step Ahead Prediction of Physiological Signals

by

Hisham M. W. S. ElMoaqet

Chair: Dawn Tilbury

Significant interest exists in the potential to use continuous physiological monitoring to prevent respiratory complications and death, especially in the postoperative period. Smart alarm-threshold based systems are currently used with hospitalized patients. Alarms are generated from these devices when an adverse event *has occurred* with the value(s) of the monitored physiological variable(s) being outside the range of normal one(s). These systems have been suggested as cost effective, non-invasive monitoring techniques and have been shown to reduce the need for intensive care admissions.

Despite clinical observations and research studies to support benefit from smart alarm monitoring systems, several concerns remain. For example, a small difference in a threshold may significantly increase the alarm rate. A significant increase in alarm related adverse outcomes has been reported by health care oversight organizations. Also, it has been recently shown in many clinical cases that the signaled alarms are indeed late detections for clinical instability leading to a delayed recognition and less successful clinical intervention.

This dissertation advances the state of art by moving from just *monitoring* towards

prediction of physiological variables. Moving in this direction introduces research challenges in many aspects. Although existing literature describes many metrics for characterizing the prediction performance of time series models, these metrics may not be relevant for physiological signals. In these signals, clinicians are often concerned about specific regions of clinical interest typically defined by deviations from the normal behavior or by the occurrence of critical patterns of abnormality. Existing prediction performance metrics that typically describe the error between predicted and actual signals do not provide insight into the performance in predicting regions of clinical interest in these signals. This dissertation develops and implements different types of metrics that can characterize the performance in predicting clinically relevant regions in physiological signals. The developed metrics address research questions that arise when the performance of a time series prediction model of a physiological signal is to be evaluated.

In the era of massive data, biomedical devices are able to collect a large number of synchronized physiological signals recording a significant time history of a patient's physiological state. Directionality between physiological signals and which ones can be used to improve the ability to predict the other ones is an important research question. This dissertation uses a dynamic systems perspective to address this question. Metrics are also defined to characterize the improvement achieved by incorporating additional data into the prediction model of a physiological signal of interest.

Although a rich literature exists on time series prediction models, these models traditionally consider the (absolute or square) error between the predicted and actual time series as an objective for optimization. This dissertation proposes two modeling frameworks for predicting clinical regions of interest in physiological signals. The physiological definition of the clinically relevant regions is incorporated in the model development and used to optimize models with respect to predictions of these regions.

CHAPTER I

Introduction

The objective of this dissertation is to develop a research framework for both evaluating and optimizing predictive models of physiological time series, and to apply these methods to patient data.

1.1 Motivation

Recent developments in biomedical monitoring systems and computing technologies have enabled collecting and recording long time histories of medical signals. As a result, there has been a significant increase in the importance of continuous physiological monitoring in preventing respiratory complications and death, especially in the postoperative period.

Smart alarm-threshold based monitoring is the state of the art technology used with hospitalized patients. These systems provide alarms when an adverse event *has occurred* with the value(s) of the monitored physiological variable(s) outside the range of normal one(s). Using smart alarm-threshold based monitoring systems in general care floor units has reduced the need for intensive care unit admissions (*Taenzer et al.*, 2010). They have been also suggested as cost effective, non-invasive techniques for monitoring high risk patient populations (*McFadden et al.*, 1996).

Despite clinical observations and research studies to support benefit from smart

alarm monitoring systems, several concerns exist. For example, a small difference in a threshold may increase the alarm rate by a factor of 10 (*Watkinson and Tarassenko*, 2012), resulting in excess false positive alarms (if alarm limits are set too high) and harm (if set too low) (*Bliss and Dunn*, 2000). The *Joint Commission* (2013) and other health care oversight organizations have recently reported a significant increase in alarm related adverse outcomes. Consequently, smart monitoring systems are more likely to cause significant alarm loads, resulting in fatigue and inadequate responses. A recent study by *Lynn and Curry* (2011) investigated patterns of in-hospital deaths and discussed cases when the signaled alarms are indeed late detections for clinical instability leading to a delayed recognition and less successful clinical intervention.

The current limitations in the state of the art highlight the need to move from just *monitoring* towards *predicting* physiological signals. Moving in this directions introduces research challenges. The research in this dissertation aims at addressing these challenges through presenting a framework for predicting physiological signals.

Although time series modeling and prediction has been greatly investigated in many research fields, it is still a relatively new research area for physiological time series compared to other science and engineering fields. For example, an extensive literature exists describing time series modeling and prediction methods in manufacturing, economics, finance, metrology, and telecommunications (*Palit and Popovic*, 2006). Existing literature also describes many engineering metrics for evaluating the prediction performance of time series models. A fundamental research question in physiological time series is the performance evaluation for the predictions of these time series. Existing metrics describe the error between the predicted and reference signals without conveying information about their ability to capture regions of clinical interest in physiological signals. The research in this dissertation highlights the shortcomings of existing engineering metrics when used with physiological time series and develops new criteria for evaluating physiological predictions from time series

models.

The ability to record a large number of physiological signals using multi-channel biomedical systems motivates the need to understand the cause-effect relationship between signals. Understanding the directionality between physiological signals is crucial for the development of multi-variable predictive models that incorporate signals directly affecting the evolution of clinically relevant events in physiological signals. The research in this dissertation addresses this directionality from a dynamic systems perspective. Metrics are also defined to characterize the improvement achieved by incorporating additional data into the prediction models of a physiological signal of interest.

Engineering literature extensively describes models and strategies for multi-step ahead prediction of time series. These methods commonly consider the error between the predicted and reference time series as an objective for developing prediction models. Nevertheless, such models are not designed to capture clinically relevant regions in physiological time series. Moving to prediction in place of standard monitoring requires innovative models designed to predict clinical patterns of interest in physiological signals. The research in this dissertation proposes a new paradigm for predicting regions of clinical interest in physiological signals using a novel time series modeling framework.

The ability to develop models and algorithms for robust predictions of physiological signals is a fundamental step towards designing innovative prophylactic interventional systems to prevent occurrence of adverse outcomes. The proactive responses obtained using predictive models can be further utilized to improve patient outcomes compared to existing methods that merely depend on reactive responses. In addition to automatic interventional systems, the developed methods enable (human) advisory control, where control actions are recommended to a patient and/or clinician based on forecasted future values of physiological signals.

1.2 Contributions

The research in this dissertation develops an integrated framework as a first step towards moving to prediction systems of physiological signals in place of currently used monitoring ones. Three fundamental aspects of this problem are addressed in this dissertation. Relevant performance metrics are investigated and developed for evaluating predictions from models of these signals. The interactive dynamical relationship between multi channel data is also incorporated to improve the ability to predict these signals. Finally, novel time-series models are developed and optimized to predict clinically relevant events within these time series.

The contributions of this work can be summarized as follows.

Performance Metrics for Evaluating Predictions of Physiological Signals Standard dynamic models are typically evaluated based on the prediction error (RMSE) between the actual and the predicted signals. Nevertheless, in physiological signals, clinicians often prefer to predict regions of clinical interest in these signals rather than just match them.

This dissertation develops and implements different types of metrics that can characterize the performance of standard time series models in predicting clinically relevant regions. Three different categories of performance metrics were developed and investigated in this dissertation.

- *Performance Evaluation Over Prediction Windows*

  RMSE (as the most commonly used standard prediction performance metric) is not capable of distinguishing errors due to phase difference from errors due to magnitude difference. However, in many clinical applications, it is not as important for the predictive model to predict the exact time instance at which a clinically relevant event is going to happen. Rather, clinicians are more interested in predicting the occurrence of such an event within a specific prediction

widow. Inspired by this application area, this dissertation proposes a performance metric for evaluating predictions of physiological signals over prediction windows. Using a prediction model (or a set of models), one can use the observed time history up to a specific time instance to generate successive multi-step ahead predictions over a window of interest. If a clinically relevant event that occurred over this window is predicted by the model(s), then this is considered a successful window prediction. The proposed metric looks at all possible scenarios resulting from comparing successive window predictions with corresponding reference ones. Evaluating predictions using this metric characterizes the overall ability of the predictive model to capture clinically relevant events in the time series of interest over the desired window length. Additionally, this metric provides the basis for comparing different types of models with respect to predictions of clinical patterns of interest in a physiological time series.

- *Performance Evaluation for the Longest Horizon That Can Predict a Clinically Relevant Event*

  An important factor that needs to be taken into consideration when evaluating predictive capability of dynamical models in a clinical setting is the longest prediction horizon at which clinically relevant events can be captured. Of course, if an adverse event can be predicted far in advance, there will be a better opportunity for clinical intervention to prevent such an event. In contrast, if the predictive model didn't capture the critical event or if it predicted the event just before its occurrence, there might not be enough time for medical intervention.

  Driven by the clinical interest of predicting clinical events as early as possible, this dissertation develops the longest horizon that can predict a clinically relevant event as a metric for evaluating predictive models for physiological time series. Considering the onset time of any clinically relevant event, a predic-

tion model (or a set of models) with varying prediction horizons can be tested starting from the most recent time instance before the onset of this event and going backwards to find the longest horizon that can still predict this event. Since such events typically vary with respect to the longest horizon that can predict them, the analysis of the predictions obtained by any predictive model results in a statistical distribution that quantifies the overall predictive power with respect to this metric.

- *Fixed Horizon Multi-Step Ahead Prediction Metrics*

  Previously introduced metrics consider different prediction horizons either to evaluate window based predictions or find the longest horizon able to predict clinical events. On the other hand, clinicians are also interested in evaluating the physiological predictions with respect to a fixed prediction horizon (k time steps). Thus, a predictive model can be directly evaluated by looking k steps ahead at each time instance and seeing if it captures a clinically relevant event that occurs after k steps. Evaluating predictions for a fixed horizon addresses the effectiveness of the predictions at this horizon as well as their ability to capture the onset of clinically relevant events at this horizon. Clinically relevant events in a physiological signal are often physiologically defined with respect to critical levels of abnormality (exceed or drop below a threshold for a minimum time duration) or with respect to deviations from the operating baseline of the signal. Accordingly, this dissertation addresses this metric in two ways. First, we consider evaluating predictive models for their ability to capture abnormal deviations from the signal baseline. Second, we consider evaluating predictive models for their ability to predict critical levels.

As a case study, the developed metrics are used for characterizing the performance and addressing limitations of standard auto-regressive models developed for predicting critical oxygen saturation levels in the blood. Using the metric of the largest horizon

that can predict critical events, it was shown that 56.2% of the critical desaturation events in the time series of patients were able to be predicted 10-60s ahead of time with the standard auto-regressive models. Although prediction performance shows excellent ability to predict *normal* signal levels over 20s and 60s prediction windows, a significant decrease in the ability to predict *critical* signal levels was noticed over 60s prediction windows compared to 20s windows. Predictions of critical signal levels obtained with a fixed horizon of 20s are associated with a significant phase lag. This contribution is discussed in Chapters IV and V.

Incorporating Interactive Relation Between Physiological Variables into Predictions of Physiological Signals Multi-channel biomedical devices record synchronized physiological signals. Understanding the causality relationship between simultaneously recorded signals can significantly improve the ability to predict them.

This dissertation addresses the cause-effect question between different simultaneously recorded physiological signals from a dynamic systems point of view. Given two physiological signals of interest A and B, we developed one dynamic model in which A is an input to B and another dynamic model in which B is an input to A. Then the significance of the coefficients for the two modeling choices was investigated to understand the directionality between these physiological variables.

Furthermore, to assess the effect of including B into the prediction model of A, we define metrics to characterize the improvement in capturing regions of clinical interest in A with the inclusion of B.

As an application, the cause-effect relationship between pulse rate (PR) and blood oxygenation (SpO_2) (both collected by pulse oximetry monitoring systems) was investigated. No significant improvement was noticed in the ability to predict critical desaturation levels in the blood using PR data. Results indicate oxygen in blood is an effective input to the pulse rate rather than vice versa. This contribution is further

discussed in Chapter IV.

Novel Auto-regressive Models for Predicting Regions of Clinical Interest in Physiological Signals Finally, the research in this dissertation advances the state of art in structured time-series prediction models by developing novel models designed to capture clinically relevant patterns in physiological signals. A dynamic auto-regressive structure with direct prediction strategy is used to build a framework for optimizing time-series models capable of predicting regions of clinical interest in physiological signals. Unlike standard time-series models that minimize the sum of square errors between the predicted and reference signals, the presented modeling framework optimizes *fixed horizon prediction metrics* with respect to predictions of clinically relevant regions in physiological signals. Two broad optimization tools were considered to identify the proposed modeling frameworks.

- *Mixed integer programming (MIP).*

  This framework optimizes auto-regressive models with respect to predictions of abnormal deviations from signal base line over a fixed prediction horizon of k time steps.

- *Quadratic programming (Support Vector Machine Optimization).*

  This framework optimizes auto-regressive models with respect to predictions of critical signal levels over a prediction horizon of k time steps in a physiological time series. The relative paucity of the critical levels of abnormality in a time series was addressed using a cost-sensitive learning scheme together with different statistical measures to address the imbalance in the optimization problem.

Two case studies are considered for this contribution. First, using MIP formulation, an auto-regressive modeling framework was optimized to predict abnormal desaturation levels in the blood over a prediction horizon of 20s. Then, another auto-regressive modeling framework was developed using SVM formulation to predict 20s

ahead critical desaturation levels in the blood. To address the issue of the rarity of critical desaturation events in the second problem, different statistical metrics are considered for optimization. Results of both case studies show a significant improvement in predicting clinically relevant events. Detailed discussion of this contribution is presented in Chapter V.

1.3 Dissertation Outline

The first part of this dissertation addresses different evaluation criteria that can characterize the performance of a standard predictive model in generating near term future predictions that capture clinically relevant patterns in physiological time series. This part also considers multi-channel data using a dynamic systems perspective and defines metrics to characterize the ability to improve predictions of clinically relevant events using multi-variate time series models. The second part addresses designing models that optimize the proposed metrics to develop a modeling framework with an improved ability to capture different types of clinically relevant events in physiological time series.

In Chapter II, the concepts and tools used throughout this dissertation are defined. Prior work in the area of time series prediction models and metrics is presented with a brief review for the physiological signals and the clinical events investigated in the case studies presented in this dissertation.

In Chapter III, a brief description of the medical data sets used in this research is presented with a description of the data channels and signals specifications. Data pre-processing techniques are also presented.

Chapter IV presents two performance metrics for evaluating predictions of physiological signals. We develop the window based prediction performance metric and the metric for the largest horizon that can predict a clinically relevant event. Using a dynamic systems perspective, we investigate the ability to improve predictions

of clinically relevant events in a time series using additional data channels. Subsequently, we present a case study for predicting critical blood oxygenation levels using the developed metrics. We also investigate the ability to improve the performance in predicting critical blood oxygenation levels using pulse rate dynamics. In this chapter, we use standard dynamical models but we show how standard evaluation metrics that have been commonly used in analyzing engineering systems may not be relevant for physiological ones.

Chapter V develops time series models optimized to capture clinically relevant events. First, we start by presenting fixed horizon prediction metrics for characterizing the performance of predictive models in capturing clinically relevant events over a fixed horizon. Then we use these metrics for developing two modeling frameworks capable of predicting two types of clinical events in a physiological time series. First, we use a mixed integer programming framework for identifying auto-regressive models capable of predicting (clinically defined) abnormal deviations from the baseline of a physiological signal. Then, we use a quadratic optimization framework for identifying auto-regressive models capable of predicting physiologically defined critical levels of abnormality in physiological signals.

Chapter VI summarizes the dissertation and research contributions and then outline directions for future work.

CHAPTER II

Background

In this chapter, we review the literature in time series prediction models and metrics for evaluating predictions obtained from these models. Recent literature in physiological signals' prediction is also discussed with a brief review for the physiological signals and the clinical outcomes investigated in the case studies presented in this dissertation.

2.1 Time Series Modeling

In structured time series models, the signal y_{t+1} at time $t+1$ ($t = n, ..., N$, where N denotes the total number of data samples available for modeling) depends on previously observed signals y_{t-i}

$$y_{t+1} = f(y_t....y_{t-n+1}) + e_{t+1} \tag{2.1}$$

where $f(.)$ denotes the functional dependency between past and future observations of time series y, n denotes the model order (or the embedding dimension (*Casdagli et al.*, 1991)), that is the number of past values used to infer future values, and e_{t+1} represents the term that includes modeling error, disturbances, and/or noise. Equation (2.2) doesn't impose any assumptions on the dynamics of the modeling error e_{t+1}

and thus represents the general form of structured time series models. The dynamics of the error residuals at each time step e_{t+1} can be statically independent (white noise) or could be correlated (coloured noise). In the latter case, the noise correlation can be utilized to model $e_{t+1} = a_{t+1} + g(a_t, a_{t-1}, ..., a_{t-m+1})$ where a_{t+1} is the unmodeled part of error (white noise) and $g(.)$ represents the functional dependence between the m most recent modeling errors. In this dissertation, e_{t+1} is assumed to be white noise.

In model identification, we seek to estimate (learn) the best model $\hat{f}$ that can generate the time series data. Assuming a specific model structure, a standard dynamical model is typically identified by minimizing the sum of squares of the one-step ahead prediction error residuals $\|y_{t+1} - \hat{y}_{t+1}\|$ (*Ljung*, 1987) such that

$$\hat{y}_{t+1} = \hat{f}(y_t....y_{t-n+1}) \tag{2.2}$$

where $\hat{y}_{t+1}$ is the one-step ahead prediction and $\hat{f}(.)$ is the learned (one step ahead) dynamical model. The model $\hat{f}(.)$ can also be updated with time as needed to address any changes in the functional dependence between time series measurement in non-stationary signal cases.

The field of time series modeling has been influenced significantly by linear statistical dynamical models such as auto-regressive (AR) models in which the function $f(.)$ is a linear combination of the n most recent observations and auto-regressive with moving average (ARMA) models that assume $g(.)$ to be a moving average of the m most recent error residuals in addition to the AR part of this type of model. Other linear statistical models such as Box-Jenkins (BJ) and Output Error (OE) models have been also used (*Ljung*, 1987). On the other hand, several non-linear time series models were proposed in the literature. In non-linear time series modeling, researchers often seek these models to be both easily interpreted and identified. Examples of non-linear models that were used successfully to fit dynamical models are non-linear auto-regressive dynamic models and block structured non-linear models

such as Hammerstein-Wiener dynamic models (*Pintelon and Schoukens*, 2012; *Ljung*, 1987). Nowadays, Monte Carlo simulation and bootstrapping methods are used in non-linear time series modeling. Being less restrictive about the distribution of the error process causes these methods to be preferred over classical non-linear modeling techniques (*Ben Taieb et al.*, 2012).

Recently, and with the huge advancements in computing technologies, machine learning models have drawn a significant research attention and have established themselves as serious contenders to classical statistical models in the research community of time series modeling (*Ben Taieb et al.*, 2012; *Ahmed et al.*, 2010; *Palit and Popovic*, 2006; *Zhang et al.*, 1998). These models, also called data driven models (*Mitchell*, 1997), are non-linear models which use only historical data to learn the dependency between past and future observations. For example, artificial neural networks have been successfully used to model non-linear time series (*Lapedes and Farber*, 1987; *Werbos*, 1988). Later, other models were proposed and used such as decision trees, support vector regression, and nearest neighbor regression (*Alpaydin*, 2014; *Hastie et al.*, 2005).

2.2 Multi-Step Ahead Time Series Prediction

Time series prediction has been the focus of research in many domains. One step ahead forecasts make use of the current and observed values of a particular variable to estimate its expected value for the next time step following the latest observation. On the other hand, predicting two or more steps ahead is considered a multi-step ahead prediction problem. Unlike one step ahead, multi-step ahead forecasting is more difficult due to various additional complications such as accumulation of errors, reduced accuracy, and increased uncertainty (*Weigend*, 1994; *Sorjamaa et al.*, 2007; *Ben Taieb et al.*, 2012).

The most commonly used strategies for multi-step ahead prediction rely either on

iterated (recursive) or direct strategies (*Atiya et al.*, 1999; *Chevillon*, 2007; *Sorjamaa et al.*, 2007; *Bao et al.*, 2014b). Sections 2.2.1 and 2.2.2 review these prediction strategies and Section 2.2.3 summarizes other strategies that have been less commonly used in the literature of time series prediction models.

2.2.1 Recursive Prediction Strategy

The prediction model in the iterated (recursive) prediction strategy is constructed by means of minimizing the one-step ahead prediction error residuals. One step ahead predictions are then fed back recursively as inputs to the (same) model for obtaining successive multi-step ahead predictions (*Cheng et al.*, 2006; *Hamzaçebi et al.*, 2009; *Sorjamaa et al.*, 2007; *Kline and Zhang*, 2004; *Tiao and Tsay*, 1994). Considering a prediction horizon of K-step ahead, the recursive prediction strategy can be expressed by the following equation

$$\hat{y}_{N+k} = \begin{cases} \hat{f}(y_N, ..., y_{N-n+1}), & \text{if } k = 1. \\ \hat{f}(\hat{y}_{N+k-2}, ..., \hat{y}_{N+1}, y_N, ..., y_{N-n+1}), & \text{if } 1 < k \leq n. \\ \hat{f}(\hat{y}_{N+k-2}, ..., \hat{y}_{N+k-n}), & \text{if } n < k \leq K. \end{cases} \tag{2.3}$$

The use of previously predicted values in successive predictions causes an error accumulation problem leading to poor prediction performance over long horizons (*Chevillon*, 2007; *Ing*, 2003; *Bao et al.*, 2014a). This shortcoming can be seen more clearly when the prediction horizon K exceeds the model order n causing all prediction inputs to be forecasted values instead of actual observations.

2.2.2 Direct Prediction Strategy

The direct strategy constructs a individual predictive model for any future time step of interest (*Cheng et al.*, 2006; *Hamzaçebi et al.*, 2009; *Sorjamaa et al.*, 2007;

Kline and Zhang, 2004; *Tiao and Tsay*, 1994). Considering a prediction horizon of K-step ahead, the direct prediction model can be expressed by

$$\hat{y}_{N+K} = \hat{f}_K(y_t, ...y_{t-n+1}) \tag{2.4}$$

where $\hat{y}_{N+K}$ is obtained using the learned model $\hat{f}_K$ that uses the n most recent observations to directly compute the K-step ahead predicted values. The model $\hat{f}_K$ is typically learned by minimizing the K-step ahead error residuals. In case all time steps in the prediction window $1 \leq k \leq K$ are of interest, the direct prediction strategy constructs a specific model $\hat{f}_k$ for each prediction time step. This strategy does not use approximate values to compute the forecasts, being more immune to the accumulation of error. Nerveless, it demands larger computational times compared to recursive strategy in order to generate simultaneous predictions of large windows (*Ben Taieb et al.*, 2012).

2.2.3 Other Prediction Strategies

In addition to recursive and direct prediction, other strategies have been proposed but less frequently used. In this section, we summarize three more strategies that have been popularly discussed in the recent literature of multi-step ahead prediction.

DirRec Strategy The DirRec strategy (*Sorjamaa and Lendasse*, 2006) combines the architectures and the principles underlying the direct and the recursive strategies. DirRec computes the forecasts with different models for every horizon (like the direct strategy) and, at each time step, it enlarges the set of inputs by adding variables corresponding to the forecasts of the previous step (like the recursive strategy). Unlike the two previous strategies, the model order (embedding dimension) n is not the same for all the horizons. For a K-step ahead prediction horizon, the DiRec prediction

strategy is given by

$$\hat{y}_{N+k} = \begin{cases} \hat{f}(y_N, ..., y_{N-n+1}), & \text{if } k = 1. \\ \hat{f}(\hat{y}_{N+k-1}, ..., \hat{y}_{N+1}, y_N, ..., y_{N-n+1}), & \text{if } 1 > k \geq K. \end{cases} \tag{2.5}$$

where $\hat{f}_k$ are K learned models for each prediction time step $\in \{1, ..., K\}$. This method has shown very good prediction performance in some time series applications (*Sorjamaa and Lendasse*, 2006). However, not much research has been done with this strategy.

MIMO Strategy The previously discussed strategies are single output strategies (*Ben Taieb et al.*, 2010) since they consider the prediction models as (multiple-input) single-output functions. Recently, *Bontempi* (2008) introduced a multi-input multi-output (MIMO) strategy for multi-step ahead predictions for the goal of preserving the stochastic dependency between the predicted values within a prediction horizon. The forecasts over a prediction horizon K are returned in one step by a multiple-output model $\hat{F}$ as expressed in the following equation.

$$[\hat{y}_{N+k}, ..., \hat{y}_{N+1}] = \hat{F}(y_N, ..., y_{N-n+1}) \tag{2.6}$$

where $\hat{F} : \mathbb{R}^n \to \mathbb{R}^K$ is a vector valued function (*Micchelli and Pontil*, 2005).

DIRMO Strategy The need to preserve the stochastic dependencies between predictions within the same prediction window using one model has a drawback as it constrains all the horizons to be forecasted with the same model structure. This limitation potentially reduces the flexibility of the MIMO forecasting approach especially with long prediction horizons (*Ben Taieb et al.*, 2009).

Accordingly, the DIRMO strategy (*Ben Taieb et al.*, 2009) was proposed to preserve the most appealing aspect of both the direct and MIMo strategies through

forecasting the all the time steps within a prediction horizon K in blocks such that each block is forecasted in a MIMO fashion. Thus, the K-step ahead forecasting task is decomposed into m Multiple Output forecasting tasks $m = \frac{K}{s}$ each with an output of size s ($s \in \{1, .., K\}$).

When $s = 1$, the number of forecasting tasks $m = K$ which corresponds to the direct strategy. When $s = K$, the number of forecasting tasks $m = 1$ which corresponds to the MIMO strategy. There are intermediate configurations between these two extremes depending on the value of a parameter s. Tuning s provides a trade off between preserving larger stochastic dependency between future values and having a greater flexibility of the predictor.

The K step ahead forecasts in DIRMO strategy are generated by using m learned models $\hat{F}_p$, $p \in \{1, .., m\}$ as shown in the following equation

$$[\hat{y}_{N+p\times s}, ..., \hat{y}_{N+(p-1)\times s+1}] = \hat{F}_p(y_N, ..., y_N - n + 1) \tag{2.7}$$

where $\hat{F}_p : \mathbb{R}^n \to \mathbb{R}^s$ is a vector valued function for $s > 1$. Recent research with this method used a particle swarm optimization to tune s with a multi output support vector regression (SVR) model for multi-step ahead time series prediction (*Bao et al.*, 2014b).

Although many methods and strategies have been developed for multi-step ahead predictions, all these methods consider the error between the predictions and reference values as an objective function to obtain prediction models.

2.3 Prediction of Physiological Signals

One well-studied physiological signal is blood glucose. Diabetes has been a challenging topic for feedback control approaches for many years through building different types of models from continuous glucose monitoring (CGM) data. Several data

driven techniques have been applied for the prediction and control of glucose concentrations. For example, a radial-basis function neural network has been used to develop a predictive model for glucose levels (*Trajanoski et al.*, 1998). Also, a Kalman filter was employed to adjust the parameters of a first-principles model for the prediction and control of blood glucose (*Dua et al.*, 2006).

Several studies considered data driven auto-regressive (AR) models to build predictive models (*Sparacino et al.*, 2007; *Reifman et al.*, 2007; *Gani et al.*, 2009) for many desirable properties. Among other reasons, these models have a limited number of parameters to identify, can be fitted from relatively short data records (compared to more complex data-driven models), are computationally efficient, and are suitable for online and recursive applications. A first-order AR model, AR-1, was used after preprocessing (smoothing) the raw CGM data to remove high-frequency noise (*Sparacino et al.*, 2007). A tenth order AR Model, AR-10, was used to model CGM signals (*Reifman et al.*, 2007). The AR coefficients were identified via regularized least squares using raw (unsmoothed) signals. More recently, a thirtieth order model, AR-30, has been proposed to predict near future glucose concentrations from smoothed data (*Gani et al.*, 2009). However, while biomedical signals are becoming more readily available, not much research work related to identifying dynamic models from signals besides glucose has been conducted. Recently, a twentieth-order AR model, AR-20, has been used to model the photoplethysmogram (PLETH) signal (*Lee et al.*, 2011). The respiratory rate of patients was estimated through analyzing frequency domain characteristics of the AR-20 model. Nevertheless, this model was not used in multi-step ahead prediction of future patterns of the PLETH time series.

2.4 Pulse Oximetry Monitoring (POM)

POM devices are frequently used in general care units to record both noninvasive oxygen saturation levels in blood (SpO_2) and pulse rate (PR) at 1-2 second intervals.

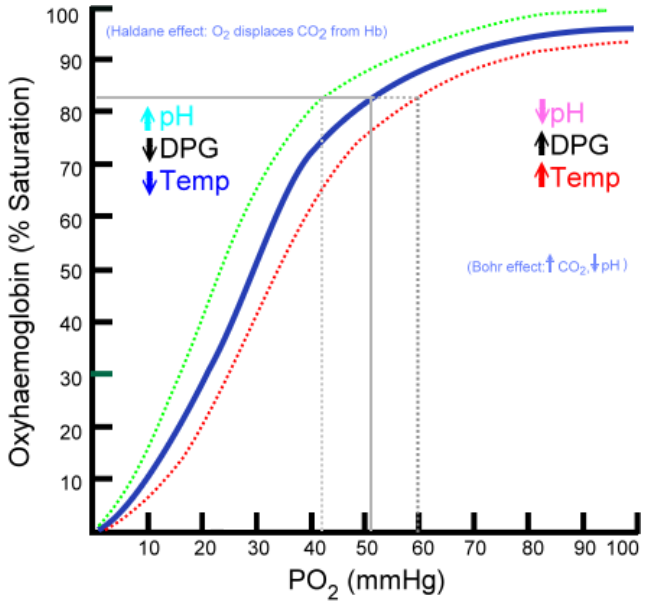


Figure 2.1: Hemoglobin saturation curve showing the sigmoid relationship between SpO_2 and pressure of oxygen in blood pO_2. At an SpO_2 of 90%, the curve starts becoming steeper, such that a small decrease in the partial pressure of oxygen results in a large decrease in the oxygen content of the blood, Figure from (*West*, 2012).

These devices have been suggested as cost effective, non-invasive techniques to reduce the risk of hypoxemic injury in high risk populations (*Eichhorn*, 2003) and to reduce costly critical unit (CCU) admissions (*McFadden et al.*, 1996; *Taenzer et al.*, 2010).

The human body requires and regulates a very precise and specific balance of oxygen in the blood to ensure sufficient oxygen delivery to tissues, primarily through binding with hemoglobin. The oxy-hemoglobin binding is characterized by cooperativity, that allows for more oxygen to be extracted by tissues that are suffering from low oxygen levels. The cooperativity between oxygen and hemoglobin molecules results in a sigmoid-shaped dissociation curve (*Mason et al.*, 2010) shown in Figure 2.1. At an SpO_2 of 90%, the curve starts becoming steeper, such that a small decrease in the partial pressure of oxygen results in a large decrease in the oxygen content of the blood (*West*, 2012). Consequently, a sudden severe reduction in SpO_2 is clinically considered to be potentially life-threatening (*Stradling and Crosby*, 1991; *Rauscher*

et al., 1991; *Kripke et al.*, 1997; *Mooe et al.*, 1996; *Epstein and Dorlac*, 1998). A threshold of $SpO_2 \leq 89\%$ has been identified as a significant marker of the need for nocturnal oxygen treatment (*Centers for Medicare and Medicaid Services*, 1993).

An Oxygen Desaturation Index of 4 percent (ODI_4) is clinically defined as the number of episodes per hour of oxygen desaturations that are greater than or equal to 4% from the base line of the signal (*Berry et al.*, 2012). The SpO_2 signals of concerned patients can be also classified according to the frequency of occurrence of these desaturation events. It has been shown that patients with an $ODI_4 \geq 5$ have a significantly higher rate of postoperative complications than those with $ODI_4 < 5$ and that the complication rate also increased with increasing ODI_4 values (*Hwang et al.*, 2008). No previous studies were found on identifying predictive models from SpO_2 data.

2.5 Multi Channel Data

One of the important questions that arises with physiological signals is the causality relationship between the signals and which ones are most likely to affect which others significantly. The availability of multi-channel biomedical devices allows synchronized physiological signals to be recorded and facilitates investigating the interactive relationship between them. For the SpO_2 and PR signals collected by POM systems, clinical observations showed that episodic drops in blood oxygen levels have direct impact on the autonomic nervous system control of the pulse rate (*Somers et al.*, 1995). Previous studies have shown transient but reproducible increases in pulse rate with episodic drops in blood oxygen, as is seen where patients suffering from sleep apnea are exposed to years of such changes, resulting in faster baseline pulse rates (*Vries et al.*, 2008; *Schulte-Frohlinde et al.*, 2002; *Griffin et al.*, 2005; *Williams and Galerneau*, 2003; *Tsuji et al.*, 1994; *Norris et al.*, 2008; *Riordan et al.*, 2009). Up to the best of our knowledge there are no physiological based models that

describe the directionality (cause-effect relationship) between SpO_2 and PR.

The high number of physiological variables collected in intensive and general care units reflects the challenge of proper extraction and interpretation of the information contained in this flood of information. Previous studies in this field have focused on statistical analysis with some efforts to use graphical methods for understanding correlations between variables (*Imhoff et al.*, 2002; *Gather et al.*, 2002). Yet, not much research has focused on dynamic model identification to quantify the cause-effect relations between these variables. Indeed, the ability to model and analyze the relationships between different physiological signals could be very useful in predicting future behavior of patient's state.

2.6 Evaluating Time Series Predictions

Although there are various methods to obtain multi-step ahead prediction models from time series data, the performance of these models in forecasting is crucial for the ability to utilize them practically. In physiological time series, root mean square error (a standard performance metric) has been used frequently to evaluate predictions of physiological systems (*Sparacino et al.*, 2007; *Reifman et al.*, 2007; *Gani et al.*, 2009). Nevertheless, clinicians are often more interested in predicting clinically relevant events in physiological signals rather than closely matching them. For example, the relative rarity of critical patterns of abnormality in a physiological signal makes it less likely for a multi-step ahead prediction model (based on minimizing the mean squared error) to capture these clinically relevant regions. In such cases, the prediction might miss these events or could capture them but with a significant phase lag (time delay). Unlike other time series applications, the phase lag of multi-step ahead predictions in physiological signals is very important to consider (*Sparacino et al.*, 2007; *Gani et al.*, 2009). Indeed, clinicians are always interested in improving the ability to predict the onset of these events since they typically require a clinical ac-

tion. For long prediction horizons, a small phase lag in capturing these events might be tolerable but a larger one will significantly limit the efficiency of the prediction.

One example of an evaluation metric well known in diabetes management is the Clarke-Error Grid (Clarke-EGA) (*Clarke*, 2005). The grid, shown in Figure 2.2, maps actual and estimated glucose concentrations into five zones, where Regions A and B are clinically acceptable, Region C may result in unnecessary corrections, Region D could lead to incorrect treatments, and Region E represents erroneous treatment. Recently, this grid has been used to assess the accuracy of time-series modeling methodologies developed to predict glucose levels using continuous glucose monitoring (CGM) data (*Gani et al.*, 2010). The grid illustrates the fact that it is not necessary to predict the blood glucose accurately on an absolute scale, but rather it is important to know whether the blood glucose will be in a normal region, or too high (hyperglycemic) or too low (hypoglycemic).

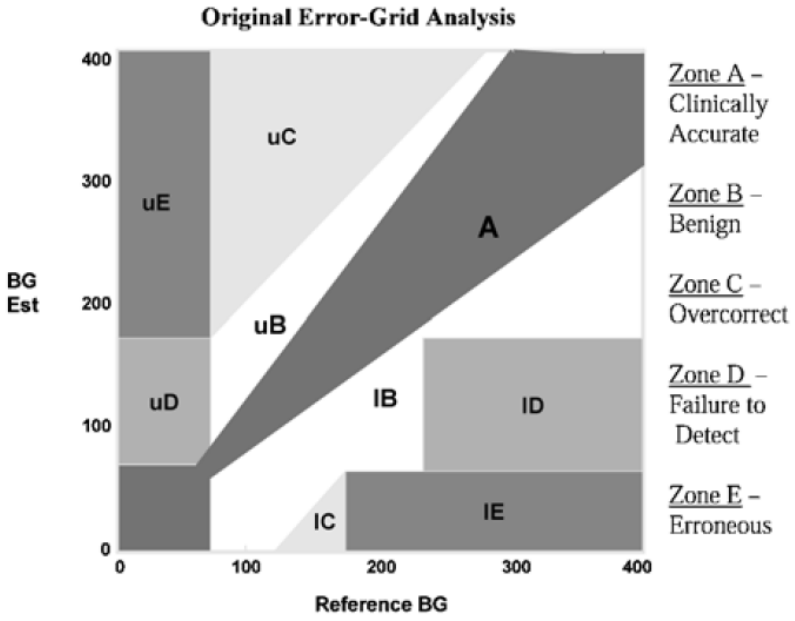


Figure 2.2: Clark Error Grid. The grid maps actual and estimated glucose concentrations into five zones, where Regions A and B are clinically acceptable, Region C may result in unnecessary corrections, Region D could lead to incorrect treatments, and Region E represents erroneous treatment, Figure from (*Clarke*, 2005).

In other time series applications, many evaluation metrics have been utilized based on the sum of (square or absolute) errors between the predictions and reference signal values (*Ben Taieb et al.*, 2012; *Bao et al.*, 2014a; *Xiong et al.*, 2013). In financial time series, the directional symmetry was used as an additional measure for evaluating the accuracy of the predictions (*Yu et al.*, 2008; *Xiong et al.*, 2013). This metric is a statistical measure of a model's performance in predicting the direction of change, positive or negative, of a time series from one time step to the next. Although this metric provides information about the qualitative behavior of one-step ahead predictions, our interest in physiological signals is beyond one step ahead predictions. In fact, we are more concerned about the ability to capture clinically relevant events over longer prediction horizons. Even if the multi-step ahead prediction didn't match the direction of change of the reference signals, it can be clinically acceptable if both are within the same clinical region. On the other hand, the multi-step prediction can match the direction of change in the reference signal but still not be able to capture a clinically relevant event.

CHAPTER III

Data Sets and Processing

This section introduces the data sets used in the research along with signal preprocessing techniques that were considered for different signals in these data sets.

3.1 Postoperative Adults

The data was collected from 119 postoperative adults following orthopedic surgery over a 3 month period in 2009. All patients were placed on postoperative POM (MASIMO RAD-8, Irvine CA) on arrival to the patient care unit per institutional policy. Immediately after termination of the patients' monitoring period, device ASCII data consisting of SpO_2 and PR were downloaded using PROFOX Oximetry Software (version PO Standard; Escondido, CA). The ODI_4 values for each patient were also computed by the PROFOX oximetry software.

This data collection was approved by the Institutional Review Board (IRB) at the University of Michigan (IRB#HUM00069035). The data sets were collected as part of a Quality Improvement study (IRB#HUM00027189) looking at the reliability and nurse response times to a postoperative oximetry based paging alert system (*Voepel-Lewis et al.*, 2013). The data obtained for both PR and SpO_2 are discrete time signals sampled every 2 seconds and quantized for a resolution of $\pm 1\%$ for SpO_2 and ± 1 bpm for PR. ODI_4 values range for the 119 patients from 0.1 to 45.9.

3.2 Data Pre-processing

3.2.1 Missing Measurements Imputation in POM signals

The raw POM signal of each patient includes some instances where the measurements are of zero amplitude. Upon inspection, the missing measurement were noticed to be independent of the observed measurements that lead to these missing values. Hence, these missing measurements were treated as missing completely at random (MCR) (*Polit and Beck*, 2008; *Baraldi and Enders*, 2010).

Zeros that extend for a duration no more than 6 sampling steps (12 seconds) were considered anomalous sensor measurements and were replaced with the most recent non-zero amplitude, similar to the zero-order-hold (ZOH) principle (*Astrom and Wittenmark*, 1997). Zeros that last for more than 6 sampling steps break the continuity of the time series and require the corresponding POM signal to be partitioned to smaller pieces that can be modeled and analyzed separately. Also, there are some instances where the PR signals are of amplitude less than 30 bpm which is clinically impossible. These false measurements were treated as missing measurements in the same way as the zero measurements.

3.2.2 POM Signals Smoothing

We considered two smoothing methods (*Sparacino et al.*, 2007, 2010; *Gani et al.*, 2009).

3.2.2.1 Moving average filters

The Simple Moving Average (SMA) filter averages a number of points from the input signal to produce each point in the output smoothed signal (*Sparacino et al.*, 2010)

$$\tilde{y}_i = \frac{1}{M} \sum_{n=0}^{M-1} y_{i-n} \tag{3.1}$$

where $\tilde{y}$ is the smooth signal, y is the raw data signal, and M is the number of points used in the moving average. The Exponentially Weighted Moving Average (EWMA) filter applies weighting factors that decrease exponentially. The weighting for each older datum point never reaches zero. The EWMA can be calculated recursively as

$$\tilde{y}_1 = y_1 \tag{3.2}$$

$$\tilde{y}_i = \alpha y_i + (1-\alpha)\tilde{y}_{i-1} \quad , i > 1 \tag{3.3}$$

where $\tilde{y}$ is the smooth signal, y is the raw data signal, and α is a constant smoothing factor between 0 and 1. A higher α discounts older observations faster.

3.2.2.2 Regularization

Regularization has been proposed previously to smooth CGM signals (*Gani et al.*, 2009). We used Tikhnov regularization approach, which yields smoothed signals $\tilde{y}$ by computing $\tilde{y} = U_d w$, where U_d denotes the integral operator and w denotes estimates of the raw signals' first derivatives. The derivatives' estimates yield excellent data smoothing and do not introduce lag on the smooth signal relative to the original raw signal.

To estimate the signal's derivatives w, we minimized the functional $f(w)$, given by

$$f(w) = \|y - U_d w\|^2 + \lambda_d^2 \|L_d w\|^2 \tag{3.4}$$

where y denotes the $N \times 1$ vector of the raw time series signal, U_d denotes the $N \times N$ integral operator, w represents the $N \times 1$ vector of first order differences (the rate of change in raw measurements), λ_d represents the data regularization parameter, and L_d denotes a well-conditioned matrix chosen to impose smoothness constraints on the derivative of the raw signal.

For a chosen L_d, the quality of smoothing is determined by λ_d. When $\lambda_d = 0$, no

regularization is performed resulting in the original raw data y. As λ_d increases, the solution w (and hence $\tilde{y}$) increasingly satisfies the imposed smoothness constraint, resulting at the same time in larger deviations from the raw data.

Figure 3.1 shows smoothed SpO_2 signals for 6-minutes data of Patient No. 12. For the SMA filter shown in dotted red line, $M = 5$, for EWMA filter shown in dash-dot blue line $\alpha = 0.35$, and for the regularized signal shown in dashed green line $\lambda_d = 20$.

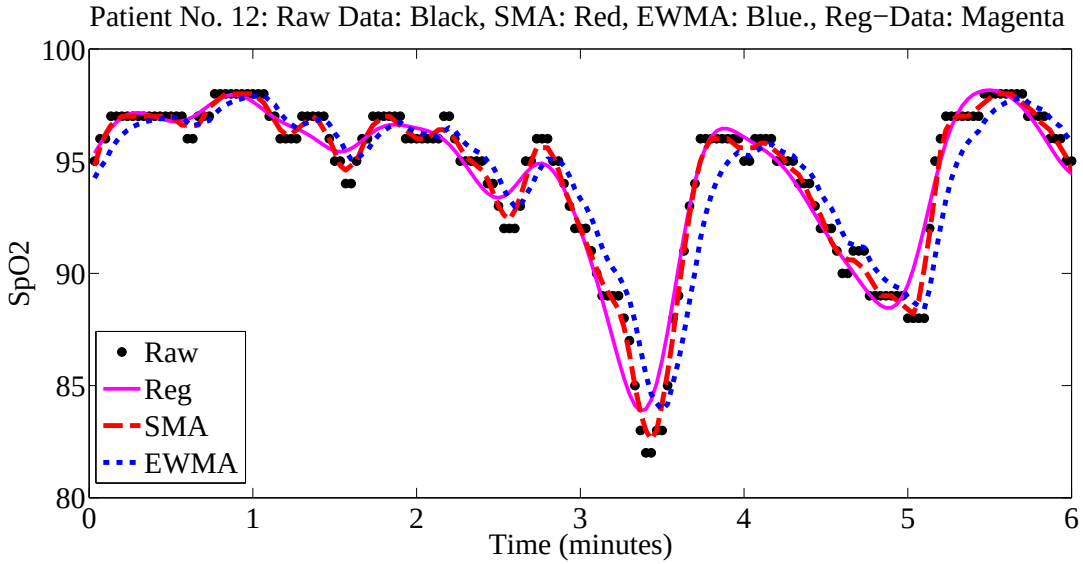


Figure 3.1: Raw data and smoothed data for the SpO_2 signal of a postoperative patient (Patient No. 12). Raw data: black, simple moving average: red, exponentially weighted moving average: blue, regularized data: magenta. Regularization yields excellent data smoothing and do not introduce lag on the smooth signal relative to the original raw signal.

CHAPTER IV

Characterizing Performance of Predictive Models for Physiological Signals

The work presented in this chapter appears in proceedings of the *2013 American Control Conference (ACC)* (*ElMoaqet et al.*, 2013b) and the *Journal of Physiological Measurement* (*ElMoaqet et al.*, 2014a).

4.1 Introduction

In physiological time series, clinicians are often interested in evaluating predictive models for their ability to capture critical levels of abnormality more than their ability to exactly match reference signals. Sections 2.1 and 2.2 reviewed modeling and prediction techniques that have been proposed in time series literature. Nevertheless, these techniques have been traditionally evaluated for the (absolute or square) error between the prediction and reference signals.

This chapter proposes methods for evaluating predictions of clinically relevant events in physiological signals. The main contribution of this chapter is defining new performance metrics for evaluating predictions of existing (standard) time series models with respect to their ability to predict regions of clinical interest in physiological time series. We present two performance metrics for evaluating predictions. First,

we develop a window based performance metric for evaluating predictions of critical levels of abnormality within a physiological signal. Second, we present another metric that characterizes the longest horizon at which a clinically relevant event can be predicted. Subsequently, we use a dynamic systems perspective to investigate the ability to improve predictions of clinically relevant events in a time series using additional data channels. As a case study, the developed metrics are used for evaluating predictions of critical blood oxygenation levels in SpO_2 time series. Then, we incorporate the pulse rate signal into the SpO_2 prediction process. In this chapter, we use standard auto-regressive dynamical models but we address the shortcomings of standard evaluation metrics.

4.2 Evaluating Predictions of Critical Levels of Abnormality in Physiological Signals

In this section, we consider evaluating dynamical models of physiological signals for their ability to predict critical levels of abnormality (exceeding or dropping below a critical signal threshold y_{cr}). In some clinical cases, the physiological event of interest is defined by breaching y_{cr} for a minimum duration Δ time steps. We will define the proposed metric for the general case of $\Delta > 1$ time step.

First, we define a utility function *Score*$(S, y_{cr}, \Delta,$*Flag)* in Algorithm 1. The input arguments for this function are a physiological time series segment S, the corresponding critical signal level y_{cr}, the minimum duration (Δ time steps), and a flag to specify whether critical low or high signal levels are of interest. The function returns the *clinically relevant events (defined by start and end times)* in the segment S or *Null* if no clinically relevant events are present in this segment.

Consider a time series y in the interval $[t_i, t_f]$ with an arbitrary $t \in [t_i, ..., t_f]$ and a prediction window of K steps corresponding to time KT_s, where T_s is the sampling

Algorithm 1 Function for scoring clinically relevant events in physiological time series

Data: Time Series Segment S
Input: y_{cr}, Critical signal level
Input: Δ Minimum duration that defines a clinical event
Input: Flag $\in$ {'Critical Low', 'Critical High'}

```
function SCORE(S, y_cr , Δ, Flag)
    If Flag= 'Critical Low' Then
        [ClinicalEvents]=Find([S ≤ y_cr] ≥ Δ)
    Else IF Flag= 'Critical High' Then
        [ClinicalEvents]=Find([S ≥ y_cr] ≥ Δ )
    end If
    If NotEmpty([ClinicalEvents]) Then
        return [ClinicalEvents]
    Else
        return Null
    end If
end function
```

interval of the discrete signal y. Assuming a dynamic model $\mathcal{M}$ with a multi-step ahead prediction strategy, one can obtain the set of predictions $\hat{y}_{t+k}$ where $1 \leq k \leq K$. A prediction vector PV_t can then be constructed as shown in Equation (4.1).

$$PV_t = \begin{bmatrix} \hat{y}_{t+1} & \hat{y}_{t+2} & \dots & \hat{y}_{t+K} \end{bmatrix} \tag{4.1}$$

On the other hand, the reference vector (RV_t) over the same window can be expressed by Equation (4.2).

$$RV_t = \begin{bmatrix} y_{t+1} & y_{t+2} & \dots & y_{t+K} \end{bmatrix} \tag{4.2}$$

The vectors PV_t and RV_t are then passed to the *Score* function. The prediction at time t is evaluated by comparing comparing PV_t and RV_t with respect to capturing the critical signal level y_{cr} for a minimum duration of Δ time steps. Repeating this procedure for all time instances $t \in [t_i, t_f]$ develops the proposed evaluation metric summarized in Algorithm 2.

The prediction grid used for assessing the quality of the prediction process is

Algorithm 2 Evaluating predictions of critical signal levels over prediction windows

Data: Time Series y_t, $t \in \{t_i, ..., t_f\}$
Input: Predictive Model $\mathcal{M}$
Input: K, Prediction horizon (window length)
Input: y_{cr} Critical signal threshold
Input: Δ Minimum duration that defines a clinical event
Input: Flag $\in$ {'Critical Low', Critical High'}

while $t \in \{t_i, ..., t_f\}$ **do**
    for $k = 1 : 1 : K$ **do**
        compute $\hat{y}_{t+k}$
    end for
    $PV_t = \begin{bmatrix} \hat{y}_{t+1} & \hat{y}_{t+2} & \dots & \hat{y}_{t+K} \end{bmatrix}$
    $RV_t = \begin{bmatrix} y_{t+1} & y_{t+2} & \dots & y_{t+K} \end{bmatrix}$
    [PredEvent]=Score($PV_t, y_{cr}, \Delta,$ Flag)
    [RefEventt]=Score($RV_t, y_{cr}, \Delta,$ Flag)
    If NotEmpty([PredEvent]) & NotEmpty([RefEvent]) **Then**
        t maps to Region **A**
    Else If NotEmpty([PredEvent]) & Empty([RefEvent]) **Then**
        t maps to Region **B**
    Else If Empty([PredEvent] & NotEmpty([RefEvent]) **Then**
        t maps to Region **C**
    Else
        t maps to Region **D**
    end If
    $t = t + 1$
end while

shown in Table 4.2. Pred.= +1 denotes a window prediction of a clinically relevant event (NotEmpty([PredEvent])in Algorithm 2) and Pred.= −1 denotes a window prediction of normal signal levels. The same analogy applies to Ref.= +1 and −1 using the reference signal.

Table 4.1: Prediction Grid for proposed metric

	Ref.= +1	Ref.= −1
Pred.= +1	A	B
Pred.= −1	C	D

Points in Regions A and D on the main diagonal represent points of good prediction, points in Region C off-diagonal represent the ones at which the model fails to detect critical events, and points in Region B off-diagonal represent false prediction points.

Figure 4.1 shows an example illustrating the different regions of the prediction grid. The prediction of blood oxygenation signal SpO_2 over a prediction window of 60s is considered in this example. The critical low threshold for this signal is $y_{cr} = 89\%$. Note that a critical desaturation event starts as soon as $y \leq y_{cr} = 89\%$ ($\Delta = 1$ time step) and continues until the signal recovers to a level higher than y_{cr} (*Oliver and Flores-Mangas*, 2006).

The proposed grid can be considered as a statistical binary classification function (*Altman and Bland*, 1994a,b) with sensitivity, specificity, positive predictive value, negative predictive value, and accuracy used as statistical measures to evaluate its performance. We assign Regions A, B, C, D as the true positive, false positive, false negative, and true negative areas respectively. Mathematically, this can be expressed

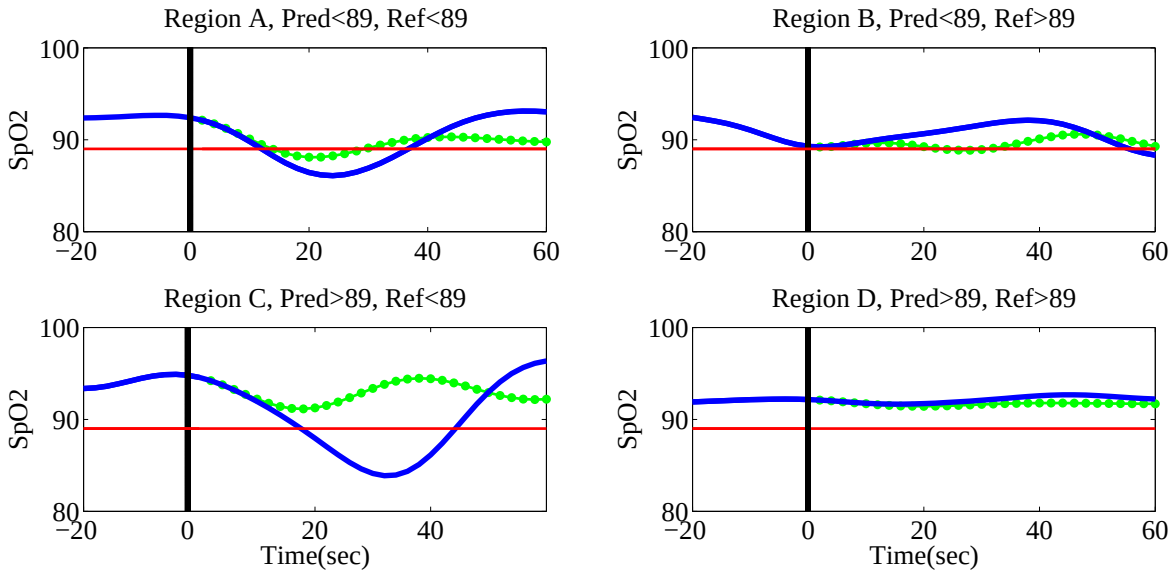


Figure 4.1: Examples of prediction grid regions, blue: actual, green-dotted: prediction. The grid can considered a binary classification function (*Altman and Bland*, 1994a).

as follows.

$$\text{Prediction Sensitivity (TPR)} = \frac{|A|}{|A|+|C|} \times 100\% \tag{4.3}$$

$$\text{Prediction Specificity (TNR)} = \frac{|D|}{|B|+|D|} \times 100\% \tag{4.4}$$

$$\text{Positive Predictive Value (PPV)} = \frac{|A|}{|A|+|B|} \times 100\% \tag{4.5}$$

$$\text{Negative Predictive Value (NPV)} = \frac{|D|}{|C|+|D|} \times 100\% \tag{4.6}$$

$$\text{Accuracy (ACC)} = \frac{|A|+|D|}{|A|+|B|+|C|+|D|} \times 100\% \tag{4.7}$$

where $|A|$, $|B|$, $|C|$, and $|D|$ are the the number of points in Regions A, B, C, and D.

Using Equations (4.3-4.7), we can characterize the performance of a time series model in predicting critical levels of abnormality within any physiological time series of interest.

4.3 Evaluating Predictions for the Longest Horizon that Can Predict Clinically Relevant Events

Another way to look at the predictive capability of a model of a physiological signal is to evaluate how early an adverse event can be predicted.

Considering the onset time of any clinically relevant in a time series, a predictive model (or a set of models) with varying prediction horizons can be tested starting from the most recent time instance before the onset of this event and going backwards to find the longest prediction horizon that still captures this event. Figure 4.2 illustrates using this metric for evaluating how early a critical event starting at time t_s can be predicted. The set of prediction horizons are $1 \leq k \leq k_m$, where k_m is the maximum tested prediction horizon, are tested iteratively to find the longest one that is able to predict the critical events that starts at time t_s.

This metric can be used to test all critical events within a time series as summarized in Algorithm 3. Using the *Score* function described earlier, the start time (t_s) and the end time (t_e) for the events at the which y_{cr} is breached for a minimum time of Δ time steps can be found. For each test horizon k, the dynamic model $\mathcal{M}$ (of order n) uses the n most recent observations up to time $t = t_s - k$ to generate the k-steps ahead prediction $\hat{y}_{ts}$. To reduce signal noise effects, the proposed approach evaluates the model at each test prediction horizon for continuous accurate predictions of critical signal levels within each event of interest.

The analysis of the predictions obtained by any predictive model with this metric results in a statistical distribution that quantifies the overall predictive power of this model. As an example, Figure 4.3 shows a histogram of the largest prediction horizon that can predict critical desaturation events within the SpO_2 time series of a postoperative patient using an auto-regressive model of order 10 (AR-10). 20 critical events (57.1% of the total number of events) were able to be predicted 20 - 60 seconds

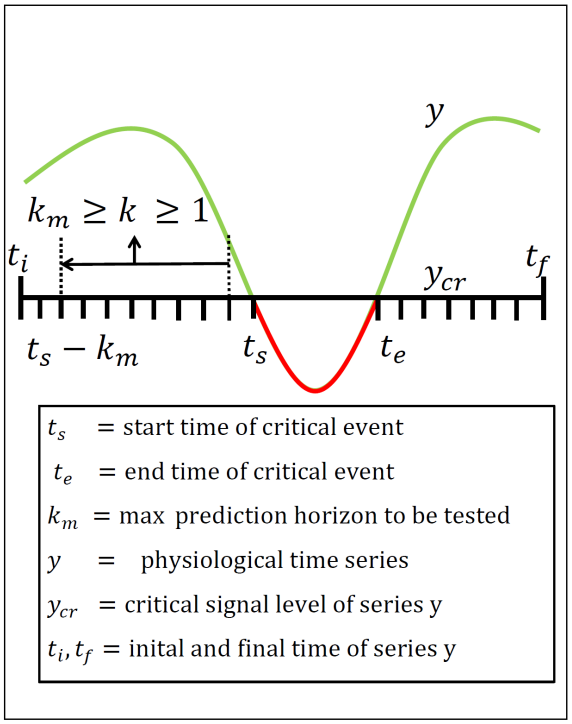


Figure 4.2: Illustration for testing predictive models for the largest horizon that can predict a critical event. After identifying the start time of the critical event, we test iteratively a set of prediction horizons $1 \leq k \leq k_{max}$ to find the largest prediction horizon that can predict the critical event.

Algorithm 3 Evaluating predictions of the longest horizon that can predict a critical signal level

Data: Time Series y_t, $t \in \{t_i, ..., t_f\}$
Input: Predictive Model $\mathcal{M}$
Input: k_{max}, Max test prediction horizon
Input: y_{cr} Critical signal threshold
Input: Δ Minimum duration that defines a clinical event
Input: Flag $\in$ {'Critical Low', Critical High'}

[ClinicalEvents]=Score(y, y_{cr}, Δ, Flag)
[NumEvents]=size([ClinicalEvents])
for $i = 1 : 1 :$ NumEvents **do**
    t_s: start time of event i
    for $k = k_{max} : -1 : 1$ **do**
        $\hat{y}_{ts}$=Predict($[y_{ts-n-k+1}, ..., y_{ts-k}]$, $\mathcal{M}$, k)
        $\hat{y}_{ts+1}$=Predict($[y_{ts-n-k+2}, ..., y_{ts-k+1}]$, $\mathcal{M}$, k)
        ...
        $\hat{y}_{ts+\Delta-1}$=Predict($[y_{ts-n-k+\Delta}, ..., y_{ts-k+\Delta-1}]$, $\mathcal{M}$, k)
        If $\{\hat{y}_{ts}, ...\hat{y}_{ts+\Delta-1}\}$ *breaches* y_{cr}
            return k
            break
        end If
    end for
end for

ahead of time using this type of model.

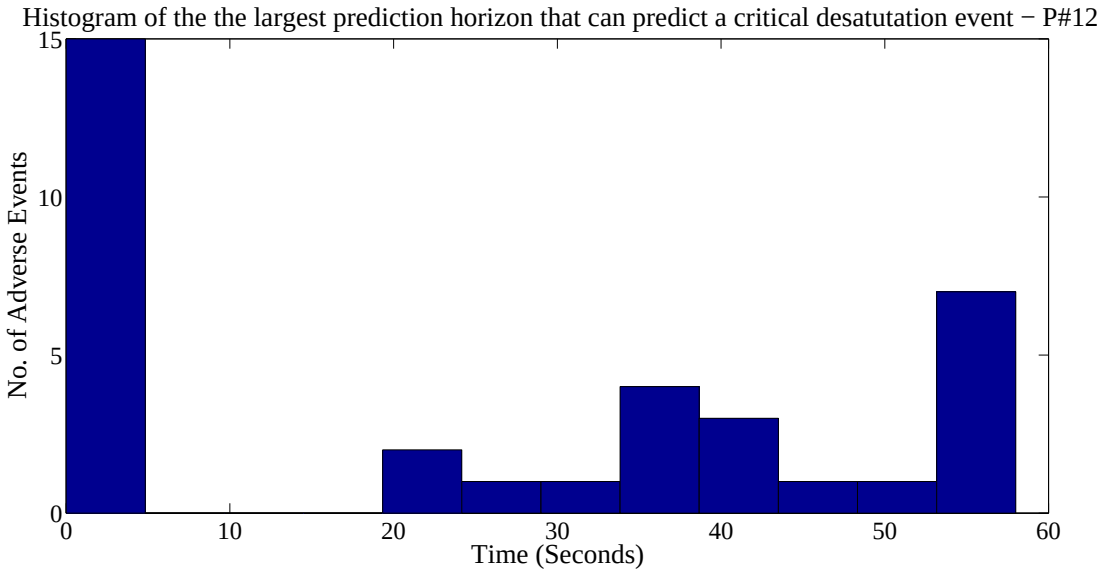
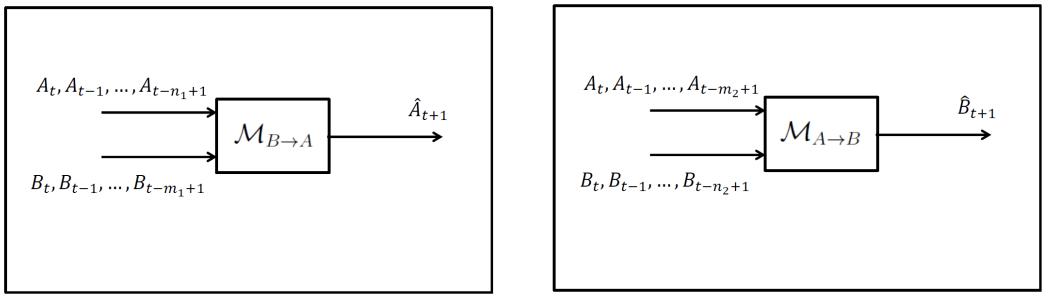


Figure 4.3: Histogram of the longest prediction horizon that can predict critical desaturation events within the SpO_2 time series of a postoperative patient using an AR-10 model.

4.4 Evaluating Predictions with Multi-Channel Data

In this dissertation, we consider a dynamic systems perspective to understand the cause-effect relationship between physiological signals. Figure 4.4 shows the approach considered for two physiological signals of interest A and B.



(a) Choice (1): B is input to A.

(b) Choice (2): A is input to B.

Figure 4.4: Modeling choices to study the directionality between two physiological signals A and B.

To understand the directionality between these two physiological signals A and B, we developed dynamic models ($\mathcal{M}_{A\rightarrow B}$) in which A is an input to B and other dynamic models in which B is an input to A ($\mathcal{M}_{B\rightarrow A}$). n and m in Figure 5.1 represent the order of the output and input polynomials (*i.e.*, the number of historical output and input observations needed to predict the next output value). The significance of the coefficients for the two modeling choices was investigated to understand the directionality between these physiological variables.

Although improved RMSE results might be achieved due to the additional degrees of freedom in the predictive model with additional data, the main goal of including additional data channels is to improve the ability to predict regions of clinical relevance within a physiological signal of interest. To evaluate the improvement in predicting regions of clinical relevance, we define the following metrics:

1. Evaluate the improvement in prediction sensitivity TPR of critical signal levels over different prediction windows due to the additional data channels.

2. Determine whether the inclusion of the additional data channels allows prediction of critical signal levels with a longer prediction horizon (improves the earliest time at which critical signal levels can be predicted).

Sections 4.5.1.2 and 4.5.3.3 present a detailed case study that applies the proposed approach for understanding the cause-effect relationship between SpO_2 and PR, both collected using POM systems. The dynamic model ($\mathcal{M}$) used to understand the directionality between these signals is a standard linear auto-regressive model with exogenous input (ARX model). We built ARX models in which SpO_2 is an output to the PR output, and other ARX models in which PR measurements are input and the SpO_2 measurements are output. Using the proposed method, our analysis shows that the two physiological variables might be best characterized by the SpO_2 measurements being an input (cause) and the PR values as an output (effect). Although SpO_2 ARX

models with PR input show a slightly improved RMSE for predicted SpO_2 signals, the proposed metrics show no significant improvement in the ability to predict critical desaturation events using PR data.

4.5 Case Study: Predicting Critical Desaturations in SpO_2 Time Series

This section presents a case study for using data driven dynamic models with blood oxygenation signals (SpO_2) recorded by POM devices. Linear autoregressive discrete time models are used to predict near future oxygen saturation levels of the hemoglobin in blood. Standard modeling methods are used in identifying dynamic systems models for these physiological signals. The metrics proposed in this dissertation are then used to evaluate the performance of the identified models in predicting critical oxygen desaturations in the blood. Using the proposed SpO_2 prediction metrics, we show that the combination of predictive models and frequent pulse oximetry measurements can be used as a warning of critical oxygen desaturation events that might have adverse effects on the health of patients. Moreover, we investigate the effect of including pulse rate dynamics (PR) as an input to SpO_2 models. We show no improvement in predicting critical desaturations if PR dynamics are included in the SpO_2 predictive models. Our results indicate oxygen in blood is an effective input to the pulse rate rather than vice versa.

We used SpO_2 and PR data (Section 3.1) for 119 postoperative adult patients generated by the pulse oximetry system. For demonstrating our results, 10 representative patients are presented here. The study was based on $15,000$ points of the data sets of each patient. The first 7500 points were used for model estimation and the second 7500 points were used for evaluating prediction results.

4.5.1 Time Series Modeling

4.5.1.1 Autoregressive models to predict future values of SpO_2

We have considered three scenarios to build AR Models.

Basic AR Modeling In AR modeling, the predicted signal $\hat{y}_t$ at time t ($t = n+1, ..., N$ where N denotes the total number of data samples available for modeling) is inferred as a linear combination of previously observed signals y_{t-i}

$$\hat{y}_t = \sum_{i=1}^{n} \theta_i y_{t-i} \tag{4.8}$$

where θ denotes the vector of AR coefficients to be determined, and n denotes the order of the model (*i.e.*, the number of previously observed SpO_2 measurements y_{t-i} used to predict a future SpO_2 value $\hat{y}_t$). The AR models are identified by finding the coefficients θ that best describe the dependencies in the time-series y. These coefficients are obtained based on the least squares fit (*Ljung*, 1987).

Basic AR modeling with smoothed data Equations (3.1), (3.2), or (3.4) can be used to smooth the raw time series and then the smoothed series can be used to identify AR Models.

Regularized AR modeling with smoothed data Regularization of AR models is performed by imposing a smoothness constraint on the least squares fit of the coefficients in Equation (4.8), resulting in the regularized least squares functional $g(\tilde{\theta})$, given by

$$g(\tilde{\theta}) = \|\tilde{y} - U_m\tilde{\theta}\|^2 + \lambda_m^2\|L_m\tilde{\theta}\|^2 \tag{4.9}$$

where $\tilde{y}$ denotes the $(N-m) \times 1$ vector of smoothed SpO_2 data, U_m denotes the $(N-m)\times m$ design matrix, $\tilde{\theta}$ represents the $m\times 1$ vector of regularized AR coefficients,

λ_m represents the model regularization parameter, and L_m denotes a well conditioned matrix chosen to impose smoothness on the AR coefficients.

4.5.1.2 Auto-regressive Models with Exogenous Input (ARX)

To study the interactive relationship between the oxygen saturation in blood (SpO_2) and the pulse rate (PR), we assume linear system models as approximations to the full nonlinear model around an operating condition. Therefore, we can use standard system identification techniques to fit linear autoregressive models with exogenous input (ARX) between these physiological variables. For an ARX(n,m), the predicted signal $\hat{y}_t$ at time t ($t = max(n, m) + 1, ..., N$ where N denotes the total number of data samples available for modeling) is inferred as a linear combination of previously observed output y_{t-i} and previously observed inputs u_{t-i}

$$\hat{y}_t + \sum_{i=1}^{n} a_i y_{t-i} = \sum_{i=0}^{m} b_i u_{t-i} \tag{4.10}$$

where n and m denote the order of output and input polynomials. The ARX models are identified by finding the coefficients a_i and b_i that best describe the dynamics of the output time-series y. These coefficients are estimated from input/output data using least squares fit principle (*Ljung*, 1987).

To investigate the interaction between SpO_2 and PR signals, we have built dynamic models using the approach discussed in Section 4.4.

4.5.2 Multi-Step Ahead SpO_2 Prediction

4.5.2.1 Linear Prediction

Future values of a discrete-time signal characterized by a linear identified time-series model are forecasted k steps into the future using historical data such that the predicted output $\hat{y}_{N+k}$ of a time series y can be seen as a linear function of previous available measurement of this time series $\hat{y}_{N+k} = f(y_N, y_{N-1}, ..., y_0)$. k is called the

prediction horizon, and corresponds to predicting output at time kT_s, where T_s is the sampling time of the discrete signal (2 seconds for POM data). For example, the one-step ahead predictor of an AR model can be expressed by the following equation:

$$\hat{y}_{N+1} = \sum_{i=1}^{m} \theta_i y_{N-i+1} \tag{4.11}$$

where y, θ can be replaced by $\tilde{y}, \tilde{\theta}$ as appropriate. Also, the same equation is used with ARX models but with the inclusion of the input terms. k-step ahead predictions are obtained using recursive prediction strategy by applying Equation (4.11) recursively.

4.5.2.2 Prediction Evaluation Metrics

Root Mean Square Error (RMSE) Linear dynamic models are fitted using least square error principle and it is a common practice in AR modeling to assess the prediction capability by calculating the root mean square error (RMSE) between the predicted and reference measurements of the time series. RMSE can be expressed as follows

$$RMSE(\hat{y}, \tilde{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \tilde{y}_i)^2} \tag{4.12}$$

where $\hat{y}$ is the predicted signal, $\tilde{y}$ is the reference smoothed data signal, and n is the number of predicted output measurements. We define the best fit model as the model that provides the lowest mean square error on the validation data set.

SpO_2 Prediction Grid Following Section 4.2, we developed the SpO_2 prediction grid shown in Table 4.2 to evaluate the ability of the AR model to predict critical desaturation events over prediction windows K in the the test time series for each patient. The critical signal level is chosen as $y_{cr} = 89\%$ due to its clinical significance (*Centers for Medicare and Medicaid Services*, 1993). Since a critical desaturation starts as soon as $y \leq y_{cr} = 89\%$, Pred.$= +1$ and -1 can be expressed by Pred.$\leq 89\%$

and $\geq 89\%$ respectively. Similar analogy applies to Ref.$= +1$ and Ref.$= -1$ in the reference SpO_2 signal.

Table 4.2: Prediction grid for K-seconds prediction windows.

	Ref. $\leq 89\%$	Ref. $> 89\%$
Pred. $\leq 89\%$	A	B
Pred. $> 89\%$	C	D

4.5.3 Results

4.5.3.1 Performance of different AR Modeling Techniques

We have developed software codes to build the three types of AR models presented in Section 4.5.1.1. The flexibility within the codes helped to test SpO_2 signals with different smoothing and modeling schemes.

We present here an example for Patient No. 12; a patient who experienced a very high frequency of desaturation events ($ODI_4 = 45.9$). A model order $n = 10$ and a prediction horizon $k = 10$ steps (20s) are illustrated. For signal smoothing, we used a moving average filter (MVA) of span $M = 5$ and compared the results to regularization using a regularization coefficient $\lambda_d = 20$. Figure 4.5 shows 20-seconds ahead prediction results for Basic-AR-10 from raw data compared to Basic-AR-10 and Reg-AR-10 from MVA smoothed data. Figure 4.6 shows the performance of the 20-seconds ahead prediction for the Basic-AR-10 and Reg-AR-10 models, both built from the regularized signal. Smoothing via regularization showed superior results compared to the moving average smoothing. For the Basic-AR-10 model, the noise in the signal and the discretization effects limited the prediction capability of this type of model. Table 4.3 summarizes RMSE performance for 20-seconds ahead predictions of different types of AR-10 models identified for patient No. 12 where it can be clearly seen that smoothing via regularization enhanced the prediction accuracy of the AR

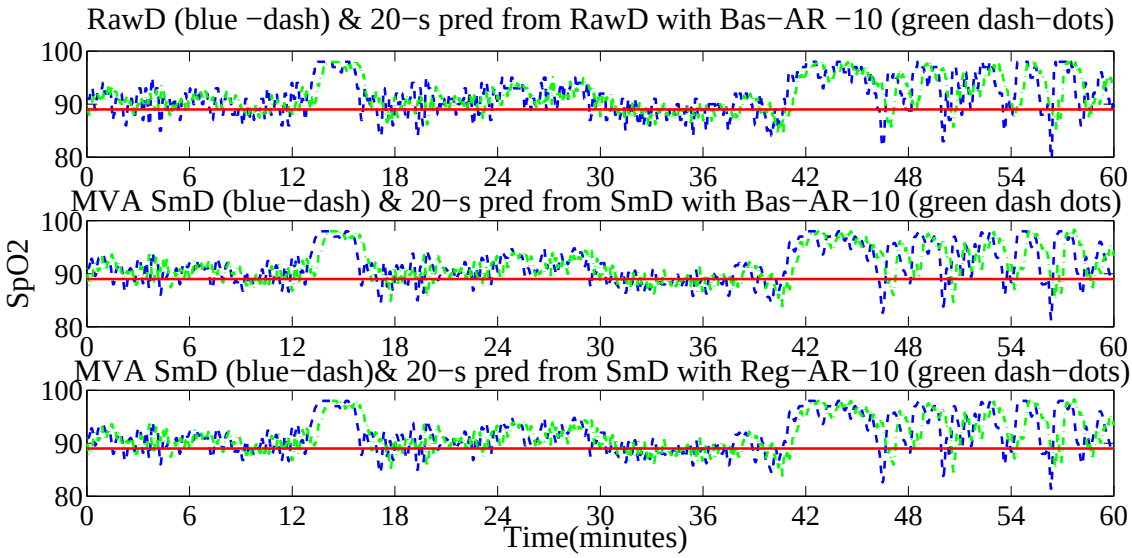


Figure 4.5: 20-seconds ahead prediction results using moving average smoothing, Patient No. 12. Discretization limited the prediction capability of AR Model. No significant improvement in prediction performance was noticed with Moving Average (MVA) smoothing.

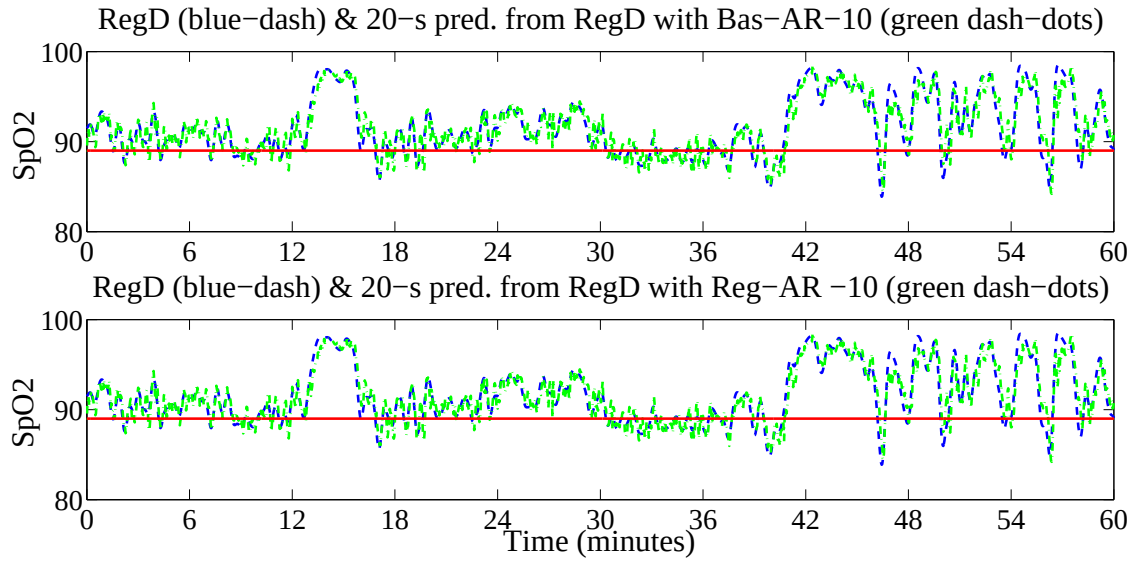


Figure 4.6: 20-seconds ahead prediction results using smoothing via regularization, Patient No. 12. Regularization significantly improved prediction performance compared to moving average smoothing methods.

models. Our results show that the dynamics of SpO_2 signals can be captured with an AR-10 model constructed from the regularized signal. Comparing the RMSE of $20s$ ahead predictions from AR-10 with the RMSE of predictions obtained using a higher model order such as AR-30 didn't show a statistically significant improvement (p-value$>$ 0.9).

Table 4.3: 20-s ahead RMSE of AR-10 for Patient No. 12

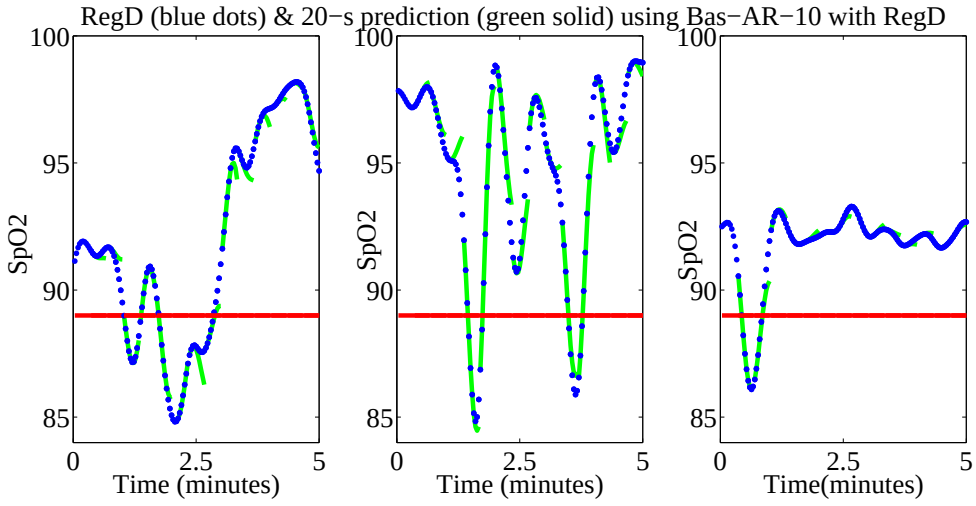
AR-Type	RMSE
Basic AR-10 with raw data	2.8873
Basic-AR-10, MVA-Smoothed Data,$M = 20$	2.8302
Reg-AR-10, MVA Smoothed-Data,$M = 20$	2.8292
Basic AR-10, Reg-Data,$\lambda_d = 20$	**0.9257**
Reg-AR-10, Reg-Data,$\lambda_d = 20$	1.2896

4.5.3.2 Evaluating SpO_2 Predictions

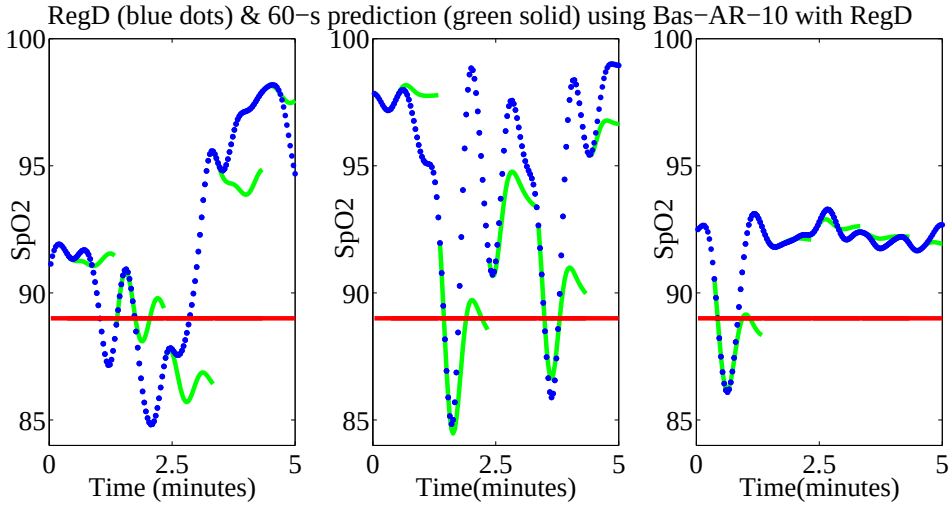
Two prediction windows (intervals) were considered $K = 20$s and 60s respectively. Figures 4.7(a) and 4.7(b) show the performance of the identified AR-10 models in the prediction intervals of 20 and 60 seconds respectively. Although the accuracy of prediction in the 60-seconds intervals is not as good as the 20-seconds ones, the AR-10 model is still able to capture critical desaturation events that will happen in the 60-seconds prediction intervals. The performance in predicting critical regions in both prediction intervals needs to be evaluated.

Table 4.4: Prediction grid for $20s$ prediction windows, Patient No. 12

	Ref. $\leq 89\%$	Ref. $> 89\%$
Pred. $\leq 89\%$	712	40
Pred. $> 89\%$	96	6642



(a) AR-10 model performance in 20-seconds prediction intervals for three patients. The prediction performance needs to be evaluated with respect to capturing critical signal levels over this prediction interval.



(b) AR-10 model performance in 60-seconds prediction intervals for three patients. Although increasing the prediction decreased the prediction accuracy, the model is still predicting critical destructions. The performance in predicting critical regions needs to be evaluated.

Figure 4.7: AR-10 model performance in 20s and 60s prediction intervals for three patients.

Table 4.5: Prediction grid for $60s$ prediction windows, Patient No. 12

	Ref. $\leq 89\%$	Ref. > 89
Pred. $\leq 89\%$	759	126
Pred. $> 89\%$	577	6008

Table 4.6: Prediction grid results for 20s prediction windows using AR-10 models

P#	$\|A\|$	$\|B\|$	$\|C\|$	$\|D\|$	TPR	TNR	PPV	NPV	ACC
10	0	0	0	7490	NA	100%	NA	100%	100%
12	712	40	96	6642	88.1%	99.4%	94.7%	98.6%	98.2%
18	124	1	10	7355	92.5%	99.9%	99.2%	99.9%	99.9%
37	339	20	22	7109	93.9%	99.7%	94.4%	99.7%	99.4%
59	232	12	21	7225	91.7%	99.8%	95.1%	99.7%	99.6%
80	81	9	6	7349	93.1%	99.9%	90.0%	99.9%	99.8%
89	71	0	7	7412	91%	100%	100%	99.9%	99.9%
118	38	2	5	7445	88.4 %	100%	95.0%	99.9%	99.9%
125	1181	84	99	6126	92.3%	98.6%	93.4%	98.4%	97.6%
136	130	13	12	7335	91.5%	99.8%	90.9%	99.8%	99.7%

Applying the proposed metric on the AR-10 model identified for patient No. 12, the prediction grid results for 20 and 60 seconds ahead in Tables 4.4 and 4.5 respectively show that the specificity (TNR) for both prediction intervals are excellent (99.4%, and 97.4% respectively) indicating that the model was able to identify successfully almost all of areas out of the dangerous zones in both prediction intervals. Also, the 60-seconds ahead prediction grid shows a relatively high positive predictive value (PPV) of 85.8% which reflects a high likelihood that the predicted desaturation events will be actual ones in both prediction intervals. On the other hand, the prediction grid results for 20 seconds prediction windows show a high sensitivity ($TPR = 88.1\%$) but for a prediction interval of 60 seconds it decreased to 56.7% indicating that the

Table 4.7: Prediction grid results for 60s prediction windows using AR-10 models

P#	$\|A\|$	$\|B\|$	$\|C\|$	$\|D\|$	TPR	TNR	PPV	NPV	ACC
10	0	0	0	7470	NA	100%	NA	100%	100%
12	759	126	577	6008	56.8%	97.9%	85.8%	91.2%	90.6%
18	126	5	89	7251	58.6%	99.9%	96.2%	98.8%	98.7%
37	359	17	174	6920	67.4%	99.8%	95.5%	97.5%	97.4%
59	244	7	136	7083	64.2%	99.9%	97.2%	98.1%	98.1%
80	70	13	80	7307	46.7%	99.8%	84.3%	98.9%	98.8%
89	72	14	66	7318	52.2%	99.8%	83.7%	99.1%	98.8%
118	38	9	45	7378	45.8%	99.9%	80.9%	99.4%	99.3%
125	1213	82	503	5672	70.7%	98.6%	93.7%	91.9%	92.2%
136	143	14	163	7150	46.7%	99.8%	91.1%	97.8%	97.6%

predictive model is still able to capture more than 50% of the critical events over the longer prediction window. Tables 4.6 and 4.7 show prediction grid results with TPR, TNR, PPV, NPV, and ACC values for 20 and 60 seconds ahead predictions respectively for ten patients. As shown in these tables, the values of TNR for both prediction intervals are close to 100% for all patients. Also, in general the PPV results are high in both intervals. In both prediction horizons, the AR model for Patient No. 10 who didn't experience any critical desaturation events didn't predict any false critical ones.

4.5.3.3 Effect of Pulse Rate (PR)

We have developed software codes to investigate the interactive relationship between SpO_2 and PR measurements. The regularized SpO_2 and PR were used to construct these models. The software codes identify dynamic systems models with different orders (n,m) using 7500 points of the two time series and the models are

validated on 7500 points of the time series that are different from the ones used for modeling. For efficient computation, the software codes select the model order that yields the lowest RMSE using candidate model orders n and m ranging from 1 to 15.

Best fit modeling results are shown in Figure 4.8. The distribution of the ARX model coefficients (n,m) shows that the effect of SpO_2 historical measurements on the PR output values is larger than the effect of the historical PR values on the SpO_2 measurements. The effect of PR as an input affecting SpO_2 dynamics can be analyzed through the number (m) and magnitude (b_i) of the input coefficients of the PR measurements. The PR input has a very short memory effect on SpO_2 (max time lag $m = 3$) and the coefficients have negligible values (max $|b_i| = 0.00003$). Although the model order could be freely selected, the low values of the coefficients (b_i) represent low dependencies between PR and SpO_2 measurements in a model that assumes PR as an input to SpO_2. Including the PR data in predicting SpO_2 in linear models is not effective. In contrast, it can be seen easily that SpO_2 is an effective input to PR.

Therefore, according to this analysis, these two physiological variables might be best characterized by the SpO_2 measurements being an input (cause) and the PR values as an output (effect). This agrees with previous studies and clinical observations that indicated pulse rate changes lag behind the changes in the SpO_2 (*Somers et al.*, 1995).

However, our goal is to assess whether better prediction of critical events can be achieved by incorporating PR dynamics in the SpO_2 models. To compare SpO_2 prediction results with and without the input from PR dynamics, we considered an ARX(10,10) and compared prediction results for this model with the previously obtained prediction results with the AR-10 model. Thus, we tested these models over 60s prediction intervals with the proposed SpO_2 prediction grid as shown in Table 4.8. Comparing these results to the previously obtained ones with AR-10 models in

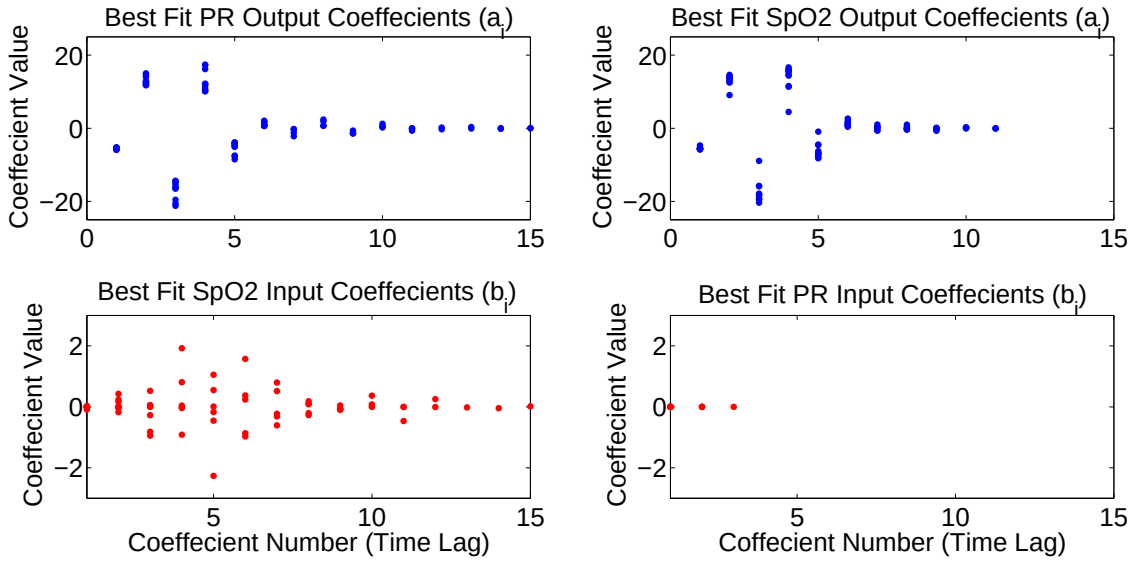


Figure 4.8: Distribution of Best Fit ARX model coefficients for two modeling choices for SpO_2 and PR. The effect of SpO_2 historical measurements on the PR output values is larger than the effect of the historical PR values on the SpO_2 measurements.

Table 4.7 shows no improvement in the TPR of the predictions of the critical events in the SpO_2 time series of patients (p-value=0.814).

4.5.3.4 Predictive Capability Based on the Longest Horizon that Can Predict Critical Desaturation Events

We can now look once again at the identified linear models but this time trying to inspect the ability of these models to predict adverse events by evaluating how early these models are able to predict critical events. Figure 4.9 shows the histograms of the duration between the time AR-10 and ARX(10,10) models predict the critical event and that actual time of occurrence of the desaturation event reflecting how early a critical event can be predicted using both types of models. The red histogram of AR-10 predictions show that 46 critical events (56.2% of the total number of events) were able to be predicted 10 - 60 ahead of time using this type of model. However, looking at same histogram but for ARX(10,10), no improvement in detecting these

Table 4.8: Prediction grid results for 60s prediction windows using ARX(10,10)

P#	$\|A\|$	$\|B\|$	$\|C\|$	$\|D\|$	TPR	TNR	PPV	NPV	ACC
10	0	0	0	7470	NA	100%	NA	100%	100%
12	672	2	664	6133	50.3%	99.9%	99.7%	90.2%	91.1%
18	125	4	90	7252	58.1%	99.9%	96.9%	98.8%	98.7%
37	358	17	175	6920	67.2%	99.8%	95.5%	97.5%	97.4%
59	233	20	147	7070	61.3%	99.7%	92.1%	98%	97.8%
80	65	7	85	7313	43.3%	99.9%	90.3%	98.9%	99.8%
89	72	6	66	7326	52.2%	99.9%	92.3%	99.1%	99%
118	36	28	47	7359	43.4%	99.6%	56.3%	99.4%	99%
136	134	9	172	7155	43.8%	99.9%	93.7%	97.7%	97.6%
125	1226	65	490	5689	71.4%	98.9%	95%	92.1%	92.6%

adverse events can be noticed due to the inclusion of the PR dynamics in the models (p-value 0.986). These results also agree with our aforementioned results.

4.6 Summary

This chapter proposed new approaches for evaluating predictions of physiological signals. The metrics presented evaluate predictions of physiological signals for their ability to predict critical levels of abnormality within these signals. First, evaluating predictions over prediction windows of fixed lengths is discussed. Then, evaluating predictions for the largest horizon that can be predict critical events is considered. Furthermore, a dynamic systems approach is presented for investigating the interactive relationship between multi-channel physiological data. Metrics are defined to characterize the improvement in prediction using additional data channels.

A case study was presented for evaluating predictions of critical desaturation events in the SpO_2 time series of patients. While the models used in this case study

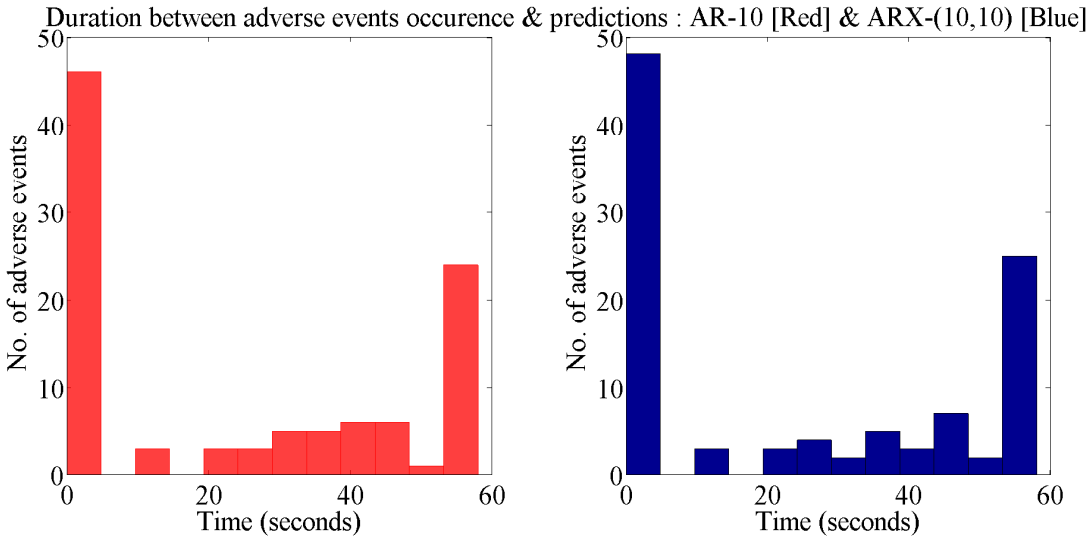


Figure 4.9: Duration between AR-10 & ARX-(10,10) predictions and adverse events occurrence for 10 patients. No improvement on the largest horizon that can predict a critical desaturation events can be noticed due to the inclusion of the PR dynamics in the models (p-value 0.986).

were standard linear models, they provided the insight needed to show the importance of introducing prediction evaluation metrics that are more suitable for physiological systems. The case study presented was based on 15000 points (4.17 hrs) of the data sets of each patient. The first 7500 points were used for model estimation and the second 7500 points were used for evaluating prediction results. Future work may include evaluating prediction results using a population based model instead of individualized (patient specific) models to reduce the amount of time needed for monitoring patients before starting to make real time predictions. Also, future efforts may include considering shorter data lengths for building models that can be updated over time as more data becomes available. The prediction results and proposed metrics were presented with AR models but in our other work we investigated ARMA models and artificial neural networks (ANN) and no significant improvements in predicting critical events were found (*ElMoaqet et al.*, 2013a).

CHAPTER V

Models to Predict Clinically Relevant Events in Physiological Signals

The work presented in this chapter appears in proceedings of the 2014 *IEEE Conference on Medical Measurements and Applications* (*ElMoaqet et al.*, 2014b) and also has been submitted for journal publication (*ElMoaqet et al.*).

5.1 Introduction

The research in this chapter advances the state of art in time-series prediction models by developing novel models optimally designed to capture clinically relevant patterns in physiological signals. New metrics are considered for optimizing time series models with respect to predictions of clinically relevant events instead of (standard) prediction error based metrics.

We consider two problems in this chapter. First, we consider predicting abnormal deviations from (operating) signal baseline. Second, we consider predicting only critical signal levels within the time series. The second problem considers more extreme signal levels which are rare events within physiological signals. Thus, to address the issue of the relative paucity of such events in physiological data, different statistical metrics are considered for optimization.

In both problems, we consider a fixed prediction horizon (k time steps) for developing the proposed models. Note that small letter k is used to stress that a fixed prediction horizon is used for developing models. The selection of a fixed horizon is driven by the clinical interest of evaluating all physiological predictions in a time series with respect to the same prediction horizon. Moreover, fixing the prediction horizon addresses the effectiveness of the predictions in capturing the onset of the events of interest.

5.2 Predicting Abnormal Deviations from Signal Baseline

5.2.1 k-Step Ahead Dynamically Adjusted Threshold Prediction Metric

In the dynamics of physiological time series, abnormal deviations from signal (operating) baseline are often defined by either a reduction or increase in the signal value that exceeds a specific percentage or a threshold from the current baseline. In such cases, the signal baseline is the reference from which deviations are compared. In fact, abnormal deviations from signal baseline have been used to define several disease conditions and also have been linked to several health complications (for example: blood pressure, air flow, and SpO_2) (*Pool et al.*, 2009; *Swigris et al.*, 2009; *American Academy of Sleep Medicine*, 2007). Depending on the evolving baseline of the signal and the amount of change in the signal value, these types of events may or may not result in more extreme critical signal levels. Nevertheless, these events themselves are of clinical interest to be predicted ahead of time.

Therefore, we define a dynamically adjusted threshold metric to evaluate physiological predictions based on their ability to predict abnormal deviations from the signal baseline. Without loss of generality, consider an abnormal deviation defined by a significant *reduction* from the current signal baseline $g(\mu_t)$, where g is real valued function that computes the (dynamic) threshold that is clinically considered as an

abnormal deviation from the baseline at time t (μ_t). At each time step t, the dynamic threshold performance metric can be evaluated by forecasting the signal k steps into the future to obtain $\hat{y}_{t+k}$ and checking whether or not the predicted signal value is able to capture abnormal deviations from the current baseline μ_t.

The proposed performance metric can be expressed as a binary classification function (*Altman and Bland*, 1994a,b) that considers all k-steps ahead predictions of a physiological time series y. The k-steps ahead dynamically adjusted threshold performance metric is shown in table 5.1. The binary function regions A, B, C and D are assigned to true positive (TP), false positive (FP), false negative (FN), and true negative (TN) regions respectively.

Table 5.1: k-steps ahead dynamically adjusted threshold metric

	$y_{t+k} \leq g(\mu_t)$	$y_{t+k} > g(\mu_t)$
$\hat{y}_{t+k} \leq g(\mu_t)$	A	B
$\hat{y}_{t+k} > g(\mu_t)$	C	D

Furthermore, we can use sensitivity (TPR), specificity (TNR) and positive predictive value (PPV) in Equations (4.3-4.5) as statistical measures to evaluate the performance of the dynamic models in predicting k-steps ahead abnormal events.

5.2.2 Case Study: A Novel Model for Predicting Abnormal Desaturations in SpO_2

Having developed the metric that characterizes the performance of prediction models in capturing abnormal deviations from signal baseline, we will now use it in developing a time series model that optimizes this metric with respect to predictions of such clinical events within the time series.

The proposed model is an auto-regressive prediction model implementing the direct prediction strategy such that the k-step ahead prediction at any time t can be

predicted as a linear combination of the last n available measurements as follows

$$\hat{y}_{t+k} = \sum_{i=1}^{n} \phi_i y_{t-i+1} \tag{5.1}$$

where $\phi_1, \phi_2, ..., \phi_n$ are coefficients of the k-step ahead prediction model. Since the proposed model implements direct prediction strategy (discussed in Section 2.2.2), the model coefficients $\phi_1, \phi_2, ..., \phi_n$ depend on the prediction horizon of interest k.

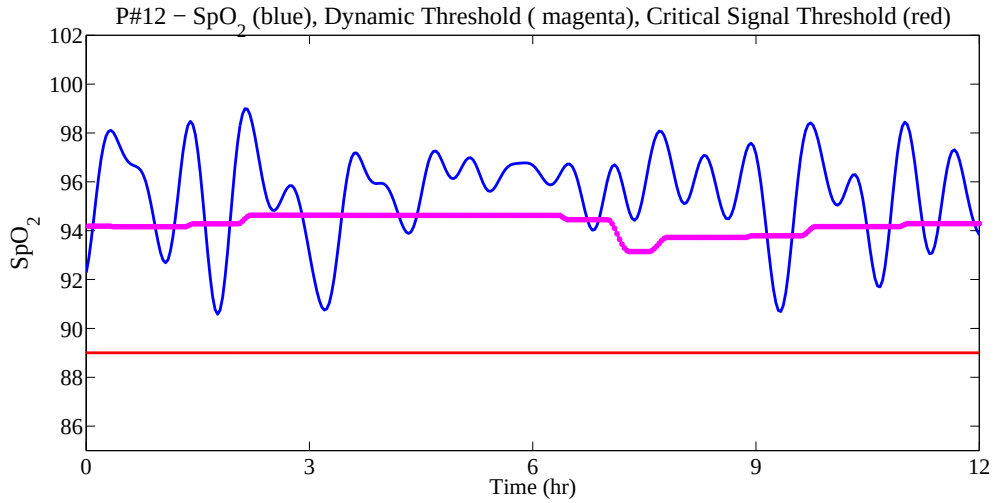
Recognizing that the clinical definition of the baseline abnormal deviations needs to be incorporated in the model optimization, we will present the model development with a case study of predicting abnormal deviations from SpO_2 baseline. Here, we need to clinically define the SpO_2 baseline and abnormal deviations from this baseline.

SpO_2 baseline: Following *Oliver and Flores-Mangas* (2006) and *Lee et al.* (2004), the baseline can be computed as the moving average over a window of 5 minutes of SpO_2 data after all the data lower than the ninety-fifth quartile are filtered such that only the top 5% of the samples are considered for calculating the mean.

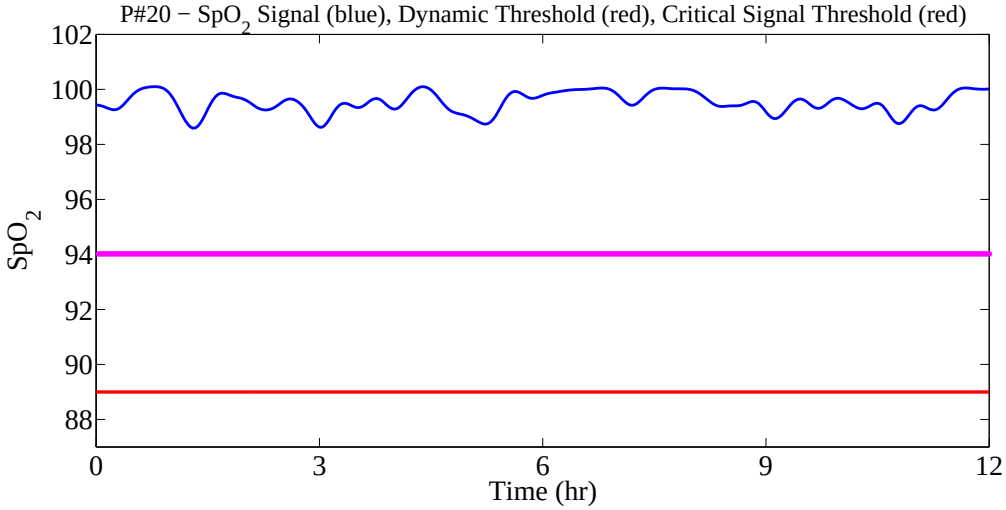
Abnormal deviations from SpO_2 baseline: Following *Berry et al.* (2012) and *Swigris et al.* (2009), abnormal events in the SpO_2 time series are desaturations in which SpO_2 falls 4 points or more below the signal baseline . Thus, the (dynamic) threshold for abnormal events at each time step t is defined by $g(\mu_t) = \mu_t - 4$. It can be noticed from the SpO_2 baseline definition that no significant changes in baseline are expected in k time steps (near future). Thus, we will compare predictions at time $t + k$ with the dynamic threshold computed at the current time t since this threshold can be readily computed using the available SpO_2 measurements.

Figure 5.1 illustrates the dynamically adjusted threshold compared to the critical signal threshold over two SpO_2 windows for two patients. The SpO_2 time series in Figure 5.1(a) shows abnormal desaturations that breach the adaptive threshold while

no abnormal desaturations exist over the SpO_2 signal shown in Figure 5.1(b).



(a) Dynamically adjusted threshold captures abnormal desaturation levels in the SpO_2 time series of a patient that experiences a high frequency of desaturation events.



(b) No noticeable changes are observed in the dynamically adjusted threshold and no abnormal events noticed for a patient with a low frequency of destauration events.

Figure 5.1: Illustration of the dynamically adjusted threshold $g(\mu_t)$ compared to the critical signal threshold y_{cr} for the SpO_2 time series of two patients.

To maximize the ability to predict abnormal deviations from signal baseline in k-step ahead predictions, we formulated a mixed integer programming problem (MIP). The elements of this optimization problem are discussed as follows.

5.2.2.1 Parameters/Inputs

The inputs to the optimization problem are as follows.

- $y_t \in R$, $t = 1, ...N$; regularized SpO_2 signal measurements in the time period $\{1, ..., N\}$ (training data).
- $z_t \in \{0, 1\}$, $t = 1, ..., N - k$; indicators for y_t whether or not an abnormal desaturation event happens at y_{t+k} (i.e dynamic threshold check for observed signal).

$$z_t = \begin{cases} 1, & \text{if } y_{t+k} \leq \mu_t - 4. \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

5.2.2.2 Decision variables

The decision variables of the model optimization problem are defined as follows.

- ϕ_i , $i = 1, .., n$; coefficients of the proposed dynamic model of order n.
- $\hat{y}_{t+k} \in R$, $t = n, ..., N - k$; k-step ahead predicted SpO_2 measurements at each time step t.
- $\hat{z}_t \in \{0, 1\}$, $t = n, ..., N - k$; indicators for $\hat{y}_{t+k}$ whether or not an abnormal desaturation event is predicted at time $t + k$ (i.e. dynamic threshold check for predicted measurements).

$$\hat{z}_t = \begin{cases} 1, & \text{if } \hat{y}_{t+k} \leq \mu_t - 4. \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

5.2.2.3 Optimization Model: Minimizing FN

To optimize the metric in Table 5.1, we will consider in this problem minimizing FN (*i.e.,* minimize instances at which the k-step ahead prediction fails to predict an abnormal signal level that occurs in the reference SpO_2 signal at this horizon). Note that minimizing FN (Region C) is equivalent to maximizing TP (Region A) of the prediction grid in Table 5.1. The optimization model can be expressed as follows.

$$\min \sum_{t=n}^{N-k} \underbrace{(1-\hat{z}_t)z_t}_{FN(\text{Region } C)} \tag{5.4}$$

subject to:

$$\hat{y}_{t+k} = \sum_{i=1}^{n} \phi_i y_{t-i+1} \tag{5.5}$$

$$0 \leq \hat{y}_{t+k} \leq 100, t = n, .., N-k \tag{5.6}$$

$$\hat{z}_t \geq (\mu_t - 4 - \hat{y}_{t+k})/100, t = n, ..., N-k \tag{5.7}$$

$$1 - \hat{z}_t \geq (\hat{y}_{t+k} - \mu_t + 4)/100, t = n, ..., N-k \tag{5.8}$$

$$\hat{z}_t \in \{0, 1\}, t = n, ..., N-k \tag{5.9}$$

In this model, objective function (5.4) is a linear function. Equation (5.5) expresses the k-step ahead SpO_2 prediction at each time t as a linear combination of the n most recently observed measurements up to time t. Inequality (5.6) ensures the predicted oxygen saturation level is neither negative nor exceeds the max possible SpO_2 saturation level. Inequalities (5.7, 5.8) provide a compact representation of binary labels $\hat{z}_t$ based on k-step ahead predictions $\hat{y}_{t+k}$ at each time step t. This representation is equivalent to the definition of $\hat{z}_t$ in Equation (5.3) but is more convenient to optimization solvers.

It should be mentioned here that the set up of this optimization problem prevents the solution that trivially predicts abnormal signal levels at all time steps resulting in

$TPR = 100\%$ while not being able to predict any normal signal levels ($TNR = 0\%$ or equivalently a false positive rate $FPR = 100\% - TNR = 100\%$. First, the objective function is set to a cost minimization problem. Second, the inequalities (5.7, 5.8) are defined such that z_t and $\hat{z}_t$ are set to 1 whenever y_{t+k} and $\hat{y}_{t+k}$ respectively represent an abnormal signal level and zero otherwise. Although FP predictions (Region B) are not explicitly included in the optimization model, the formulation of the optimization problem implicitly rejects trivial prediction models.

5.2.3 Results and Discussion

We used SpO_2 signals described in Section 3.1. For the analysis of this study, 2500 continuous points were used for constructing predictive dynamic models from the SpO_2 time series of each patient and distinct 2500 points were used for evaluating predictions. To account for inter-patient variability, models were individualized for each patient. For demonstrating our results, 10 representative patients are considered here.

In Section 4.5.3.1, we have shown that an AR-10 is sufficient to model the dynamics of SpO_2 signals and that higher order models might not be needed for them. Therefore, in this study, we fix the order for the proposed model to $n = 10$ to compare its performance with standard AR model of the same order. For illustrating results, we use a prediction horizon of 20 seconds ($k = 10$) to construct the proposed predictive models and compare them with the performance of the standard AR-10 models in predicting abnormal desaturations in the blood.

5.2.3.1 Minimizing FN (Region C) of the Dynamic Threshold Metric

Considering the objective function in Equation (5.4), we can fit models with maximized TPR. Figure 5.2 compares 20s ahead predictions using the proposed AR-10 model with predictions obtained using standard AR-10 model for Patients Nos. 68,

83 and 125. The red line in each of the subplots in the figure represents the dynamic threshold with respect to SpO_2 baseline at different time instances. As seen in this figure, the 20s ahead predictions obtained from the proposed model, shown in green, are better able to detect regions of abnormal desaturations through magnifying the prediction response in areas of abnormal events to maximize the number of true predictions in these regions. Thus, significant under-predictions in some areas of abnormal events as well as significant over-predictions in some areas of normal SpO_2 levels are noticed compared to standard AR-10. Nevertheless, the objective is to maximize the ability to predict abnormal SpO_2 events rather than minimizing the prediction error.

RMSE values for predictions obtained from both types of models are shown in the last column of Tables 5.2 and 5.3. As expected, RMSE values for predictions obtained from proposed model are not as good as RMSE from standard AR-10 model. For Patients Nos. 68, 83 and 125, Figure 5.2 illustrates the increase in RMSE. On the other hand, the proposed model has better ability to capture abnormal changes of SpO_2 in blood. Using the proposed AR-10, prediction sensitivity (TPR) increases from 69.7%, 69.4% and 84.7% to 95.5%, 99.1% and 98.9% respectively showing significant improvement in predicting clinically relevant events.

In order to evaluate the overall predictive capability of the proposed AR-10 model compared to the standard AR-10 model, we need to construct the dynamic threshold prediction grid and analyse the performance of the predictive models through the statistical measures of the grid TPR, PPV and TNR. Tables 5.2 and 5.3 show dynamic grid results for 20s ahead predictions obtained from proposed AR-10 and standard AR-10 respectively. As shown in these tables, there is a significant increase in prediction sensitivity (TPR) for the proposed AR-10 over the standard AR-10. This increase in TPR indicates that the proposed model is better able to explain the dynamic behavior for abnormal changes of SpO_2 in blood. Furthermore, although

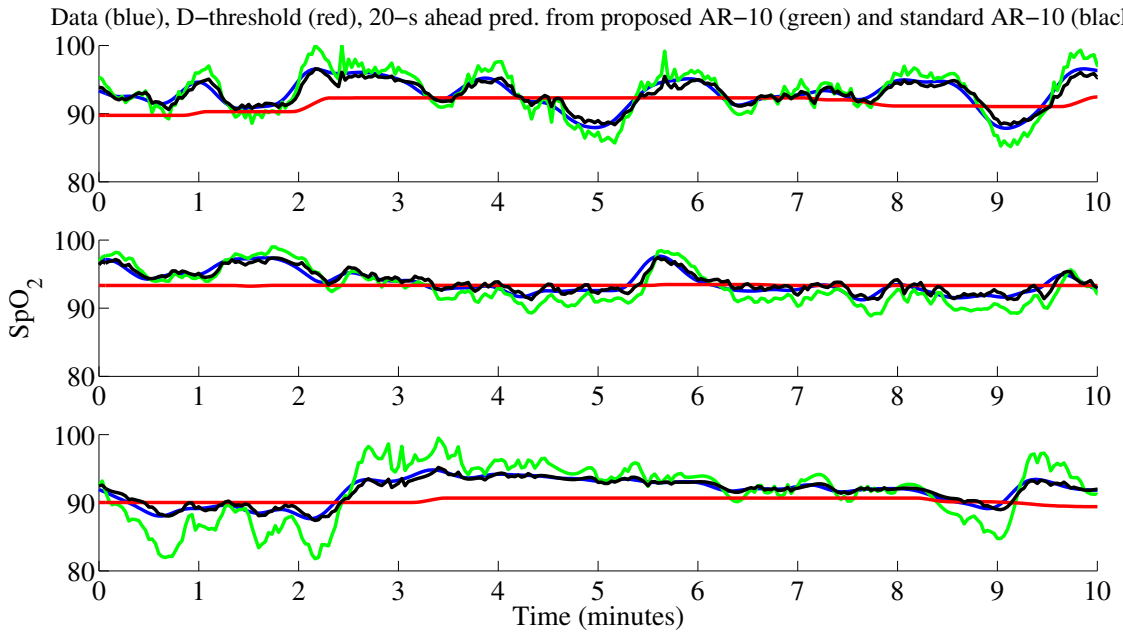


Figure 5.2: 20s ahead predictions using proposed AR-10 compared to standard AR-10 for P# 68, 83, 125

there is a slight decrease in prediction specificity (TNR) for the proposed model, it is still able to identify regions of normal behavior of SpO_2 dynamics in an acceptable way. However, inspecting the positive predictive values (PPV) shows that the improved performance in predicting abnormal events is accompanied with increase in the rate of false predictions reflected by decreased PPV. These results show that the objectives of minimizing both false negatives and false positives (regions C and B in the dynamic grid) are competing ones.

5.2.3.2 Incorporating FP (Region B of the Dynamic Threshold Metric) Explicitly in the Optimization Model

To investigate the ability to improve the prediction performance by explicitly incorporating FP in the optimization model, we modified the objective function in Equation 5.4 as follows

Table 5.2: 20s ahead dynamic grid results and RMSE for proposed AR-10 model

P#	$\|A\|$	$\|B\|$	$\|C\|$	$\|D\|$	TPR	PPV	TNR	RMSE
9	241	262	26	1812	90.3%	47.9%	87.4%	1.36
10	181	21	55	2084	76.7%	89.6%	99%	0.47
12	275	267	12	1787	95.8%	50.7%	87%	1.28
42	222	34	56	2029	79.9%	86.7%	98.4%	0.7
59	106	84	5	2146	95.5%	55.8%	96.2%	0.69
68	211	204	10	1916	95.5%	50.8%	90.4%	1.17
83	214	216	2	1909	99.1%	49.8%	89.8%	1.13
89	158	133	18	2032	89.8%	54.3%	93.9%	0.78
99	199	31	8	2103	96.1%	86.5%	98.5%	0.95
125	259	197	3	1882	98.9%	56.8%	90.5%	1.92

Table 5.3: 20s ahead dynamic grid results and RMSE for standard AR-10 model

P#	$\|A\|$	$\|B\|$	$\|C\|$	$\|D\|$	TPR	PPV	TNR	RMSE
9	162	75	105	1999	60.7%	68.4%	96.4%	0.67
10	192	17	44	2038	81.4%	91.9%	99.2%	0.37
12	219	71	68	1983	76.3%	75.5%	96.5%	0.63
42	216	33	62	2030	77.7%	86.7%	98.4%	0.62
59	84	22	27	2208	75.7%	79.2%	99%	0.42
68	154	64	67	2056	69.7%	70.6%	97%	0.55
83	150	28	66	2097	69.4%	84.3%	98.7%	0.50
89	113	21	63	2144	64.2%	84.3%	99%	0.49
99	198	22	9	2112	95.7%	90%	99%	0.63
125	222	23	40	2056	84.7%	90.6%	98.9%	0.42

$$\min \sum_{t=n}^{N-k} \underbrace{(1-\hat{z}_t)z_t}_{FN(\text{Region } C)} + \underbrace{\hat{z}_t(1-z_t)}_{FP(\text{Region } |B|)} \tag{5.10}$$

The modified objective function in Equation (5.10) is a two sided objective function that minimizes both FN (Region C) and FP (Region B) of the dynamic threshold metric. The goal of the optimization model is to fit the time series model by *minimizing* the number of mismatches between the predicted and reference k-step ahead SpO_2 values. The solution returned for the optimization problem in this setting was a null model that avoids making any predictions to maintain the lowest number of false predictions. Indeed, maximizing the sensitivity of the abnormal level predictions is accompanied with a larger increase in false predictions. Thus, the total value of the objective function with contributions from the two terms in Equation (5.10) is higher than the one returned by a null model that simply suggests to predict the mean of the signal at each time step.

5.2.4 Summary

This section presented a new approach for modeling and predicting the dynamics of physiological signals. Recognizing the clinical definition for abnormal changes in a physiological signal, a performance metric based on an adaptive threshold is defined to evaluate predictive models for their ability to capture baseline abnormal deviations in a physiological signal. Then, the new performance metric is used to identify a novel modeling framework that optimizes this metric. As a case study, the modeling approach was applied to SpO_2 time series to develop auto-regressive models that maximize the ability to predict abnormal events over a prediction horizon of $k = 20s$. Our results show that the proposed model is better able to capture the dynamic behavior for abnormal changes of SpO_2 compared to standard autoregressive models.

Results show that minimizing false negative predictions is accompanied with an

increase in the rate of false positive prediction due to the competing nature between the regions of false negatives and true positives. Future work may include improving the solution with a modified cost function that allows incorporating false positive predictions explicitly in the optimization problem. Alternatively, the false positive predictions may be constrained at a fixed level and then a prediction model can be fitted to maximize prediction sensitivity while not exceeding the maximum allowed rate of false positives.

5.3 Predicting Critical Signal Levels

In this section, we consider the problem of developing time series models optimized with respect to predictions of critical levels of abnormality within a physiological time series. First, a performance metric is proposed for evaluating multi-step ahead predictions of critical levels of abnormality over a fixed prediction horizon k time steps. The proposed metric addresses the fact that it is more important to predict regions of critical signal levels than the absolute amplitudes of these signals. Subsequently, this metric is used to build a framework for optimizing auto-regressive models capable of predicting regions of clinical interest in physiological signals. The model structure used is a standard auto-regressive structure that implements a direct prediction strategy but the major contribution is optimizing this predictive model to capture clinically relevant events in time series. To address the issue of the relative paucity of critical events in physiological data, different statistical metrics are considered for optimization. To account for inter-patient variability, the proposed models are individualized for each patient.

5.3.1 k-Step Ahead Prediction Metric for Critical Signal Levels

k-step ahead prediction has been typically used in evaluating predictive models of physiological time series (*Sparacino et al.*, 2007; *Reifman et al.*, 2007; *Gani et al.*,

2009). The classical definition of this metric doesn't address the ability to predict critical signal levels. This can be further illustrated by looking at Figure 5.3 where we would like to perform multi-step ahead prediction for the signal y at time t using the observed time history of y. For example, $\hat{y}_{t+k_2} \leq y_{cr}$ is needed for this prediction to be successful. On the other hand, $\hat{y}_{t+k_1} > y_{cr}$ indicates that $\hat{y}_{t+k_1}$ mispredicts the critical signal value $y_{t+k_1} \leq y_{cr}$.

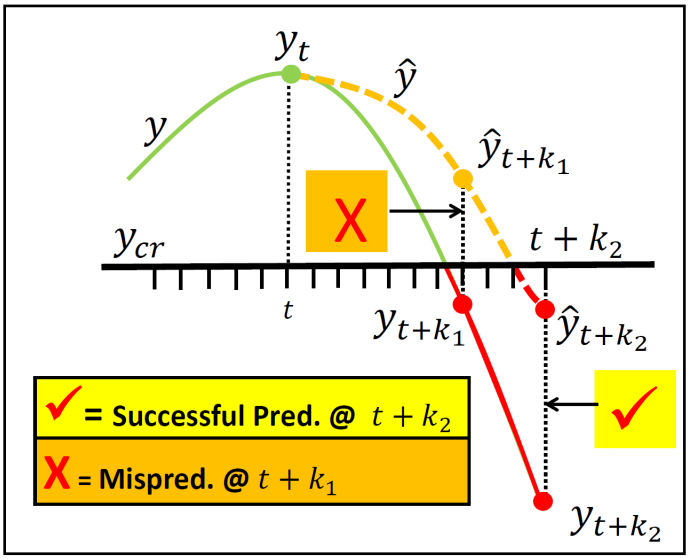


Figure 5.3: Modified k-step ahead Metric

Accordingly, we need to modify the traditional k-step ahead definition to ensure predicting clinically relevant regions in physiological signals. Considering a physiological time series in the interval $[t_i, t_f]$, we can denote the occurrence of a critical signal level k steps after time instance $t \in [t_i, t_f]$ by Equation (5.11)

$$\gamma_t = \begin{cases} +1, & \text{if } y_{t+k} \leq y_{cr}. \\ -1, & \text{if } y_{t+k} > y_{cr}. \end{cases} \tag{5.11}$$

Similarly, we can denote the prediction state by $\hat{\gamma}_t$ defined by Equation (5.12)

$$\hat{\gamma}_t = \begin{cases} +1, & \text{if } \hat{y}_{t+k} \leq y_{cr}. \\ -1, & \text{if } \hat{y}_{t+k} > y_{cr}. \end{cases} \tag{5.12}$$

Considering a critical signal threshold y_{cr} has been breached k steps from time instance t, our target is to evaluate whether or not the prediction $\hat{y}_{t+k}$ is able to capture this critical signal threshold. Repeating this recursively to test all time instances $t \in [t_i, t_f]$ results in the k-step ahead prediction grid shown in Figure 5.4 where points in Regions A and D on the main diagonal represent points of good prediction, points in Region C off-diagonal represent the ones at which the model fails to detect critical events, and points in Region B off-diagonal represent false prediction points. Regions A, B, C, D are assigned as the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) areas respectively.

$k-$ Step Ahead Prediction Metric	$\gamma = 1$	$\gamma = -1$
$\hat{\gamma} = 1$	A	B
$\hat{\gamma} = -1$	C	D

Figure 5.4: Proposed k-step ahead prediction metric.

The proposed metric can be considered as a statistical binary classification function (*Altman and Bland*, 1994a,b) with sensitivity (TPR), specificity (TNR), positive predictive value (PPV), and accuracy (ACC) in Equations (4.3-4.5, 4.7) used as statistical measures to evaluate its performance.

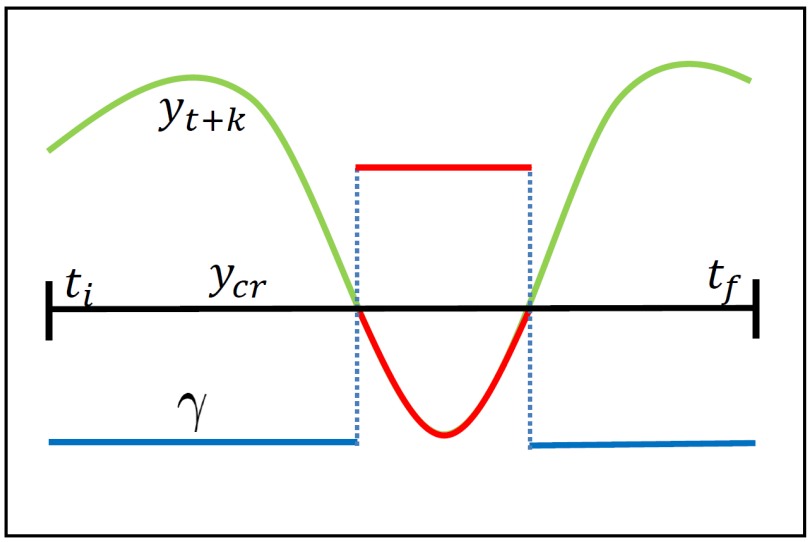


Figure 5.5: Predicting clinically relevant regions

5.3.2 A Framework for Optimizing k-Step Ahead Predictions of Critical Patterns

Having redefined the k-step ahead metric for threshold based physiological signals, we will next use it in identifying an AR modeling framework for optimizing models for critical signal levels in these signals. The inputs to the proposed model at each time instance t will be the n most recent observations while the output will characterize the *physiological state* of $\hat{y}_{t+k}$ instead of the *predicted value*. We can express this model by Equation (5.13)

$$\hat{z}_t = \sum_{i=1}^{n} \phi_i y_{t-i+1} = \Phi^T Y_t \tag{5.13}$$

where $\Phi = [\phi_1 ... \phi_n]$ is a vector of autoregressive model coefficients of length n and $Y_t = [y_t ....y_{t-n+1}]$ contains the n most recent observation up to time t. $\hat{z}_t > 0$ indicates prediction of a critical signal level ($\hat{\gamma}_t = 1$) while $\hat{z}_t < 0$ indicates a normal prediction ($\hat{\gamma}_t = -1$). Using this formulation will allow identifying models optimized to predict regions of clinical relevance as shown in Figure 5.5.

We will pose the model identification problem as a support vector machine (SVM) problem (*Bishop et al.*, 2006) in which the target is to identify the n^{th} dimensional decision hyperplane defined by Φ that maximizes the margin for predictions of critical and normal signal levels while correctly classifying them. Recognizing that the distance at any prediction instance t to the decision surface is given by $\frac{\gamma_t \Phi^T Y_t}{\|\Phi\|}$, the maximum margin solution is given by Equation (5.14)

$$\underset{\Phi}{\text{argmax}} \left\{ \frac{1}{\|\Phi\|} \min_t \left[\gamma_t \Phi^T Y_t \right] \right\} \tag{5.14}$$

The solution of this optimization problem would be very complex and so it is converted to an equivalent quadratic programming (QP) problem (*Bishop et al.*, 2006) as shown in Equations (5.15, 5.16)

$$\underset{\Phi}{\text{argmin}} \ \frac{1}{2}\|\Phi\|^2 \tag{5.15}$$

$$\text{s.t. } \gamma_t(\Phi^T Y_t) \geq 1, \ t = n, ..., N-k \tag{5.16}$$

In practice, however, the predictions of critical and normal signal levels might not be linearly separable in the space of auto-regressive inputs. To allow misclassified predictions, the objective function (5.15) needs to be modified to a soft support vector machine that maximizes the margin while softly penalizing the misclassified predictions as shown in Equation (5.17)

$$\underset{\Phi,\zeta_t}{\text{argmin}} \ \frac{1}{2}\|\Phi\|^2 + \alpha \sum_{t=n}^{N-k} \zeta_t \tag{5.17}$$

$$\text{s.t. } \gamma_t(\Phi^T Y_t) \ \geq \ 1 - \zeta_t, \ t = n, ..., N-k \tag{5.18}$$

$$\zeta_t \ \geq \ 0, \ t = n, .., N-k \tag{5.19}$$

where $\zeta_t,\ t = n, ..., N-k$ are slack variables introduced to relax the hard margin constraint and allow misclassified predictions such that $0 \leq \zeta_t \leq 1$ indicates a correct prediction at time t while $\zeta_t > 1$ indicates a misclassified prediction at this time instance. α is a regularization parameter that controls the trade-off between complexity of decision hyperplane and the number of misclassified predictions. Thus, Equation (5.17) can be solved to obtain the AR model that minimizes misclassified predictions (maximize ACC in Equation (4.7)). Since the critical levels of abnormality are generally rare events within the time series of patients, this solution will not address the ability to predict critical signal levels. To account for the statistical imbalance in this problem, the formulation was adjusted by adding different weights for both classes as shown in Equation (5.20)

$$\arg\min_{\Phi,\zeta_t} \frac{1}{2}\|\Phi\|^2 + \alpha\beta_+ \sum_{t:\gamma_t=1} \zeta_t + \alpha\beta_- \sum_{t:\gamma_t=-1} \zeta_t \tag{5.20}$$

$$\text{s.t. } \gamma_t(\Phi^T Y_t) \geq 1 - \zeta_t,\ t = n, ..., N-k \tag{5.21}$$

$$\zeta_t \geq 0,\ t = n, .., N-k \tag{5.22}$$

where β_+ and β_- are weighting parameters for the predictions of critical and normal signal levels respectively.

5.3.3 Optimizing Prediction Framework

The quadratic optimization problem in Equation (5.20) can be solved by constructing the Lagrange dual and solving the dual optimization problem to identify the predictive model. On the other hand, we need to optimally identify this model to maximize the ability to predict clinically relevant events. To do so, we first need to tune the objective function regularization parameter (α), class weights (β_+, β_-) as well as selecting the optimal model order (n).

The classical way of considering ACC (Equation (4.7)) as a performance objective would not lead to the maximum ability to predict the critical signal levels. Naturally, tuning the objective function to maximize the sensitivity would be the ideal solution. However, the fact that the four regions of the prediction grid are collectively exhaustive ones would indicate that maximizing performance over any region in the grid would affect performance over the other ones. Thus, we identified AR models by tuning the proposed model framework with respect to five different performance measures considering different regions of the prediction grid.

$M1$**: Maximizing Sensitivity (TPR)** This performance measure considers true positives (region A of the prediction grid) only to identify models with best ability to predict critical signal levels in the patient's time series.

$M2$**: Maximizing Precision (PPV)** This performance measure considers false positives (region B of the prediction grid) only and so leading to models with the fewest possible false predictions.

$M3$**: Combining Sensitivity and Specificity (BAC)** Maximizing the sensitivity alone would potentially affect the specificity and so we consider optimizing our models for maximum mathematical mean of sensitivity and specificity which is the Balanced Accuracy (BAC) (*García et al.*, 2009), defined by Equation (5.23).

$$BAC = \frac{TPR + TNR}{2} \tag{5.23}$$

$M4$**: Combining Sensitivity and Precision (F_μ-score)** The F_μ-score (*Powers*, 2011) was used to consider both true and false positives (regions A and B of the

prediction grid). For any $\mu > 0$, The F_μ-score can be expressed by Equation (5.24)

$$F_\mu = (1+\mu^2)\frac{TPR.PPV}{TPR+\mu^2 PPV} \tag{5.24}$$

where $\mu > 1$ weights sensitivity higher than precision while $\mu < 1$ emphasizes precision more than sensitivity. The selection of μ depends on the nature of application and how important is maximizing sensitivity of the predictions compared to their precision. In this dissertation, results for F_3-score are reported.

$M5$: Maximizing Area Under Receiver Operating Characteristic Curve (AUC) The ROC curve is a graphical plot used to illustrates the performance of a classifier as its classification threshold is varied (*Zweig and Campbell*, 1993). The curve is created by plotting the TPR against the false positive rate ($FPR = 100\% - TNR$) at various threshold settings.

5.3.4 k-Step Ahead Prediction Algorithm

To simplify tuning the objective function, we set $\beta_- = 1$ and β_+ to be proportional to $\frac{N_-}{N_+}$ where N_+ and N_- are the number of positive and negative labels respectively (critical and normal signal levels associated with y_{cr}). Using this approach enables cost-sensitive learning (*Tang et al.*, 2009) and Equation (5.20) can now be expressed as follows

$$\underset{\Phi,\zeta_t}{\operatorname{argmin}} \ \frac{1}{2}\|\Phi\|^2 + \alpha\eta\frac{N_-}{N_+}\sum_{t:\gamma_t=1}\zeta_t + \alpha\sum_{t:\gamma_t=-1}\zeta_t \tag{5.25}$$

$$\text{s.t. } \gamma_t(\Phi^T Y_t) \ \geq \ 1-\zeta_t, \ t = n, ..., N-k \tag{5.26}$$

$$\zeta_t \ \geq \ 0, \ t = n, .., N-k \tag{5.27}$$

where η controls the class weight for the critical predictions. The smallest allowable value $\eta_{min} = \frac{N_+}{N_-}$ will assign equal weights to both classes. $\eta \leq \eta_{min}$ were not considered since these will provide higher weights on predictions of normal signal levels. Note that this formulation assumes that critical signal levels exist within the data available for developing the prediction model.

The prediction performance of the proposed AR modeling approach (with respect to any of the performance measures) directly depends on the tunable parameters of the objective function (η and α). Very small α values likely generate underfitted models incapable of predicting critical signal levels while very high α values will likely generate overfitted models that can't generalize over independent data sets. Also, the weighting parameter η needs to be tuned with respect to the performance measure of interest. Larger η values are more likely to generate models with higher TPR than lower ones by giving higher weight to the predictions of critical signal levels. Unlike previous studies that select model order based on the error between predicted and reference signals, we consider selecting the model order with respect to the ability to predict critical signal levels in time series. Algorithm 4 was used for tuning α, η, n, as well as training and validating the proposed AR modeling framework. A nested cross validation scheme was used for selecting the model order n and the objective function tunable parameters (η and α). The algorithm starts by specifying the desired prediction horizon k as well as the range of search for n, η, and α. Selecting any of the performance measures ($Mi, i = 1, ...5$) as a training objective, the algorithm uses K-fold cross validation to select the parameters that maximize the performance with respect to the desired training objective. Finally, the final model is applied to an independent test set in which the predictions are evaluated with respect to all performance measures $M1, ...M5$.

Algorithm 4 AR models optimized for k-step ahead predictions of critical signal levels

Data: Training: TrData, Testing: TsDat
Input: k, prediction horizon
Input: $\alpha_{test} = \{\alpha_{min}, ..., \alpha_{max}\}$
Input: $\eta_{test} = \{\eta_{min}, ..., \eta_{max}\}$
Input: $n_{test} = \{n_{min}, ..., n_{max}\}$
Input: K for K-fold cross validation
Input: i for training performance $Mi, i = 1, ..., 5$
Output: Mi on TrDat
Output: $M1, M2, ..., M5$ on TsDat

Compute N_+, N_- from TrDat
for each $\alpha \in \alpha_{test}$ **do**
    for each $\eta \in \eta_{test}$ **do**
        for each $n \in n_{test}$ **do**
            for $L = 1$ to K **do**
                Hold out data in fold L for validation
                Solve Eq. (5.25) using TrDat $\setminus$ L
                Evaluate Mi on fold L (Mi_L)
            end for
            $Mi(\alpha, \eta, n) = \frac{\sum_{L=1}^{K} Mi_L}{K}$
        end for
    end for
end for
$\alpha_{Mi}, \eta_{Mi}, n_{Mi} = \text{argmax}_{\alpha,\eta,n}\, Mi$
Solve Eq. (5.25) with $(\alpha_{Mi}, \eta_{Mi}, n_{Mi}\ )$ using TrDat **return** Mi on TrDat & $M1, M2, ..., M5$ on TsDat

5.3.5 Case Study: Predicting Critical Desaturations in SpO_2 Time Series

We used SpO_2 and PR data (Section 3.1) for 119 postoperative adult patients generated by the pulse oximetry system. For each patient, $13,000$ points were used for training SpO_2 predictive models and separate $2,000$ points were used for reporting test results. We demonstrate results for $k = 20$s ahead predictions of critical desaturations characterized by the critical SpO_2 threshold $y_{cr} = 89\%$. To illustrate the results of the proposed prediction algorithm, 10 representative patients were selected with major critical events in both the training and test sets.

5.3.5.1 Prediction Framework Optimization Results

Using Algorithm 4, an exhaustive search was performed for the $3D$ space formed by the different parameters' choices in order to select the optimal ones with regard to the 5 statistical measures presented earlier. The range of α tested was $\{e^{-9}, ..., e^{+1}\}$ with 11 evenly spaced values in the log space between e^{-9} and e^{1} while the range of η tested was from $\eta_{min} = \frac{N_+}{N_-}$ to $\eta_{max} = 1$ with a step size of 0.05. The range of n tested was from 1 to 10 with a step size of 1. To make sure the models are trained sufficiently to tune these parameters, we used a 10-fold cross validation ($K = 10$).

Upon completion of cross validation, the set of parameters that maximize the prediction performance along each statistical measure are used to build predictive models with the whole training data. Finally, the models were evaluated on the test sets.

For the critical threshold $y_{cr} = 89\%$ with SpO_2 time series, our results show that the best ability to predict critical signal levels was achieved by maximizing over the $M1 = TPR$ and $M3 = BAC$ performance measures. Results for optimizing the models with each of the performance measures $M1, M2, ..., M5$ are presented hereafter followed by an in depth discussion for these results.

5.3.5.2 Maximal Sensitivity ($M1$) AR-Models

Table 5.4: Maximal TPR ($M1$) AR models

Optimal α, η, n for max Sensitivity AR models			
P#	$log(\alpha)$	η	n
12	1	0.81	9
18	1	0.91	6
21	1	0.86	5
63	1	0.80	9
86	1	0.92	9
93	1	0.82	6
94	1	0.97	9
97	1	0.86	7
111	1	0.95	9
113	1	0.99	9

The optimal set of parameters that maximize TPR for each patient are shown in Table 5.4. High α values were needed to increase the cost of missed predictions as well as high η values to increase the weights of the positive class significantly over the negative class to obtain the highest possible TPR.

Results for maximal TPR AR models are shown in Table 5.5 where the TPR performance is noticed to be excellent. TNR and BAC are generally high. The AUC values for the predictions are also excellent. However, the high TPR is accompanied with a decrease in PPV which also affects the F_3-score values.

Table 5.5: Training and Test Performance for Maximal Sensitivity ($M1$) AR Models

Maximal Sensitivity ($M1$) AR Models

P#	TrDat Performance	TsDat Performance					
	TPR	TPR	TNR	PPV	BAC	F_3-Score	AUC
12	**96.9%**	**98.2%**	51.1%	35.7%	74.6%	0.84	0.88
18	**100.0%**	**82.5%**	95.3%	42.3%	88.9%	0.75	0.94
21	**95.1%**	**93.1%**	96.5%	28.1%	94.8%	0.76	0.99
63	**100.0%**	**100.0%**	98.4%	11.4%	99.2%	0.56	1.00
86	**98.6%**	**100.0%**	87.5%	7.1%	93.8%	0.43	0.97
93	**96.8%**	**85.0%**	97.1%	23.0%	91.1%	0.67	0.96
94	**96.0%**	**94.3%**	94.0%	57.7%	94.2%	0.89	0.99
97	**92.5%**	**100.0%**	94.5%	6.0%	97.3%	0.39	0.99
111	**93.1%**	**100.0%**	98.8%	14.3%	99.4%	0.63	1.00
113	**88.8%**	**97.0%**	92.4%	40.2%	94.7%	0.85	0.98

5.3.5.3 Maximal Precision ($M2$) AR-Models

The optimal set of parameters that maximize PPV for each patient are shown in Table 5.4. Generally lower α values compared to maximal TPR models were needed which decreased the cost of missed predictions to identify these models. very low η values ($\approx \eta_{min}$) were needed to maximize the precision of the AR models indicating that highest precision occur with nearly equal class weights.

Results for maximal precision AR model are shown in Table 5.7. High PPV values are generally not achievable with a critical signal level $y_{cr} = 89\%$. Indeed, using PPV as a modeling metric would lead to models that ignore predictions of critical patterns similar to the cases of Patients Nos. $63, 86, 97, 111, 113$. Its worthy to be mentioned that higher PPV could have been obtained with $\eta < \eta_{min}$ but these will further decrease TPR. Nevertheless, TNR of the predictions are excellent and the AUC values are also high while low BAC and F_3-scores are noticed.

Table 5.6: Maximal PPV ($M2$) AR models

Optimal α, η, n for max-precision AR models			
P#	$log(\alpha)$	η	n
12	-1	0.06	7
18	-7	0.06	6
21	1	0.01	5
63	-6	0.05	6
86	0	0.07	6
93	1	0.02	7
94	1	0.02	6
97	-4	0.11	5
111	-4	0.05	8
113	-4	0.04	7

Table 5.7: Training and Test Performance for Maximal Precision ($M2$) AR Models

Maximal Precision ($M2$) AR Models

P#	TrDat Performance	TsDat Performance					
	PPV	TPR	TNR	PPV	BAC	F_3-Score	AUC
12	**50.0%**	15.0%	97.4%	**61.9%**	56.2%	0.16	0.85
18	**38.5%**	33.8%	99.3%	**65.9%**	66.5%	0.35	0.85
21	**95.0%**	13.8%	99.8%	**50.0%**	56.8%	0.15	0.98
63	**66.7%**	0.0%	100.0%	**NA**	50.0%	NA	0.79
86	**55.4%**	0.0%	99.9%	**NA**	50.0%	NA	0.98
93	**93.3%**	65.0%	99.9%	**86.7%**	82.4%	0.67	0.95
94	**88.2%**	3.8%	99.9%	**75.0%**	51.8%	0.04	0.98
97	**50.0%**	0.0%	100.0%	**NA**	50.0%	NA	0.96
111	**66.7%**	0.0%	100.0%	**NA**	50.0%	NA	0.98
113	**84.2%**	0.0%	100.0%	**NA**	50.0%	NA	0.96

5.3.5.4 Maximal BAC ($M3$) AR-Models

The optimal set of parameters that maximize BAC for each patient are shown in Table 5.8. As shown in this table, optimal α values were similar to maximal TPR models but generally lower η values than maximal TPR models were needed to maximize BAC.

Results for maximal BAC AR models are shown in Table 5.9. High TPR and TNR are noticed and slightly improved PPV and F_3-score compared to maximal sensitivity AR models. High AUC values were also noticed with these models.

Table 5.8: Maximal BAC ($M3$) AR models

Optimal α, η, n for max BAC AR models			
P#	$log(\alpha)$	η	n
12	1	0.81	9
18	1	0.86	9
21	1	0.66	6
63	1	0.80	9
86	1	0.77	10
93	1	0.67	9
94	1	0.82	9
97	1	0.96	10
111	1	0.95	10
113	1	0.99	9

5.3.5.5 Maximal F_μ-score ($M4$) AR-models

Setting $\mu = 3$ generated models with generally better results on the sensitivity side of the predictions compared to lower μ values. The optimal set of parameters that maximize F_3-score for each patient are shown in Table 5.4. Optimal η values are higher than the corresponding ones of maximal PPV models and lower than those of the maximal TPR models.

Results for maximal F_3-score AR models are shown in Table 5.10. Of course, the prediction sensitivities are not as good as maximal TPR models but on the hand, using the F_3 score generally improved the precision. Its also noticed that F_μ with $\mu > 3$ might be needed for Patient Nos. $63, 96$ to improve TPR. The personalized modeling framework provides flexibility to select the best F_μ-score that would result in acceptable TPR and PPV.

Table 5.9: Training and Test Performance for Maximal BAC ($M3$) AR Models

Maximal BAC ($M3$) AR Models

P#	TrDat Performance	TsDat Performance					
	BAC	TPR	TNR	PPV	BAC	F_3-Score	AUC
12	**94.8%**	98.2%	51.1%	35.7%	**74.6%**	0.84	0.88
18	**98.6%**	81.3%	95.8%	44.5%	**88.5%**	0.75	0.92
21	**96.1%**	93.1%	97.8%	38.0%	**95.4%**	0.81	0.99
63	**98.7%**	100.0%	98.4%	11.4%	**99.2%**	0.56	1.00
86	**94.4%**	89.5%	88.5%	6.9%	**89.0%**	0.41	0.97
93	**95.6%**	85.0%	97.4%	24.6%	**91.2%**	0.68	0.96
94	**94.0%**	94.3%	94.4%	59.3%	**94.4%**	0.89	0.99
97	**93.1%**	100.0%	95.0%	6.5%	**97.5%**	0.41	0.99
111	**93.1%**	100.0%	98.9%	15.4%	**99.4%**	0.65	1.00
113	**92.3%**	97.0%	92.4%	40.2%	**94.7%**	0.85	0.98

Table 5.10: Training and Test Performance for Maximal F_3-score AR Models

Maximal F3-Score AR Models							
P#	TrDat Performance	TsDat Performance					
	F3-Score	TPR	TNR	PPV	BAC	**F3-Score**	AUROC
12	**0.58**	65.6%	85.4%	55.4%	75.5%	**0.64**	0.87
18	**0.79**	70.0%	97.0%	49.1%	83.5%	**0.67**	0.90
21	**0.80**	79.3%	99.5%	71.9%	89.4%	**0.78**	0.99
63	**0.76**	0.0%	100.0%	NA	50.0%	**NA**	1.00
86	**0.68**	84.2%	93.8%	11.5%	89.0%	**0.52**	0.98
93	**0.76**	80.0%	98.9%	42.1%	89.4%	**0.73**	0.95
94	**0.78**	91.2%	96.2%	67.4%	93.7%	**0.88**	0.99
97	**0.61**	28.6%	99.3%	12.5%	63.9%	**0.25**	0.99
111	**0.55**	100.0%	99.4%	25.0%	99.7%	**0.77**	1.00
113	**0.82**	91.0%	95.2%	50.0%	93.1%	**0.84**	0.98

Table 5.11: Maximal F_3-score ($M4$) AR models

Optimal α, η, n for max F_3-score AR models			
P#	$log(\alpha)$	η	n
12	1	0.11	8
18	0	0.11	8
21	1	0.16	6
63	1	0.05	6
86	1	0.32	10
93	1	0.17	10
94	1	0.32	10
97	1	0.16	8
111	1	0.20	8
113	1	0.59	9

5.3.5.6 Maximal AUC ($M5$) AR-Models

The optimal set of parameters that maximize AUC for each patient are shown in Table 5.13. Optimal η values are either very high similar to the corresponding ones in the maximal TPR models or very low and close to those of the maximal PPV models.

Results for maximal AUC models ares shown in Table 5.12. Excellent AUC values are noticed but the TPR values varied from excellent (100%) to very poor (0%). Of particular interest to note that the AUC as a modeling metric doesn't guarantee very good TPR and may result in models that completely ignore critical predictions. Nevertheless, the high AUC could be always be utilized to select a different classification threshold at which these models can perform with higher TPR while maintaining an acceptable FPR.

Table 5.12: Training and Test Performance for Maximal AUC ($M5$) AR Models

Maximal AUC ($M5$) AR Models

P#	TrDat Performance	TsDat Performance					
	AUC	TPR	TNR	PPV	BAC	F_3-Score	AUC
12	**0.98**	0.0%	100.0%	NA	50.0%	NA	**0.89**
18	**1.00**	81.3%	95.7%	43.9%	88.5%	0.75	**0.93**
21	**0.98**	93.1%	95.9%	25.2%	94.5%	0.73	**0.99**
63	**1.00**	100.0%	98.3%	10.5%	99.1%	0.54	**1.00**
86	**0.98**	0.0%	100.0%	NA	50.0%	NA	**0.99**
93	**0.99**	85.0%	97.2%	23.6%	91.1%	0.67	**0.96**
94	**0.98**	94.3%	93.8%	56.8%	94.1%	0.88	**0.99**
97	**0.99**	0.0%	100.0%	NA	50.0%	NA	**0.99**
111	**0.99**	0.0%	100.0%	NA	50.0%	NA	**1.00**
113	**0.97**	97.0%	92.9%	42.0%	95.0%	0.86	**0.98**

Table 5.13: Maximal AUC ($M5$) AR models

Optimal α, η, n for max AUC AR models			
P#	$log(\alpha)$	η	n
12	0	0.01	7
18	1	0.91	10
21	1	0.96	9
63	1	0.90	9
86	1	0.02	10
93	1	0.97	9
94	1	0.97	6
97	1	0.01	5
111	1	0.00	9
113	1	0.99	8

5.3.6 Discussion of Results

For the critical threshold $y_{cr} = 89\%$ with SpO_2 time series, our results show that the best ability to predict critical signal levels was achieved by maximizing over the TPR and BAC performance measures. Although there are differences between n values that maximize BAC and TPR, the most noticeable effect on these measures was through η values. Figure 5.6 shows η effect on the cross validation TPR, BAC, and AUC for Patient No. 21 while fixing n and α to the values that maximize each of these measures. Both TPR and BAC increase as η increases but the BAC is maximized at at a lower η value than TPR. In all patients, maintaining high TPR needed relatively high η values. The high AUC over all η values indicate excellent classification ability for the algorithm. Nevertheless, the non-significant change in this measure with η doesn't guarantee high TPR of the maximal AUC AR-model.

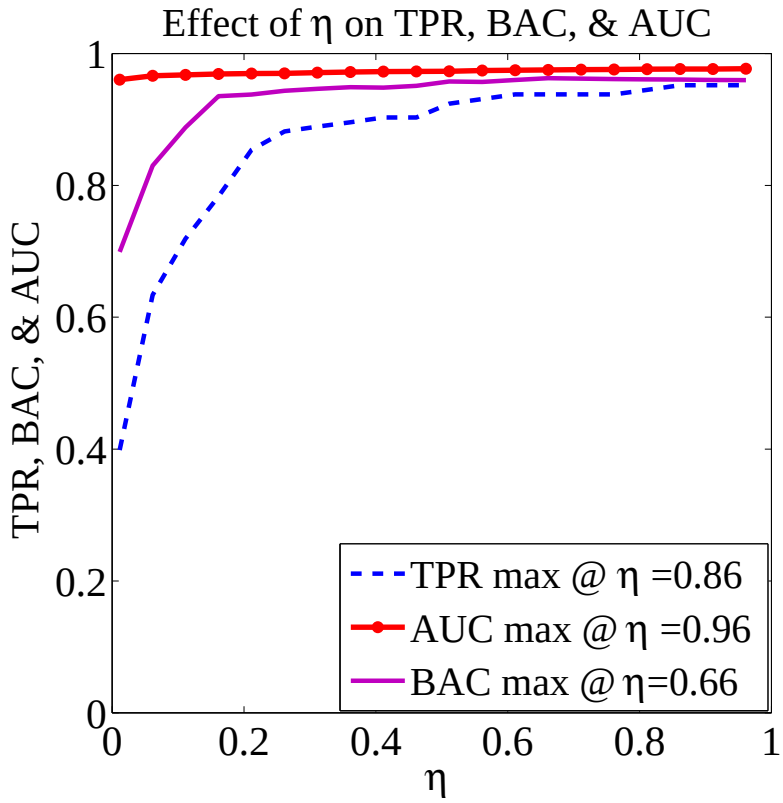


Figure 5.6: Cross validation performance measures vs. η. BAC is maximized at a smaller η value. No significant change in AUC as η changes.

The range of α, η, and n were sufficient to maximize over the different performance measures $M1, ..., M5$. For example, our results show that better prediction results couldn't be achieved with $\eta_{max} > 1$. On the other hand, avoiding $\eta < \frac{N_+}{N_-}$ guarantees excluding solutions that would give higher class weight for the normal predictions. For the range of α, we didn't allow $\alpha > e^1$ to avoid potentially overfitted models. Also, our results show that high model orders ($n > 10$) may not be needed for short term predictions of critical patterns in SpO_2 time series. This is also favorable since higher order models are more easily affected by noise in measurements.

The main computational effort in the algorithm is in using cross validation to find the optimal set of parameters α, η, and n. Indeed, this is comparable to standard AR models where similar techniques are used to find model order and regularization parameter in Equations (4.8, 4.9). Moreover, since the most noticeable effect on different statistical measures is through η, we can improve the computational efficiency of this algorithm by setting α and n values a priori and then modify the algorithm to select the optimal η values that maximize any of the performance measures. This would be a slightly suboptimal approach but with an accurate tuning for η, a sufficiently good performance could be obtained. Once these parameters are set, the model can be trained and used for online prediction efficiently.

Of particular interest is the dependence of the prediction results on y_{cr}. Setting this threshold to very low/ high values will generally decrease the likelihood to see events associated with it. Thus, higher η values will be needed which also affects PPV. Although the predictions obtained for less extreme thresholds might have higher PPV, this improvement should be evaluated with respect to the clinical significance of predicting levels associated with the less extreme thresholds. Another key factor that would influence the selection of y_{cr} is the patient's health state. The choice to fit individualized models would enable them to be tailored to patient needs with respect to setting y_{cr} as well as the performance measure needed for model

optimization.

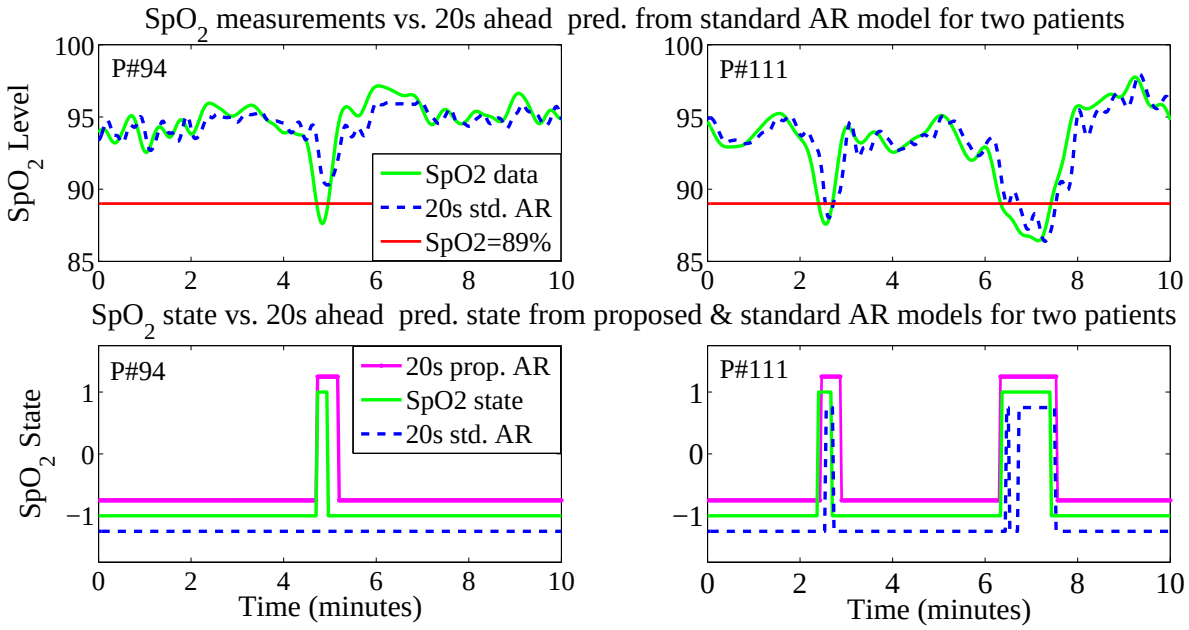


Figure 5.7: Comparison between 20s ahead predictions using standard and maximal BAC AR models

The selection of y_{cr} affects not only the performance measure needed for fitting the model, but also the amount of data needed to build it. Extremely low/ high thresholds will require monitoring patients for longer times in order to incorporate rare occurrences, associated with the selected y_{cr}, for building the models. Thus, the personalized predictive modeling framework could be used to build models at earlier monitoring stages (with less data and using less extreme thresholds). Then, these models could be gradually adapted with more extreme thresholds depending on the patient state after longer monitoring times.

Overall, the statistical measures presented show the tradeoff between prediction results obtained through optimizing different regions (or combination of regions) of the prediction grid in Figure 5.4. The selection of any of these measures depends on the clinical application, nature of the physiological signal, and the critical signal level desired for building the models.

5.3.7 Comparison to Standard AR Models

For a standard k-step ahead AR model, the k-step ahead predicted signal $\hat{y}_{t+k}$ at time $t+k$ is inferred as a linear combination of previously observed signals y_{t-i+1} up to time t

$$\hat{y}_{t+k} = \sum_{i=1}^{n} \phi_i y_{t-i+1} \tag{5.28}$$

where $\Phi = [\phi_1, ..., \phi_n]^T$ denotes the vector of AR coefficients to be determined, n denotes the order of the model.

Considering N sampled data points of a time series, the coefficients Φ in Equation (5.28) are estimated using regularized least squares fit (*Ljung*, 1987; *Bishop et al.*, 2006). A regularization parameter λ is typically used to decrease model complexity and allow better generalization over independent data sets (*Bishop et al.*, 2006). For each patient time series, the model order n and λ are tuned optimally using cross validation (*Bishop et al.*, 2006; *Geisser*, 1993).

For comparison with the proposed modeling framework, individualized standard AR models were identified for each patient using $13,000$ data points and the models were evaluated over 2000 points of the SpO_2 time series of each patient. The data sets are identical to the ones used in evaluating the proposed algorithm. Models with order $n \in \{1, 2, 3, ..., 10\}$ and $\lambda \in \{e^{-1}, e^{0}, e^{1}\}$ were tested to select the optimal set of parameters for the minimum $RMSE$ using cross validation. Our results show that no significant improvement could be obtained with $n > 10$ or $\lambda < e^{-1}$.

Figure 5.7 shows a comparison between the performance of the 20s ahead predictions of the maximal BAC AR model and the standard LSE AR model for two patients (P#94,111). The proposed algorithm shows a better ability in predicting critical levels than the standard AR model while maintaining an excellent capability in predicting regions of normal signal levels. For P#94, the proposed AR model was able to predict the critical SpO_2 levels (without any phase lag) while the standard

AR model failed to predict them. For the first critical desaturation epoch shown for P#111, the 20s prediction for the first occurrence of a critical signal level is delayed by 4s for the proposed model compared to 10s of delay for the standard AR model. For the second critical epoch, the 20s ahead prediction of the proposed AR model captured the onset of the critical epoch without any phase lag compared to 6s of delay for the standard AR model. The standard AR model fails to predict the critical levels for 8s after that before successfully predicting the remaining critical level instances in the second critical desaturation epoch.

Table 5.14: Training and Test Performance for LSE AR Models

Standard LSE AR Models							
P#	TrDat Performance	TsDat Performance					
	$RMSE$	TPR	TNR	PPV	BAC	F_3-score	$RMSE$
12	**1.14**	13.3%	98.8%	75.3%	56.1%	0.15	**1.17**
18	**0.74**	66.3%	98.9%	70.7%	82.6%	0.67	**0.76**
21	**0.96**	51.7%	99.8%	78.9%	75.8%	0.54	**0.99**
63	**0.89**	0.0%	100.0%	NA	50.0%	NA	**0.91**
86	**0.71**	0.0%	99.5%	0.0%	49.8%	NA	**0.73**
93	**1.18**	75.0%	99.7%	71.4%	87.3%	0.75	**1.21**
94	**0.73**	63.5%	98.9%	83.5%	81.2%	0.65	**0.74**
97	**1.10**	0.0%	100.0%	NA	50.0%	NA	**1.13**
111	**1.03**	0.0%	100.0%	NA	50.0%	NA	**1.05**
113	**0.98**	36.0%	98.9%	63.2%	67.4%	0.38	**1.01**

To evaluate the ability of the standard AR model to predict the critical signal levels in the patients' time series, the proposed k-step ahead prediction metric was applied to the predictions of the standard AR model and the results are shown in Table 5.14. The proposed algorithm outperforms the standard AR model in predicting

critical desaturation levels. Prediction sensitivity for maximal TPR and maximal BAC AR models are significantly better than standard AR models ($p < 0.0001$). This significant improvement in TPR is found in two main aspects. First, the proposed algorithm is able to predict critical signal levels that are not predicted by the standard model. Second, a negligible (if any) lag in the predictions of critical signal levels by the proposed algorithm is noticed compared to the more pronounced phase lag of the standard AR model predictions.

On the other hand, the standard AR model shows a superior TNR compared to the proposed model indicating higher ability to predict normal regions than the critical signal levels. Minimizing the mean square error in the standard AR model restricts its ability to predict large or sudden deviations from the mean of the signal ($y \leq y_{cr} = 89$). Interestingly, with the significantly improved TPR, the proposed algorithm maintained a generally high TNR. The competing nature of the TPR and PPV objectives caused the standard AR model to show higher PPV compared to the proposed algorithm. In fact, our results show that the prediction performance of the standard AR model is very close to the maximal PPV AR model in the presented framework.

5.3.8 Summary

This section presented a comprehensive framework for multi-step ahead predictions of critical levels in physiological signals. A performance metric for evaluating k-step ahead predictions was proposed and used to build a prediction framework for auto-regressive models capable of predicting regions of clinical interest in physiological signals. Using the proposed algorithm, a significant improvement was achieved in the ability to predict critical signal levels compared to standard auto-regressive models.

The development of this framework enables more interaction with clinicians to get

better insight into the prediction of specific critical patterns. Also, the development of such high fidelity predictive models for critical signal levels is a fundamental step for designing novel interventional systems that can automatically apply necessary therapeutic treatment to prevent adverse outcomes associated with these critical levels. Clinical evaluation for the prediction performance at different signal thresholds will enable selecting the ones that can improve the performance and maximize the clinical use of the prediction algorithm.

CHAPTER VI

Conclusions and Future Work

6.1 Conclusions

The objective of this dissertation is to develop a research framework for both evaluating and optimizing predictive models of physiological time series. Moving from monitoring to predicting physiological signals introduces research challenges in designing models that are able to make accurate near-term predictions for clinically relevant events as well as evaluating predictions with respect to these events. The main contributions can be divided into three categories.

6.1.1 Performance Metrics for Evaluating Predictions of Physiological Signals

Inspired by the clinical preference of predicting regions of clinical interest in physiological signalise, several performance metrics were developed and applied to clinical data sets.

Chapter IV considered characterizing performance over prediction windows. This metric addresses applications where it is a greater clinical interest to predict the occurrence of clinically relevant events within a specific prediction widow rather than the exact instances at which these event will occur. A clinically relevant event is

physiologically defined by breaching a critical threshold for a predefined minimum duration.

Of particular interest in clinical applications is the ability to predict clinically relevant events as early as possible. Recognizing the high importance of this aspect in a clinical setting, Chapter IV develops another metric for characterizing the performance of a time series model of a physiological signal using the longest horizon that can predict clinical events of interest in this signal. The analysis of the predictions obtained by any predictive model results in a statistical distribution that quantifies the overall predictive power of this model with respect to this metric.

Chapter V considers evaluating predictions of clinically relevant events with respect to a fixed prediction horizon. Most importantly, this metric characterizes the performance in capturing the onset of clinically relevant events over the horizon of interest. Two main types of clinically relevant events are defined in the clinical literature. Thus, Chapter V considers two methods for defining fixed horizon metrics. First, a metric is considered for predictions of abnormal deviations from the signal baseline. Second, a metric is developed for predictions of critical levels of abnormality.

Chapter IV applies the developed metrics for a case study of characterizing the performance of standard auto-regressive models in predicting critical oxygen saturation levels in the blood. Using the metric of the largest horizon that can predict clinically relevant events, it was shown that 56.2% of the critical desaturation events in the time series of patients were able to be predicted 10-60s ahead of time using the standard auto-regressive model. Although prediction performance shows excellent ability to predict normal signal levels, the standard model has a limited ability to predict critical signal levels. Analysis with the fixed horizon metric, discussed in Chapter V, shows that predictions of critical signal levels over a fixed horizon of 20s are associated with a significant phase lag that limits the performance over this horizon.

6.1.2 Incorporating Directionality into Predictions of Physiological Signals

Chapter IV considers a dynamic systems perspective to understand the directionality between any two physiological signals A and B. A dynamic model is developed in which A is an input to B and another dynamic model is developed with B as an input to A. The significance of the coefficients for the two modeling choices was investigated to understand the directionality between these physiological signals.

For any physiological signal of interest, the goal is to incorporate additional data channels that improve the ability to predict regions of clinical relevance in that signal. Two metrics are defined in Chapter IV to characterize the improvement in predicting regions of clinical interest with multi-channel data:

1. Evaluate the improvement in prediction sensitivity TPR of critical signal levels over different prediction windows due to the additional data channels.

2. Determine whether the inclusion of the additional data channels allows prediction of critical signal levels with a longer prediction horizon (improve the earliest time at which critical signal levels can be predicted).

A case study was considered for investigating the cause-effect relationship between pulse rate (PR) and blood oxygenation (SpO_2) dynamics. No significant improvement was noticed in the sensitivity (TPR) of 20s and 60s predictions of critical desaturation levels in the blood. Also, including PR data in the SpO_2 predictive models didn't increase horizons that predict critical desaturation events. Results indicate oxygen in blood is an effective input to the pulse rate rather than vice versa.

6.1.3 Novel Models for Predicting Regions of Clinical Interest in Physiological Signals

Characterizing the performance of standard models of physiological signals revealed the need to design new models optimized to capture clinically relevant events in these signals. A framework for optimizing time series models with respect to predictions of clinically relevant events was developed. Two types of clinically defined regions were considered.

6.1.3.1 Predicting Abnormal Deviations from Signal Baseline

A mixed integer programming framework was used for optimizing models in this problem. A cost minimization optimization problem was formulated to minimize false negatives FN in the k-step ahead dynamically adjusted threshold metric in Table 5.1. Due to the dependence of the optimization problem setup on the the clinical definition for the baseline of the signal, this problem was presented with a case study of predicting abnormal deviations from SpO_2 baseline. The clinical literature was used to define both SpO_2 baseline and abnormal deviations from this baseline.

Models that minimize FN (maximizing TPR) showed significant improvement in the ability to predict 20s ahead abnormal SpO_2 levels. Nevertheless, the improvement in TPR was accompanied (in some cases) with magnifying predictions in regions of abnormal signal levels. As a result, the improvement in TPR was associated with a decrease in PPV. The optimization problem setup was designed to reject the trivial solution that always predicts an abnormal deviation (regardless the model input) through the design of the objective function and constraints. Incorporating FP explicitly in the model optimization resulted in models that continuously predict the mean of the signal to maintain the lowest FP.

6.1.3.2 Predicting Critical Signal Levels

In the second problem, models optimized with respect to predictions of critical signal levels were presented. A quadratic programming (support vector machine) framework was used to develop models. To avoid problems associated with magnifying predictions in regions of critical signal levels, the models were defined to directly predict the physiological state associated with the clinical definition of the critical signal level. Note that signal levels in the second problems are much more extreme ones than abnormal deviations from baseline. Thus, to address the issue of the relative paucity of such events in physiological data, different statistical metrics are considered for optimization. Considering the prediction metric in Figure 5.3, the models were optimized for sensitivity (TPR), precision (PPV), balanced accuracy (BAC), F-score, and AUC. The selection of any of these statistical metrics depends on the clinical application, nature of the physiological signal, and the critical signal level desired for building the models.

A case study was presented for developing an auto-regressive modeling framework to predict critical desaturation events in the SpO_2 time series. For the critical threshold $y_{cr} = 89\%$ with SpO_2 time series, our results show that the best ability to predict critical signal levels was achieved by maximizing over the TPR and BAC performance measures. Significant improvement in the ability to predict critical signal levels was achieved compared to standard AR models. The improvement in TPR was found in two main aspects. First, the proposed modeling framework is able to predict extreme levels when predictions of the standard model are not low enough to capture them. Second, a negligible phase lag in the predictions of critical signal levels is noticed compared to the more pronounced phase lag of the standard AR model predictions.

6.2 Next Step: Developing Time Series Models for Predicting Sleep Apnea

6.2.1 Sleep Apnea

Sleep apnea is a sleep disorder that is defined as a transient reduction or complete cessation of breathing during sleep (*Quan et al.*, 1999). The *American Academy of Sleep Medicine* (2007) defines the criteria for scoring a sleep apneic event by a clear decrease ($\geq 90\%$) in the air flow from the corresponding baseline for a duration $\geq$ 10s. The airflow baseline is defined as the mean amplitude of stable breathing and oxygenation in a 2-min window or the mean amplitude of the three largest breaths in case of unstable breathing (*Mendez et al.*, 2010). Clinicians usually divide sleep apnea into three major categories: obstructive, central, and mixed apnea (*American Academy of Sleep Medicine*, 2007; *De Chazal et al.*, 2003). Obstructive sleep apnea (OSA) is characterized by intermittent pauses in breathing during sleep caused by the obstruction or collapse of the upper airway. The airway is blocked at the level of the tongue or soft palate preventing air from entering the lungs in spite of continued efforts to breathe. Central sleep apnea (CSA) is a neurological condition which causes the loss of all respiratory effort during sleep. With CSA, the airway is not necessarily obstructed. Mixed sleep apnea combines components of both CSA and OSA, where an initial failure in breathing efforts allows the upper airway to collapse. Sleep apneic events are typically followed by a significant reduction in blood oxygen saturation.

Standard scoring for sleep apnea is carried out by an expert sleep clinician in a dedicated sleep lab where concerned patients undergo an overnight polysomnography (PSG). Due to the cost and relative scarcity of diagnostic sleep laboratories, it is estimated that sleep apnea is widely under diagnosed (*Young et al.*, 1997). Hence, several techniques that aim to provide a simple detection of sleep apnea with fewer and simpler measurements and without the need for specialized sleep labs have been

proposed (*Mendez et al.*, 2010; *De Chazal et al.*, 2003). Majority of studies in this area focused on using the electrocardiogram (ECG) signal since it is highly influenced during the apneic events and can be easily measured in a non-invasive way and with high signal-to-noise ratio even in a nonclinical environment (*Penzel et al.*, 2002).

Although considerable attention has been given to apnea detection methods, prediction of an impending apnea episode has been rarely reported in the literature. For example, (*Dagum and Galper*, 1995) and (*Bock et al.*, 1998) use limited data from OSA patients to to predict 1s ahead OSA espisodes. A recent study by *Le et al.* (2013) uses Dirichlet process based Gaussian process mixture (DPMG) model to predict the evolution of two signal features derived from the ECG signal. Univariate time series models are used for generating 1-3 minutes ahead predictions of each of these signals on a minute by minute basis. Then, the predictions are fed to an offline support vector machine algorithm trained against expert apnea annotations to classify if these predictions correspond to an apneic event or not. No studies were found on developing multi-channel time series models with predicting apneas as an objective for model optimization.

6.2.2 Polysomnography (PSG) Data

This data set is composed of 100 patients. For these patients, we have full polysomnography (PSG) data that were recorded at the University of Michigan Sleep Lab. The PSG data for each patient consists of 21 signal channels. In addition to the SpO_2 and PR recorded by the POM systems, we have other signals collected during the overnight stay for the patient in sleep lab. The available signal channels along with their sampling frequencies are listed in Table 6.1.

The PSG monitors human body functions during sleeping period (usually at night). The activities of brain and muscles are monitored through EEG and EMG respectively. The heart rhythm is monitored through ECG as well as PR while eye

Table 6.1: Signals available from PSG data

No.	Signal Label	Signal Name	No. of Channels	Sampling (Hz)
1	EOG	Electrooculogram	2	256
2	ECG	Electrocardiogram	3	256
3	EMG	Electromyogram	3	256
4	EEG	Electroencephalography	6	256
5	PLETH	Plethysmography	3	256
6	SNORE	Snore	1	256
7	NPRE	Nasal Pressure	1	32
8	NO	Nasal/Oral Air Flow	1	32
9	SpO_2	Blood Oxygenation	1	16
10	PR	Pulse Rate	1	16

movement is monitored through EOG. The breathing functions are monitored by NO, NPRE, and PLETH in addition to SpO_2. Nasal pressure and Nasal/Oral airflow (NPRE and NO), typically measured using transducers fitted in or near the nostrils, are used to measure the rate of respiration and to identify interruptions in breathing (*Iber*, 2007). The plethysmography has 3 signal channels. Two of them monitor the the movement of the chest and abdominal walls to evaluate pulmonary ventilation (*Konno and Mead*, 1967). The third channel, often obtained from pulse oximeters, measures the changes in blood volume in the skin. The plethysmography is used to monitor respiration, heart rate, and cardiac cycle (*Shamir et al.*, 1999; *Shelley et al.*, 2006)

The patients in this data set have a medical diagnosis for their disease states. They are diagnosed with obstructive sleep-apnea, central sleep-apnea, or normal state.

6.2.2.1 ECG Signals

The preprocessing of the ECG signal includes bandpass filtering to remove noise and retain the clinical features in the QRS complexes of the ECG signals. Following the University of Michigan Sleep Lab settings, and to minimize the effect of physiological and measurement artefact in the ECG signals, the differential signal $ECG12 = ECG1 - ECG2$ was considered. The ECG12 signal was filtered by applying a pass-band filter between $0.3 - 70$ Hz prior to applying QRS detection algorithm for extracting beat to beat intervals. The filtered ECG12 signal was used in subsequent analysis.

6.2.2.2 RR Time Series

QRS Detection We used an open source QRS detection software (*Niskanen et al.*, 2004) for annotating the R-wave occurrences in the ECG signals. Careful editing and visual inspection of the ECG signal helped to eliminate sources of errors arising from missing QRS complexes or spurious QRS detections.

After the QRS complex occurrence times have been estimated, the RR (inter-beat) intervals are obtained as the differences between successive R-wave occurrence times as shown in Figure 6.1. The n^{th} RR interval is obtained as the difference between the R-wave occurrence times $RR_n = t_n - t_{n-1}$. Accordingly, the time series of available RR intervals (t_n, RR_n) was constructed.

RR Intervals Correction Due to poor signal to noise ratio and errors in the automatically generated QRS detections, the time series of RR-intervals contained physiologically unreasonable times. The following algorithm was used to generate a corrected sequence of RR-intervals with all intervals physiologically reasonable (*Chen et al.*, 2015; *De Chazal et al.*, 2003).

A median filter of width 5 was applied to the sequence of RR-intervals to find

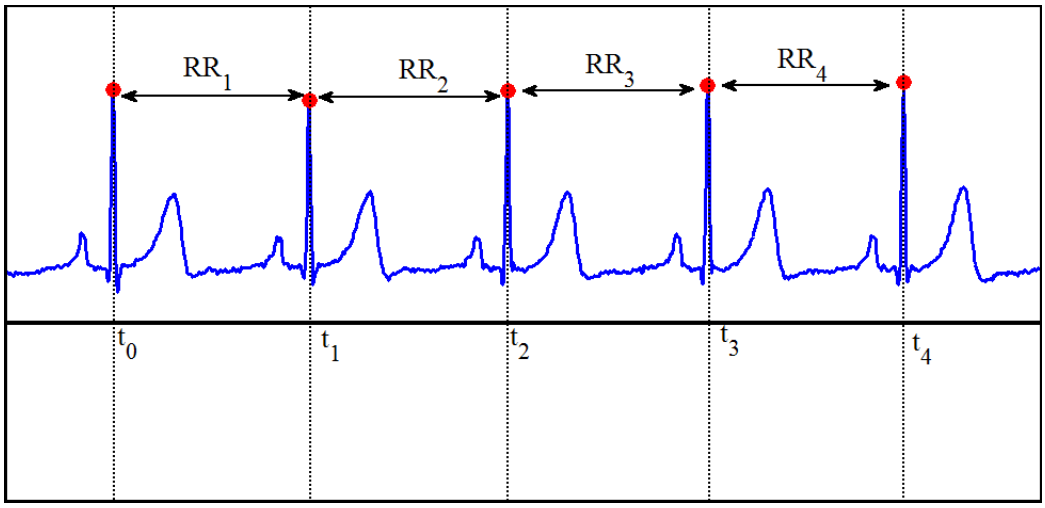


Figure 6.1: Extracting RR intervals. The n^{th} RR interval is obtained as the difference between the R-wave occurrence times $RR_n = t_n - t_{n-1}$.

suspect RR intervals. The filtered signal provided a robust estimate for the expected value for each RR interval. Any RR interval that is significantly different from the robust estimate was considered a suspect RR-interval which could be either spurious QRS detections, or missed QRS complexes.

Spurious detections were found by comparing the sum of adjacent RR-intervals with the robust RR-interval estimate. If this sum was numerically closer to the robust estimate than either of the individual RR-intervals, then a spurious detection was considered to be present. The two RR-intervals were merged to form a single RR-interval.

If the RR-interval was a factor of 1.8 times or greater than the robust estimate then this was considered as an indication that one or more QRS complexes were missed. To estimate the times of the missing QRS complexes, the RR-interval was divided by the sequence of integers $2, 3, 4, ...$ until it best matched the robust estimate of the RR-interval. The single RR-interval was then subdivided by the appropriate integer to form a series of new detections that were placed in the missed locations.

6.2.2.3 ECG Derived Respiratory (EDR) and R-Wave Amplitude Signals

During the breathing cycle, the ECG signal is affected by the variations in the relative distance between the electrode located on the chest surface and the heart, and by the changes in the thoracic impedance produced by the inflation and deflation of the lungs (*Mendez et al.*, 2009). The effect is most obviously seen as a slow modulation of the ECG amplitude at the same frequency as the breathing cycle (*Moody et al.*, 1985; *Travaglini et al.*, 1998). To access the EDR signal, the ECG signal was filtered with a median filter of 200ms for ECG baseline estimation (*Mendez et al.*, 2009; *Sörnmo and Laguna*, 2005). Then, the resulting signal was subtracted from the ECG signal to produce the baseline corrected ECG. Each sample point of the EDR signal was then obtained by calculating the area enclosed by the baseline corrected ECG in a temporal window of 100ms before and after the R-peak value of the corresponding QRS complex.

The baseline corrected ECG signal was also used to extract the time series of R-wave amplitudes using the peaks of successive QRS complexes.

6.2.3 Research Plan

For this problem, we plan to develop a multi-channel time series model for sleep apnea prediction. The primary source for detecting the occurrence of a sleep apneic event is the air flow signal (NO) (*American Academy of Sleep Medicine*, 2007). In addition to air flow, the proposed model will include ECG and SpO_2 as inputs to the prediction algorithm.

The first step is to define a robust performance metric for characterizing sleep apnea. Several algorithms will be evaluated to characterize airflow changes during apnea. After defining a relevant metric, we will formulate an optimization problem with respect to predictions of sleep apnea. In this problem, we will design the proposed model to operate over prediction windows of fixed length. The rationale behind this

selection is that the reduction in airflow needs to last for at least 10s during an apneic event. Performance of the proposed model with respect to predicting sleep apnea will be evaluated using the PSG data set.

6.3 Future Work

Future extensions to the work presented in this dissertation include new methods and applications for defining prediction performance metrics for physiological signals as well as continuing to develop predictive models that broaden the approach with new problems and different signals.

6.3.1 Prediction Performance Metrics

6.3.1.1 Engineering Metrics for Evaluating Physiological Predictions

The first step in the process of developing high fidelity predictive dynamical models of physiological signals is to define engineering metrics that are able to accurately characterize the prediction performance in capturing clinically relevant events. Current research in this area lacks the presence of robust consistent definitions for metrics that can be applied to prediction.

One major limitation is that the clinical literature has defined many critical patterns using different published standards (*Ruehland et al.*, 2009). Moreover, many clinically relevant events are defined in a way that is geared towards identifying them by an "expert eye" instead of accurately characterizing the underlying changes in the physiological signal(s) of interest during the occurrence of such events. For example, the gold standard for scoring sleep apnea and other sleep disorders are the manual annotations provided retrospectively by sleep lab technicians (*American Academy of Sleep Medicine*, 2007). A study has shown that there is a significant intraobserver and interobserver variability when used to identify these events (*Whitney et al.*, 1998). In

fact, expert evaluation and scoring for clinically relevant patterns is not only affected by human error, but also doesn't provide enough insight into the actual dynamical changes of the corresponding signal during such events.

To enable accurate predictions of physiological signals, future research is needed to fill this gap by developing performance metrics that can accurately characterize changes in physiological signals during occurrence of critical events. Defining performance metrics that can be used for online prediction is of great clinical value.

6.3.1.2 Probabilistic Metrics for Evaluating Predictions

In physiological signals, clinicians are interested in maximizing the ability to predict clinically relevant events over a specific prediction horizon of interest. In Chapter V, we considered a fixed horizon prediction metric and used it for developing models that optimize this metric with respect to the ability to predict clinically relevant events. We showed that this approach generates models with much improved performance in predicting the onset of such critical events.

Future research includes investigating probabilistic metrics for predicting clinically relevant events. For example, probabilistic information in the predictive distribution of k-step ahead models can be used to improve the ability to predict critical signal levels. Figure 6.2 illustrates a common case that could occur in threshold based evaluation for k-step ahead predictions when $\hat{y}_{t+k}$ is slightly higher than y_{cr}. A phase lag of δ time steps is noticed in predicting the onset of the critical event ($\hat{y}_{(t+\delta)+k} \leq y_{cr}$). Future research includes investigating approaches for generating robust probabilistic predictions of critical signal levels when expected values $\hat{y}_{t+k}$ are not able to capture them. Sequential Monte Carlo sampling methods can be used to generate these probabilistic predictions when no analytical form exists for the probability distribution of the k-step ahead model.

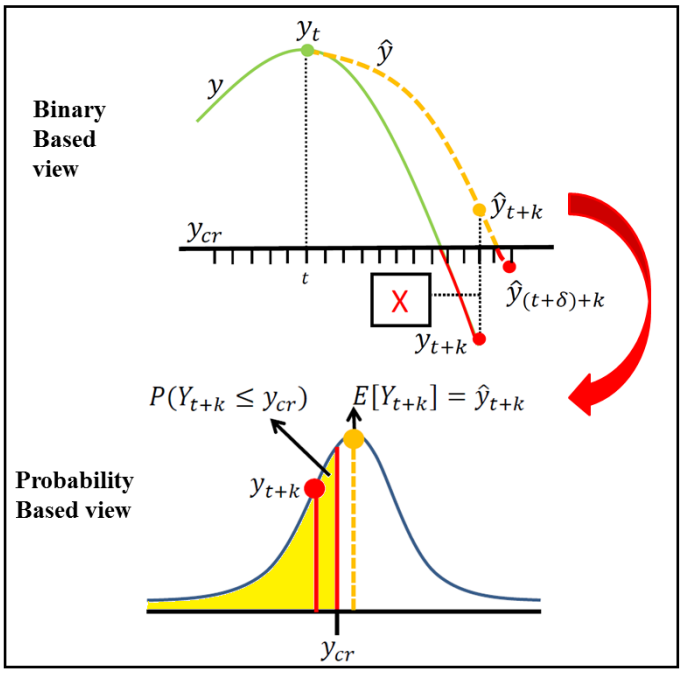


Figure 6.2: Binary vs. probabilistic based evaluation for k-step ahead predictions.

6.3.2 Models to Predict Clinically Relevant Events

6.3.2.1 Optimizing Models over Prediction Windows

In Chapter V, we identified models with respect to a fixed prediction horizon of k steps driven by the objective to maximize the ability to detect the onset of the clinically relevant events with respect to to this horizon. In some physiological signals, clinically relevant events only occur after breaching the critical signal threshold for a minimum duration of time. For example, sleep apnea is defined with a significant reduction (more than 90%) in the baseline airflow for at least 10s. Thus, it would be more convenient to consider performance over prediction windows in such cases.

Future work includes developing predictive models that operate over prediction windows. New objective functions need to considered while maintaining convexity of the optimization problem. In terms of prediction strategy, direct and multi input multi output (MIMO) strategies can be evaluated. The latter method has fewer parameters to fit but is also less flexible than the direct prediction strategy. Future

research is needed to address tradeoffs between these two methods with respect to the ability to predict clinically relevant events over the window of interest. Objective function for fitting these predictive models will consider maximizing the area between the critical threshold and signal levels below this threshold over the window of interest. Lessons learned from Chapter V can be incorporated for metrics that need to be used for optimization with rare events.

6.3.2.2 Multi-channel Dynamical Models

Multi Input Single Output (MISO) Models The developed models in Chapter V are single channel predictive models. The work can be extended to develop multi input models optimized for predicting critical signal levels within a physiological signal of interest. The metrics in Chapter IV can be used to evaluate and select channels to be incorporated in the prediction algorithm. Future research includes selecting the smallest possible set of physiological signals that maximizes the ability to predict clinically relevant events. Multiple time series will be derived from each channel to maximize the ability to predict clinically relevant events within the time series of interest. For example, RR time series, R-wave time series, and ECG derived respiratory signal (EDR) can be derived from ECG. Also, non-linear techniques representing measures of complexity in physiological time series can be used to improve the prediction algorithms.

Multi Input Multi Output (MIMO) Models Predicting regions of clinical relevance assumes direct access to the corresponding physiological signal in order to generate future predictions based on the observed history of this signal. In practice, some physiological signals are known to be invasive, uncomfortable and not very reliable. For example, Nasal Pressure and Nasal/Oral Air Flow (NPRE and NO) that are used to identify interruptions in breathing during sleep are typically measured

using sensors fitted in or near the nostrils causing these sensors themselves to disturb sleep.

Future research includes developing multi input multi output (MIMO) models to predict critical events of interest assuming no access to the corresponding physiological signal. Prediction windows over which the clinically relevant event under study occurs need to be considered. Metrics can then be defined in terms of signal changes in the other data channels during the occurrence of the critical event of interest. For each data channel to be included in the model, a metric needs to be defined to characterize the changes in the corresponding signal during the occurrence of the event of interest. Since each of the derived metrics represents the effect of the critical event of interest on the corresponding data channel, an objective function that considers these metrics together is needed to optimize a MIMO model that uses the desired data channels to predict the occurrence of the critical event as predicted by the metrics defined for each signal.

With continued research, new metrics and models will be developed for prediction of physiological signals. Using recorded physiological time series, relevant engineering metrics can be developed and used to create predictive models. Clinical assessment for the predictions can then be used to adapt or update the modeling methods or even refine the metrics. Figure 6.3 shows a schematic diagram of what it is anticipated for the future of physiological signals' prediction. The presented framework clearly motivates synergy between engineering and clinical research communities. The research approach in this dissertation has the potential to be applied to a wide range of the vast amounts of data that is currently being gathered. The broader impact of this research is the potential for improved quality of health care for monitored patients in hospitals.

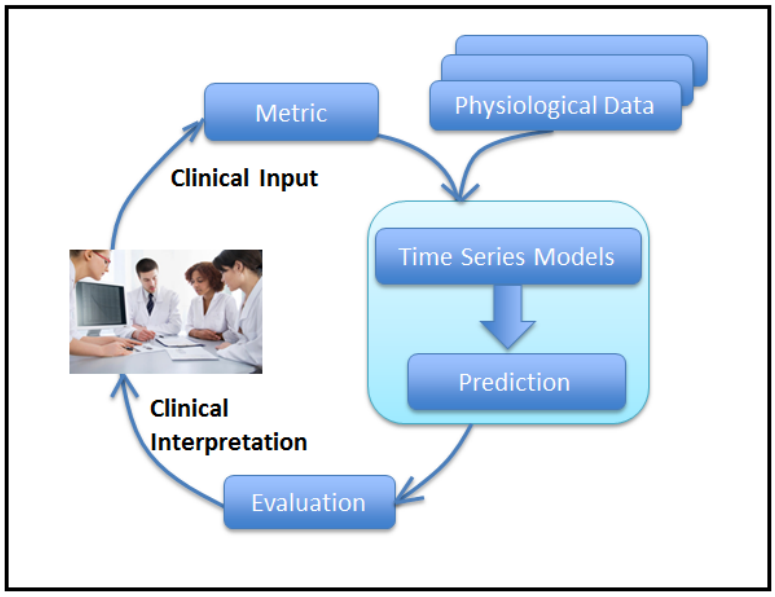


Figure 6.3: Future for physiological signals' prediction. Using recorded physiological time series, relevant engineering metrics can be developed and used to create novel predictive models. Clinical assessment for the predictions can then be used to adapt or update the modeling methods or even refine the metrics to improve and optimize the prediction performance with respect to the clinical events of interest.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahmed, N. K., A. F. Atiya, N. E. Gayar, and H. El-Shishiny (2010), An empirical comparison of machine learning models for time series forecasting, *Econometric Reviews*, *29*(5-6), 594–621.

Alpaydin, E. (2014), *Introduction to machine learning*, MIT press.

Altman, D. G., and J. M. Bland (1994a), Diagnostic tests 2: Predictive values, *British Medical Journal*, *309*(6947), 102.

Altman, D. G., and J. M. Bland (1994b), Diagnostic tests 1: sensitivity and specificity, *British Medical Journal*, *308*(6943), 1552.

American Academy of Sleep Medicine (2007), *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, American Academy of Sleep Medicine.

Astrom, K., and B. Wittenmark (1997), *Computer Controlled Systems: Theory and design*, Upper Saddle River, NJ: Prentice Hall.

Atiya, A. F., S. M. El-Shoura, S. I. Shaheen, and M. S. El-Sherif (1999), A comparison between neural-network forecasting techniques-case study: river flow forecasting, *Neural Networks, IEEE Transactions on*, *10*(2), 402–409.

Bao, Y., T. Xiong, and Z. Hu (2014a), Multi-step-ahead time series prediction using multiple-output support vector regression, *Neurocomputing*, *129*, 482–493.

Bao, Y., T. Xiong, and Z. Hu (2014b), PSO-MISMO modeling strategy for multistep-ahead time series prediction, *Cybernetics, IEEE Transactions on*, *44*(5), 655–668.

Baraldi, A. N., and C. K. Enders (2010), An introduction to modern missing data analyses, *Journal of School Psychology*, *48*(1), 5–37.

Ben Taieb, S., G. Bontempi, A. Sorjamaa, and A. Lendasse (2009), Long-term prediction of time series by combining direct and mimo strategies, in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pp. 3054–3061, IEEE.

Ben Taieb, S., A. Sorjamaa, and G. Bontempi (2010), Multiple-output modeling for multi-step-ahead time series forecasting, *Neurocomputing*, *73*(10), 1950–1957.

Ben Taieb, S., G. Bontempi, A. F. Atiya, and A. Sorjamaa (2012), A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition, *Expert systems with applications*, *39*(8), 7067–7083.

Berry, R. B., et al. (2012), Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events, *Journal of Clincial Sleep Medicine*, *8*(5), 597–619.

Bishop, C. M., et al. (2006), *Pattern recognition and machine learning*, vol. 1, springer New York.

Bliss, J. P., and M. C. Dunn (2000), Behavioural implications of alarm mistrust as a function of task workload, *Ergonomics*, *43*(9), 1283–1300.

Bock, J., D. Gough, et al. (1998), Toward prediction of physiological state signals in sleep apnea, *Biomedical Engineering, IEEE Transactions on*, *45*(11), 1332–1341.

Bontempi, G. (2008), Long term time series prediction with multi-input multi-output local learning, *Proc. 2nd ESTSP*, pp. 145–154.

Casdagli, M., S. Eubank, J. D. Farmer, and J. Gibson (1991), State space reconstruction in the presence of noise, *Physica D: Nonlinear Phenomena*, *51*(1), 52–98.

Centers for Medicare and Medicaid Services (1993), National coverage determination (NCD) for home use of oxygen.

Chen, L., X. Zhang, and H. Wang (2015), An obstructive sleep apnea detection approach using kernel density classification based on single-lead electrocardiogram, *Journal of medical systems*, *39*(5), 1–11.

Cheng, H., P.-N. Tan, J. Gao, and J. Scripps (2006), Multi-step-ahead time series prediction, in *Advances in Knowledge Discovery and Data Mining*, pp. 765–774, Springer.

Chevillon, G. (2007), Direct multi-step estimation and forecasting, *Journal of Economic Surveys*, *21*(4), 746–785.

Clarke, W. L. (2005), The original Clarke error grid analysis (EGA), *Diabetes Technology and Therapeutics*, *7*(5), 776–779.

Dagum, P., and A. Galper (1995), Time series prediction using belief network models, *International Journal of Human-Computer Studies*, *42*(6), 617–632.

De Chazal, P., C. Heneghan, E. Sheridan, R. Reilly, P. Nolan, and M. O'Malley (2003), Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea, *Biomedical Engineering, IEEE Transactions on*, *50*(6), 686–696.

Dua, P., F. J. Doyle, and E. N. Pistikopoulos (2006), Model-based blood glucose control for type 1 diabetes via parametric programming, *IEEE Transactions on Biomedical Engineering*, *53*(8), 1478–1491.

Eichhorn, J. (2003), Recognizing and preventing hypoxemic injury risk on the general care floor, *Journal of Healthcare and Risk Management*, *23*(1), 217 – 222.

ElMoaqet, H., D. M. Tilbury, and S.-K. Ramachandran (), Multi-step ahead predictions for critical levels in physiological time series, *IEEE Transcations on Cybernetics (Submitted)*.

ElMoaqet, H., D. Tilbury, and S. K. Ramachandran (2013a), Linear and non-linear models for blood oxygenation prediction: A comparison, *Technical Report of the University of Michigan-Mechanical Engineering Department*.

ElMoaqet, H., D. M. Tilbury, and S.-K. Ramachandran (2013b), Predicting oxygen saturation levels in blood using autoregressive models: A threshold metric for evaluating predictive models, *Proceedings of American Control Conference*, pp. 734–739.

ElMoaqet, H., D. Tilbury, and S. K. Ramachandran (2014a), Evaluating predictions of critical oxygen desaturation events, *Physiological Measurement*, *35*(4), 639–655.

ElMoaqet, H., D. M. Tilbury, and S.-K. Ramachandran (2014b), A novel dynamic model to predict abnormal oxygen desaturations in blood, in *Medical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on*, pp. 1–6, IEEE.

Epstein, L., and G. Dorlac (1998), Cost-effectiveness analysis of nocturnal oximetry as a method of screening for sleep apneahypopnea syndrome, *Chest*, *113*(1), 97–103.

Gani, A., A. Gribok, S. Rajaraman, W. Ward, and J. Reifman (2009), Predicting subcutaneous glucose concentration in humans:data-driven glucose modeling, *IEEE Transactions on Biomedical Engineering*, *56*(2), 246–254.

Gani, A., A. Gribok, Y. Lu, W. Ward, R. A. Vigersky, and J. Reifman (2010), Universal glucose models for predicting subcutaneous glucose concentration in humans, *IEEE Transactions on Information Technology in Biomedicine*, *14*(1), 157–165.

García, V., R. A. Mollineda, and J. S. Sánchez (2009), Index of balanced accuracy: A performance measure for skewed class distributions, in *Pattern Recognition and Image Analysis*, pp. 441–448, Springer.

Gather, U., M. Imhoff, and R. Fried (2002), Graphical models for multivariate time series from intensive care monitoring, *Statistics in medicine*, *21*(18), 2685–2701.

Geisser, S. (1993), *Predictive inference*, vol. 55, CRC Press.

Griffin, M., D. Lake, and J. Moorman (2005), Heart rate characteristics and laboratory tests in neonatal sepsis, *Pediatrics*, *115*(4), 937–941.

Hamzaçebi, C., D. Akay, and F. Kutay (2009), Comparison of direct and iterative artificial neural network forecast approaches in multi-periodic time series forecasting, *Expert Systems with Applications*, *36*(2), 3839–3844.

Hastie, T., R. Tibshirani, J. Friedman, and J. Franklin (2005), The elements of statistical learning: data mining, inference and prediction, *The Mathematical Intelligencer*, *27*(2), 83–85.

Hwang, D., N. Shakir, B. Limann, C. Sison, S. Kalra, L. Shulman, C. Souza-Ade, and H. Greenberg (2008), Association of sleep-disordered breathing with postoperative complications, *Chest*, *133*(5), 1128–1134.

Iber, C. (2007), *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, American Academy of Sleep Medicine.

Imhoff, M., F. Ronald, and U. Gather (2002), Detecting relationships between physiological variables using graphical models, *Proceedings of the American Medical Informatics Association (AMIA)*, pp. 340–344.

Ing, C.-K. (2003), Multistep prediction in autoregressive processes, *Econometric theory*, *19*(02), 254–279.

Joint Commission (2013), Medical device alarm safety in hospitals, *Sentinel Event Alert*, *50*.

Kline, D. M., and G. Zhang (2004), Methods for multi-step time series forecasting with neural networks, *Neural networks in business forecasting*, pp. 226–250.

Konno, K., and J. Mead (1967), Measurement of the separate volume changes of rib cage and abdomen during breathing, *Journal of Applied Physiology*, *22*(3), 407–422.

Kripke, D. F., S. Ancoli-Israel, M. R. Klauber, D. L. Wingard, W. J. Mason, and D. J. Mullaney (1997), Prevalence of sleep-disordered breathing in ages 40-64 years: a population based survey, *Sleep*, *20*(1), 65–76.

Lapedes, A., and R. Farber (1987), Nonlinear signal processing using neural networks: Prediction and system modelling, *Tech. rep.*

Le, T. Q., C. Cheng, A. Sangasoongsong, W. Wongdhamma, and S. T. Bukkapatnam (2013), Wireless wearable multisensory suite and real-time prediction of obstructive sleep apnea episodes, *Translational Engineering in Health and Medicine, IEEE Journal of*, *1*, 2700,109–2700,109.

Lee, J., J. P. Florian, and K. H. Chon (2011), Respiratory rate extraction from pulse oximeter and electrocardiographic recordings, *Physiological Measurement*, *32*(1), 1763–1773.

Lee, Y., M. Bister, P. Blanchfield, and Y. Salleh (2004), Automated detection of obstructive apnea and hypopnea events from oxygen saturation signal, *Proceedings of the* 26^{th} *Annual International Conference of the IEEE EMBS*, pp. 321–324.

Ljung, L. (1987), *System identification: theory for the user*, Upper Saddle River, NJ:Prentice-Hall.

Lynn, L., and P. Curry (2011), Patterns of unexpected in-hospital deaths: a root cause analysis, *Patient Safety in Surgery*, *5*, 1–24.

Mason, R., V. Broaddus, T. Martin, T. King, E. Schraufnagel, J. Murray, and J. Nadel (2010), *Murray and Nadel's Textbook of Respiratory Medicine*, Philadelphia, PA: Saunders Elsevier.

McFadden, C. J., L. T. Gutierrez, J. A. Leveque, and M. D. Anderson (1996), CPOM: alleviating the demand for ICU beds, *Journal of Nursing Management*, *27*(2), 48E – 48H.

Mendez, M., J. Corthout, S. Van Huffel, M. Matteucci, T. Penzel, S. Cerutti, and A. Bianchi (2010), Automatic screening of obstructive sleep apnea from the ECG based on empirical mode decomposition and wavelet analysis, *Physiological measurement*, *31*(3), 273.

Mendez, M. O., A. M. Bianchi, M. Matteucci, S. Cerutti, and T. Penzel (2009), Sleep apnea screening by autoregressive models from a single ECG lead, *Biomedical Engineering, IEEE Transactions on*, *56*(12), 2838–2850.

Micchelli, C. A., and M. Pontil (2005), On learning vector-valued functions, *Neural computation*, *17*(1), 177–204.

Mitchell, T. M. (1997), Machine learning. 1997, *Burr Ridge, IL: McGraw Hill*, *45*.

Moody, G. B., R. G. Mark, A. Zoccola, and S. Mantero (1985), Derivation of respiratory signals from multi-lead ECGs, *Computers in cardiology*, *12*(1985), 113–116.

Mooe, T., T. Rabben, U. Wiklund, K. Franklin, and P. Eriksson (1996), Sleep-disordered breathing in women: occurrence and association with coronary artery disease, *American Journal of Medicine*, *101*(3), 251–256.

Niskanen, J.-P., M. P. Tarvainen, P. O. Ranta-Aho, and P. A. Karjalainen (2004), Software for advanced hrv analysis, *Computer methods and programs in biomedicine*, *76*(1), 73–81.

Norris, P., P. Stein, and J. Morris (2008), Reduced heart rate multiscale entropy predicts death in critical illness: a study of physiologic complexity in 285 trauma patients, *Journal of Critical Care*, *23*(3), 399–405.

Oliver, N., and F. Flores-Mangas (2006), Healthgear: A real-time wearable system for monitoring and analyzing physiological signals, *Proceedings of the IEEE-International Workshop on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 1–4.

Palit, A. K., and D. Popovic (2006), Computational intelligence in time series forecasting.

Penzel, T., J. McNames, P. De Chazal, B. Raymond, A. Murray, and G. Moody (2002), Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings, *Medical and Biological Engineering and Computing*, *40*(4), 402–407.

Pintelon, R., and J. Schoukens (2012), *System identification: a frequency domain approach*, John Wiley & Sons.

Polit, D. F., and C. T. Beck (2008), *Nursing research: Generating and assessing evidence for nursing practice*, Lippincott Williams & Wilkins.

Pool, J. L., R. Glazer, N. Crikelair, and D. Levy (2009), The role of baseline blood pressure in guiding treatment choice, *Clinical drug investigation*, *29*(12), 791–802.

Powers, D. M. (2011), Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Quan, S., J. C. Gillin, M. Littner, and J. Shepard (1999), Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research. editorials, *Sleep*, *22*(5), 662–689.

Rauscher, J., W. Popp, , and H. Zwick (1991), Computerized detection of respiratory events during sleep from rapid increases in oxyhemoglobin saturation, *Lung*, *169*(1), 335–342.

Reifman, J., S. Rajaraman, A. Gribok, and W. Ward (2007), Predictive monitoring for improved management of glucose levels, *Diabetes Science and Technology*, *1*(4), 478–486.

Riordan, W., P. Norris, J. Jenkins, and J. Morris (2009), Early loss of heart rate complexity predicts mortality regardless of mechanism, anatomic location, or severity of injury in 2178 trauma patients, *Journal of Surgical Research*, *156*(2), 283–289.

Ruehland, W. R., P. D. Rochford, F. J. O'Donoghue, R. J. Pierce, P. Singh, and A. T. Thornton (2009), The new AASM criteria for scoring hypopneas: impact on the apnea hypopnea index, *Sleep*, *32*(2), 150.

Schulte-Frohlinde, V., Y. Ashkenazy, A. Goldberger, P. Ivanov, M. Costa, A. Morley-Davies, H. Stanley, , and L. Glass (2002), Complex patterns of abnormal heartbeats, *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, *66*(3), 031,901.

Shamir, M., L. Eidelman, Y. Floman, L. Kaplan, and R. Pizov (1999), Pulse oximetry plethysmographic waveform during changes in blood volume, *British journal of anaesthesia*, *82*(2), 178–181.

Shelley, K., D. Jablonka, A. Awad, R. Stout, H. Rezkanna, and D. Silverman (2006), What is the best site for measuring the effect of ventilation on the pulse oximeter waveform?, *Anesthesia & Analgesia*, *103*(2), 372–377.

Somers, V., M. E. Dyken, M. P. Clary, and F. M. Abboud (1995), Sympathetic neural mechanisms in obstructive sleep apnea, *Journal of Clinical Investigation*, *96*(4), 1897–1904.

Sorjamaa, A., and A. Lendasse (2006), Time series prediction using DirRec strategy., in *ESANN*, vol. 6, pp. 143–148.

Sorjamaa, A., J. Hao, N. Reyhani, Y. Ji, and A. Lendasse (2007), Methodology for long-term prediction of time series, *Neurocomputing*, *70*(16), 2861–2869.

Sörnmo, L., and P. Laguna (2005), *Bioelectrical signal processing in cardiac and neurological applications*, Academic Press.

Sparacino, G., F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli (2007), Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series, *IEEE Transactions on Biomedical Engineering*, *54*(5), 931–937.

Sparacino, G., A. Facchinetti, and C. Cobelli (2010), Smart continuous glucose monitoring sensors: On-line signal processing issues, *Sensors*, *10*(7), 6751–6772.

Stradling, J., and J. Crosby (1991), Predictors and prevalence of obstructive sleep apnea and snoring in 1001 middle aged men, *Thorax*, *46*(1), 85–90.

Swigris, J., X. Zhou, F. Wamboldt, R. du Bois, A. F. R. Keith, G. Cosgrove, S. K.Frankel, D. Curran-Everett, and K. Brown (2009), Exercise peripheral oxygen saturation (SpO_2) accurately reflects arterial oxygen saturation (SaO_2) and predicts mortality in systemic sclerosis, *Thorax*, *64*(7), 626–630.

Taenzer, A. H., J. Pyke, S. McGrath, , and G. Blike (2010), Impact of pulse oximetry surveillance on rescue events and intensive care unit transfers: a before-and-after concurrence study, *Anesthesiology*, *112*(1), 282–287.

Tang, Y., Y.-Q. Zhang, N. V. Chawla, and S. Krasser (2009), Svms modeling for highly imbalanced classification, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *39*(1), 281–288.

Tiao, G. C., and R. S. Tsay (1994), Some advances in non-linear and adaptive modelling in time-series, *Journal of forecasting*, *13*(2), 109–131.

Trajanoski, Z., W. Regittnig, and P. Wach (1998), Simulation studies on neural predictive control of glucose using the subcutaneous route, *Computer Methods and Programs in Biomedicine*, *56*(2), 133–139.

Travaglini, A., C. Lamberti, J. DeBie, and M. Ferri (1998), Respiratory signal derived from eight-lead ECG, in *Computers in Cardiology 1998*, pp. 65–68, IEEE.

Tsuji, H., F. Venditti, E. Manders, J. Evans, M. Larson, C. Feldman, and D. Levy (1994), Reduced heart rate variability and mortality risk in an elderly cohort, *The Framingham Heart Study. Circulation*, *90*(2), 878–883.

Voepel-Lewis, T., M. L. Parker, C. N. Burke, J. Hemberg, L. Perlin, S. Kai, and S. K. Ramachandran (2013), Pulse oximetry desaturation alarms on a general postoperative adult unit: A prospective observational study of nurse response time, *International Journal of Nursing Studies*, *50*(10), 1351–1358.

Vries, E. D., M. Ramratthan, S. Smorenburg, D. Gouma, and M. Boremeester (2008), The incidence and nature of in-hospital adverse events: a systematic review, *Quality and Safety in Health Care*, *17*(3), 216–223.

Watkinson, P., and L. Tarassenko (2012), Current and emerging approaches to address failure-to-rescue, *Anesthesiology*, *116*(5), 1158–1159.

Weigend, A. S. (1994), Time series prediction: forecasting the future and understanding the past, *Santa Fe Institute Studies in the Sciences of Complexity*.

Werbos, P. J. (1988), Generalization of backpropagation with application to a recurrent gas market model, *Neural Networks*, *1*(4), 339–356.

West, J. (2012), *Respiratory Physiology: The Essentials*, Baltimore, MD :Lippincott Williams & Wilkins.

Whitney, C. W., D. J. Gottlieb, S. Redline, R. G. Norman, R. R. Dodge, E. Shahar, S. Surovec, and F. J. Nieto (1998), Reliability of scoring respiratory disturbance indices and sleep staging., *Sleep*, *21*(7), 749–757.

Williams, K., and F. Galerneau (2003), Intrapartum fetal heart rate patterns in the prediction of neonatal acidemia, *American Journal of Obstetrics and Gynecology*, *188*(3), 820–823.

Xiong, T., Y. Bao, and Z. Hu (2013), Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices, *Energy Economics*, *40*, 405–415.

Young, T., L. Evans, L. Finn, and M. Palta (1997), Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women., *Sleep*, *20*(9), 705–706.

Yu, L., S. Wang, and K. K. Lai (2008), Forecasting crude oil price with an emd-based neural network ensemble learning paradigm, *Energy Economics*, *30*(5), 2623–2635.

Zhang, G., B. E. Patuwo, and M. Y. Hu (1998), Forecasting with artificial neural networks:: The state of the art, *International journal of forecasting*, *14*(1), 35–62.

Zweig, M. H., and G. Campbell (1993), Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, *39*(4), 561–577.