Sepsis Monitoring Using Contextually-Tailored Online Change Point Detection and Beyond

Nazmus Sakib
Marquette University

Follow this and additional works at: https://epublications.marquette.edu/dissertations_mu

Part of the Mathematics Commons

Recommended Citation
https://epublications.marquette.edu/dissertations_mu/1084
Sepsis Monitoring using Contextually-tailored Online Change Point Detection and Beyond

by

Nazmus Sakib

A Dissertation Submitted to the Faculty of the Graduate School, Marquette University, in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Milwaukee, Wisconsin
August, 2021
ABSTRACT
Sepsis Monitoring using Contextually-tailored Online Change Point Detection and Beyond

Nazmus Sakib, B.S., M.S.
Marquette University, 2021

Considering morbidity, mortality, and annual treatment costs, the dramatic rise in the incidence of sepsis and septic shock among intensive care unit (ICU) admissions in US hospitals is an increasing concern. The recent excruciating statistics regarding sepsis mortality, the average length of hospital stay, and annual treatment costs made sepsis treatment and research a critical domain in medical informatics. The aims of this dissertation center around four research questions. First, we discuss how we can investigate the prevalence and underlying relation of the sepsis diagnosis criteria (qSOFA and SIRS) and its implications in Medical Informatics and predictive analytics. Second, we delved into how we can develop a more sustainable medical informatics assistive solution that will help make evidence-based judgments instead of flummoxing the caregivers in decision making. Third, we aim to develop a data-driven tool as a medical informatics solution that helps ICU practitioners and researchers to monitor and intervene on the existing sepsis patients more efficiently and interactively and conduct retrospective studies to seek rationales to different sepsis scenarios in ICU. Fourth, we unravel the computational sustainability perspective of our medical informatics research. Computational Sustainability is a movement facilitated by CompSustNet—a virtual network led by Cornell University and supported by NSF—so that a novel scientific method, algorithm, or solution innovated to solve one particular problem of one domain can be repurposed for another distinct problem of another domain with a similar computational nature. Our discussion is twofold. Initially, we contextualized different perspectives and conceptual elements of Internet-of-Energy and Computational Sustainability implications. Then, we unravel how the Contextually-tailored Bayesian Online Change Point Detection Algorithm can be repurposed to address the Public Safety Power Shutoff issues impacting grid resiliency of IoE and Wildfire threat in the West of the United States.
# TABLE OF CONTENTS

Chapter 1 Introduction

1.1 Research Questions .................................................. 7
1.2 Dissertation Organization .......................................... 8
1.3 Related Publications .................................................. 12

Chapter 2 Unpacking Prevalence and Underlying Relation in qSOFA and SIRS

2.1 Theoretical Background ............................................. 15

2.1.1 Sepsis Pathophysiology ........................................... 15
2.1.2 Bedside Monitoring: qSOFA vs. SIRS ......................... 17
2.2 Data and Research Design ........................................... 20
2.3 Result ........................................................................ 25

2.3.1 Analyzing Statistical Distributions: qSOFA and SIRS ...... 25
2.3.2 Analyzing Patients’ Entire Trajectory (qSOFA) .............. 29
2.3.3 Analyzing Patients’ Entire Trajectory (SIRS) ................. 32
2.3.4 Analyzing Patients’ First Observation Only (qSOFA) ....... 34
2.3.5 Analyzing Patients’ First Observation Only (SIRS) ......... 36
2.4 Discussion ................................................................. 38

2.4.1 Theoretical Reasoning ............................................. 38
2.4.2 Medical Informatics Implications ................................. 39
2.4.3 Research Opportunities ............................................ 40
2.5 Summary ................................................................. 41

2.5.1 Objective .......................................................... 41
2.5.2 Methods .......................................................... 41
2.5.3 Results .......................................................... 42
2.5.4 Conclusions .................................................. 42

Chapter 3 State-of-the-art Medical Informatics Solutions Related to Sepsis: Existing Pitfalls and Potential Challenges 43

3.1 State-of-the-art .................................................. 43

3.1.1 Automated Sepsis Risk Assessment System: ................. 43

3.1.2 Predicting Sepsis, Septic Shock, and Sepsis-associated Poor Outcomes: .............................................. 48

3.1.3 Biomarker Discovery: ........................................... 53

3.1.4 Optimizing Therapeutics: ....................................... 57

3.2 Existing Pitfalls and Potential Challenges ......................... 62

3.2.1 Incomplete Observation ........................................ 64

3.2.2 Selection Bias .................................................... 64

3.2.3 CMIs .............................................................. 65

Chapter 4 Change Point Detection: Comparative Analysis on Methods and Utility 68

4.1 Defining and Formulating the Problem ............................... 69

4.2 Evaluation Metrics ................................................ 73

4.2.1 Evaluating the Change Point Detection Problem in the lens of Clustering: ............................................. 74

4.2.2 Evaluating the Change Point Detection Problem in the lens of Classification: .......................................... 75

4.3 Overviewing the Experimental Design ............................... 76

4.4 Results .............................................................. 78

Chapter 5 Contextually-tailored Online Change Point Detection for Sepsis Vital Sign Monitoring 82

5.1 Algorithm Inclusion Criteria ....................................... 82
5.1.1 Algorithmic Perspective ........................................... 82
5.1.2 Medical Informatics Perspective ................................ 83
5.2 Bayesian Online Change Point Detection ......................... 84
  5.2.1 Model ............................................................... 84
  5.2.2 Recursive run length posterior estimation ................... 85
  5.2.3 Underlying Probabilistic Model: ............................... 88
  5.2.4 Posterior Predictive for Conjugate Models: .................. 89
  5.2.5 Posterior Predictive for Exponential Family Models: ........ 90
  5.2.6 Message-passing Parameters: ................................ 91
  5.2.7 Change Point Prior: .......................................... 93
  5.2.8 Initial Conditions ............................................. 94
5.3 Sequential Inference Algorithm .................................... 95
5.4 Potential Challenges ................................................ 95
5.5 Contextually-tailored Bayesian Online Change Point Detection . 96
  5.5.1 Illustration ...................................................... 97

**Chapter 6** SepINav (Sepsis ICU Navigator): A Data-driven Software Tool for Sepsis Monitoring and Intervention using Contextually-tailored Bayesian Online Change Point Detection

6.1 Motivation and Significance ...................................... 101
6.2 Software Description .............................................. 103
6.3 Illustrative Examples .............................................. 106
6.4 Usability, Impact, and sustainability ............................ 111
6.5 Summary ............................................................ 119

**Chapter 7** Computational Sustainability and Internet-of-Energy

7.1 Introduction ....................................................... 121
7.2 Grid Overview of the United States .................................. 125
7.3 Load forecasting .......................................................... 134
  7.3.1 Weather and Climate of the United States ...................... 134
  7.3.2 Load Forecasting Techniques .................................... 146
7.4 Penetration of Renewable Energy .................................... 154
  7.4.1 Variability of Renewables ...................................... 154
  7.4.2 Effect of Commodity Prices .................................... 162
7.5 Understanding Energy Consumption Dynamics .................... 165
  7.5.1 Energy Disaggregation ......................................... 165
  7.5.2 Generation Expansion and Decision Analysis .................. 173
7.6 Internet-of-Energy ...................................................... 177
  7.6.1 IoE Architecture and IoT Integration ......................... 180
  7.6.2 Broader Impact .................................................. 183
  7.6.3 Challenges and Future Research Opportunities ............... 185
  7.6.4 IoE and Computational Sustainability ........................ 187

Chapter 8 WildFire and Public Safety Power Shutoffs: Implementing Contextually-tailored Online Change Point Detection 190

Chapter 9 Conclusion ...................................................... 199
  9.1 Contribution of this Dissertation .................................. 199
  9.2 Impact ................................................................. 201
    9.2.1 Immediate Impact ............................................. 201
    9.2.2 Long-term Impact ............................................. 202
  9.3 Future Works ......................................................... 203
    9.3.1 Data-driven Sepsis Monitoring and Intervention using Contextually-tailored Online Change Point Detection ...................... 203
9.3.2 WildFire and Public Safety Power Shutoffs: Implementing

Contextually-tailored Online Change Point Detection . . . 207

BIBLIOGRAPHY . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 208
List of Tables

2.1 Systemic inflammatory response syndrome (SIRS) criteria for sepsis definition. ........................................ 18
2.2 Quick Sequential Organ Failure Assessment. ..................... 19
2.3 Comparing Sepsis-2 and Sepsis-3 (SIRS vs. qSOFA) ............. 20
2.4 Statistical Distributions for qSOFA and SIRS. .................... 27
4.1 State-of-the-arts change point detection algorithms for this comparative study. ........................................ 77
4.2 Default and maximum score on covering metric for univariate series. 79
4.3 Default and maximum score on F1 metric for univariate series. .. 79
4.4 Default and maximum score on covering metric for multivariate series. ........................................ 79
4.5 Default and maximum score on F1 metric for Multivariate series. . 79
7.1 Statistically Significant Blackouts’ Cause Categories in the United States. ........................................ 144
7.2 Impact of Weather Elements in Electricity Demand and Generation. 144
7.3 Applications of Different Load Forecasting in Energy Workflow. . 147
7.4 Materials for clean energy technologies and components. .......... 164
7.5 Comparison between Hardware-based and Software-based Disaggregation Techniques. .......................... 170
7.6 Transferable Computational Techniques and Their Prospective Applications in IoE Conceptualization (1) .......... 188
7.7 Transferable Computational Techniques and Their Prospective Applications in IoE Conceptualization (2) .......... 189
List of Figures

1.1 death related to sepsis (%) over age groups due to infections, injuries, and non-communicable diseases. ......................... 2

1.2 Hospitalization cost in the United States in 2013. .................... 3

2.1 A simple outline for research design. ................................. 24

2.2 Kernel density estimation distribution of (a) qSOFA and (b) SIRS parameters to understand the prevalence. ................... 28

2.3 Kernel density estimation distribution of qSOFA parameters for both safe and qSOFA criterion met observations to understand the prevalent qSOFA parameters. ................................. 30

2.4 Facet grid illustration of Sepsis-3 (qSOFA) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario in the ICU. ................................. 31

2.5 Kernel density estimation distribution of SIRS parameters for both safe and sepsis criterion met observations to understand the prevalent SIRS parameters. ................................. 32

2.6 Facet grid illustration of Sepsis-2 (SIRS) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario in the ICU. ................................. 33
2.7 Kernel density estimation distribution of qSOFA parameters for both safe and qSOFA criterion met patients at first observations to understand the prevalent qSOFA parameters. ........................... 35

2.8 Facet grid illustration of Sepsis-3 (qSOFA) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario of patients at first observations in the ICU. .... 35

2.9 Kernel density estimation distribution of SIRS parameters for both safe and sepsis criterion met patients at first observations to understand the prevalent SIRS parameters. ......................... 37

2.10 Facet grid illustration of Sepsis-2 (SIRS) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario of patients at first observations in the ICU. ....... 37

3.1 Visualizing physiological alterations and their effects on the antimicrobial pharmacokinetics [1]. .................................................. 59

3.2 Visualizing physiological functional alterations and their effects on the antimicrobial pharmacokinetics [1]. ........................................ 59

3.3 State-of-the-art Medical Informatics Solutions for Sepsis Care. ..... 61

4.1 Illustration of change points (break points) and states (regimes) in a sample time series problem [2]. ............................................. 69
5.1 Conceptual sketch representing change point detection and run-length of Bayesian Online Change Point Detection. The top figure shows synthetic data with two partitions (the third partition is undefined/unknown). The bottom figure represents the associated run length \((r_t)\) as a time function [3].

5.2 Illustrating the message-passing algorithm where each of the nodes has mass associated with it [3].

5.3 Illustrating parameter updates that is living on a trellis of possible values.

5.4 Sequential inference of Bayesian online change point detection [3, 4].

5.5 Simulated data drawn using Gaussian distribution with variable mean and constant variance; ground-truth change points with bold red dash lines.

5.6 Implementing proposed algorithm realizing scenario presented in 5.5.

6.1 Sepsis-3 clinical criteria to diagnose patients with sepsis and septic shock.

6.2 Contextually-tailored bayesian online change point detection algorithm with prediction.

6.3 Illustration of monitoring practitioner’s interventions and patient’s trajectory in Sepsis-2.
6.4 Illustration of monitoring practitioner’s interventions and patient’s trajectory in Sepsis-3.

6.5 Structural change detected in the patient’s vital signs using bayesian online change point detection algorithm.

6.6 Illustrating the option to observe correlations between vital signs crucial for Sepsis.

6.7 Understanding Sepsis-3 in the lens of change points in qSOFA parameters’ data stream.

6.8 Understanding Sepsis-2 in the lens of change points in SIRS parameters’ data stream.

6.9 A Step Towards Evidence-based and Data-driven Decision Making in Sepsis Monitoring and Intervention.

7.1 Functional layers of IoT application, accumulating physical, data and operational layers with communication and power lines.

7.2 United States synchronous grid of 186, 411 miles.

7.3 Geographical distribution of Federal MG assessment and demonstration projects in the United States.

7.4 Distribution of generation from different sources in the United States in 2017.

7.5 Annual share trend in United States for electricity generation by source from 1950 to date.
7.6 Evolution of the generation mix contributing to the United States electricity generation over the time. ........................................ 131

7.7 Net generation distribution of electricity by type and states in the United States. ......................................................... 132

7.8 Extent of ISOs in the United States. ........................................ 134

7.9 Electricity consumption in US homes during 2018 (kWh/year). . 135

7.10 Region classification of the United States based on heating degree days. ................................................................. 137

7.11 Region classification of the United States based on cooling degree days. ................................................................. 137

7.12 Temperature anomaly from 1901 to 2015 in the contiguous 48 states.138

7.13 Anomaly between a normal atmospheric condition and forecasted catastrophic condition in the dynamics of hydropower-based system. 140

7.14 Patterns of seasonal wind variability in different regions of the United States. ......................................................... 141

7.15 Impact of Hurricanes on oil and natural gas production in the United States. ......................................................... 143

7.16 Variation in wind energy generation output for 200 and 15 wind turbines placed in dispersed positions. ............................ 145
7.17 Trade-off in flexibility of resources, though we need flexibility in resources to address variability in demand, flexible resources are costly. ................................................................. 148

7.18 Yearly US per capita consumption in kWh by renewable sources from 1999-2016 with trendlines and five year-forecast. .................. 155

7.19 The effect of sky conditions on solar panel power output. ............. 156

7.20 The wind velocity distribution over the United States. ................. 158

7.21 Year to year changes in wind power production. ....................... 159

7.22 Impact of increasing PV penetration and duck diagram interpretation of over-generation risk and ramp requirement. ..................... 160

7.23 Duck diagram interpretation of over-generation risk and ramp requirement. ................................................................. 160

7.24 Commodity prices drive wind energy prices. .............................. 163

7.25 Criticality matrix of materials used in clean energy for short-term and medium-term. ................................................................. 165

7.26 Electricity consumption by different sectors and household electricity consumption distribution in the United States. ................. 166

7.27 Iterative process for developing models to understand energy usage behavior. ................................................................. 167
7.28 Energy consumption patterns of different sectors for both monthly 
and sub-daily basis. .................................................. 168

7.29 Almost real-time household level total power provided by the ad-
vanced meter and the desired disaggregated power. .................... 169

7.30 Capacity expansion over time. .................................. 174

7.31 Impact of the critical shortage challenges in the case of the Pacific 
Northwest of the United States. .................................... 174

7.32 Impact of overestimating electricity demand resulted in a higher price.175

7.33 Essential layers of IoT deployment with the smart grid in the real-
ization of the internet-of-energy. .................................... 178

7.34 Internet-of-Energy integration. ................................. 180

7.35 benefits of adopting internet-of-energy from the functional aspect . 182

8.1 Annual acres burned in wildland fires from 1980-2020 in the US. .. 190

8.2 Overall wildfire losses in the US (2010-2019). .................... 191

8.3 Biggest wildfires (with respect to land acres) in the history of Cal-
ifornia and associated losses. ....................................... 191

8.4 Proposed dynamics for Wildfire-risk and PSPS event monitoring. .. 195

8.5 Realizing Change point detection in understanding wildfire risks in 
San Francisco, 2019. .................................................. 197
9.1 Time series depiction for the daily closing price and volume of Apple Inc. Stock presented in Yahoo Finance. It captures data for a time around the year 2000, sampled at every three-time steps in an effort to reduce the length of the series. From the contextual aspect, 09-29-2000 experienced a notable drop in stock price [5]... 242

9.2 Time series of the available amount in an anonymous person’s bank account. The data stream shows periodic attributes and significant variation with time [5]... 242

9.3 Time series representing movement switches of Honey bee between three states: waggle, right turn, and left turn [5]... 243

9.4 Bitcoin Price in US Dollar, data collected from bitcoin.com [5]... 243

9.5 Price of Brent Crude Oil per barrel in US Dollar. Data was collected from EIA (Energy Information Administration) of the United States and then sampled at every ten observations to reduce the series’s breadth [5]... 243

9.6 Time series representing monthly total business inventories in US dollars. From the contextual aspect, the effects of the financial crisis are apparent. Data were collected from the US Census Bureau [5]... 244

9.7 Time series depicts the population of Centralia, the mining town of Pennsylvania. The context is that a mine fire burning since 1962 has a significant impact, and in 1992, it experienced condemning all buildings in the town [5]... 244
9.8 Number of children born per women on average globally (Source: GapMinder) [5]. ................................................................. 244

9.9 Emissions of $CO_2$ per person in Canada. (Source: GapMinder) [5] . 245

9.10 Time series representing total construction spending in the private sector in the United States over time. (Source: United States Census Bureau) [5] ................................................................. 245


9.12 GDP over time in constant local currency in Argentina. (Source: World Bank) [5] ................................................................. 246

9.13 GDP over time in constant local currency in Croatia. (Source: World Bank) [5] ................................................................. 246

9.14 GDP over time in constant local currency in Iran. (Source: World Bank) [5] ................................................................. 246


9.16 Global $CO_2$ level per month over time. To reduce the length of the time series, collected data were sampled every four years. [5] . . . 247
9.17 Time series representing the number of home runs each year in the American League of baseball. (Source: Data retrieved from Baseball Databank). From the contextual aspect, Major League Baseball was expanded, indicating a potentially significant impact of World War-II [5].

9.18 The number of visitor entries per month through Keflavik airport, Iceland (Data Source: Icelandic Tourist Board) [5].

9.19 Monthly arrival and departure count of passengers at John F. Kennedy Airport, New York city (Data Source: Port Authority of New York and New Jersey) [5].

9.20 Monthly arrival and departure count of passengers at LaGuardia Airport, New York city (Data Source: Port Authority of New York and New Jersey) [5].

9.21 Monthly measles cases in England and Wales for a certain period of time [5].

9.22 Time series representing the annual volume of the Nile River at Aswan, Egypt. This data contextualizes the change point that happened as an effect of a dam built in 1898 [5].
9.23 Time series representing room occupancy associated parameters by Candanedo and Feldheim in 2016. To abridge the length of the series, collected data was sampled, while preparing the time-series, at every 16 observations. This multidimensional time-series captures four measurements: temperature, relative humidity, light, and CO$_2$, respectively (Data Source: UCI Repository) [5]. ........................................ 249

9.24 Ozone-depleting substances’ level in the atmosphere over time, contextualizing the impact of the Montreal Protocol enforced in September 1989 [5]. .................................................. 250

9.25 Time series representing total available rail-lines in the world in Kilometers (Data Source: World Bank) [5]. .................................................. 250

9.26 Ratner Group’s stock price time-series, sampled at every 3 original observations of the collected data, in the early 1990s (Data Source: Yahoo Finance) [5]. .................................................. 250

9.27 Time series representing the number of automated phone calls in the United States over time. In context, these numbers are being subjected to vary with the degrees of regulation imposed by the government [5]. .................................................. 251

9.28 Time series data stream representing pace and total distance covered by a runner while following an interval training program (pace indicates the runner’s alteration between running and walking in an interval) [5]. .................................................. 251
9.29 A horizontal scan line of an image (Data collected from the BSDS, image no. 126007) [5].

9.30 A horizontal scan line of an image (Data collected from the BSDS, image no. 42049) [5].

9.31 Time series representing the impact of introducing seat-belt mandate in the UK as the number of drivers killed or seriously injured experienced a gradual decline. In context, seat-belts were compulsory in the UK for the new cars starting in 1972 and were mandatory to be worn for the driver from 1983 onward [5].

9.32 Number of license plate applicants over time in Shanghai [5].

9.33 Number of workers employed in coal mines in a spectrum of hundred years in Britain [5].

9.34 Time series representing the unemployment rates in The Netherlands [5].

9.35 Fluctuation in the conversion rate between USD and Icelandic Krona, representing a sharp drop as the impact of the financial crisis (Data Source: EuroStat) [5].

9.36 The population of the United States in the spectrum of seventy years (Data Source: United States Census Bureau) [5].

9.37 Well-log data, after sampling the original observations in an interval of six to truncate the breadth of the series. [6].
9.38 A simulated series with a predefined change point. A data stream with Gaussian noise and a small trend, and a series with an offset and uniform noise are partitioned at index 146 [5].

9.39 A simulated series with a predefined change point, having a constant noise and a mean shift change point at index 97 [5].

9.40 A simulated series with a predefined change point. A series having the noise of $N(0, 1)$ and a series having the noise of $N(2, 2)$ are partitioned at index 179. This simulated series also has an outlier at index 42 [5].

9.41 A simulated series with multiple periodic components of different amplitude and a predefined change point at index 341. Contextually, the series reveals a change point since the mean shifts [5].

9.42 A simulated series without having a change point follows $N(0, 1)$ [5].
Chapter 1
Introduction

Sepsis, one of the most elusive syndromes in medical science, is a syndrome induced by infection and associated with biochemical, physiological, and pathological abnormalities as a result of an unregulated response from the human body [7, 8, 9]. In the United States, over 1.7 million adults are affected by sepsis each year, and annually, more than 970,000 cases are admitted to U.S. hospitals because of sepsis. Sepsis—directly and indirectly—contributes to annually more than 250,000 deaths, representing more than 50% of hospital deaths [8, 10, 11, 12, 13, 14]. Unfortunately, this excruciating statistics has been exacerbated over the years in recent times, identified by a 2-decade study on U.S. hospitalizations, costs, and disease epidemiology. Statistics reflect an 8.7% annual increase in the incidence of sepsis among hospitalized patients in the U.S. [11, 15, 16].

Besides the alarmingly increasing sepsis incidence and mortality rate, the average Length of Stay (LOS) in hospitals is considerably higher (approximately 75% higher than the most other conditions) for sepsis patients in the U.S.; and so is the burden associated with hospital utilization [16, 17, 18, 19]. Furthermore, [20] reported that the average LOS of septic patients is dilated compellingly in 2013, and there observed a distinct proportion with the severity of sepsis, such as 4.5 days, 6.5 days, and 16.5 days for sepsis, severe sepsis, and septic shock respectively according to the SIRS criteria. Apart from that, though accounted for 3.6% of hospital stays, sepsis portrayed 13% of total U.S. hospital costs, resulting in hospital expenses of more than $24 billion in 2013. Not surprisingly, in 2013, the cost associated with sepsis management ranked highest
among the admissions for all diseases and medical conditions, followed by osteoarthritis of $17$ billion and childbirth (medical condition) of $13$ billion [21, 22, 23]. As of now, the hospital costs associated with sepsis still hold this rank, and currently, require more than twice compared to those of other medical conditions [24]. [9, 25, 26] indicate these costs can even be exacerbated in the near future, and approach to thrice compared to other admissions.

Sepsis exhibits complicated dynamics, as compounded by host factors and pathogen factors [27]. Figure 1.1 projects the sepsis scenario in the global perspectives and illustrates the death related to sepsis (%) over the age groups (one of the host factors) due to infections, injuries, and non-communicable diseases (pathogen factors)[28]. [28] reports that neonates are the most vulnerable age group considering sepsis mortality (host perspective). It further depicts that infections are the fatal causes towards sepsis-related mortality (pathogen perspective). Like its complicated dynamics, sepsis hospitalization

![Figure 1.1: death related to sepsis (%) over age groups due to infections, injuries, and non-communicable diseases.](image-url)
costs consistently rank highest among admissions for all disease states and medical conditions. Figure 1.2 summarizes the hospitalization cost in the United States in 2013, denoting a snapshot of sepsis hospitalization burden [29]. These excruciating statistics reflect in dire need to further research automated sepsis risk assessment systems, predicting sepsis, septic shock, and sepsis-associated poor outcomes, biomarker discovery, and optimizing therapeutics.

This notable increase in mortality rate and annual health care expenditure (affected by LOS) made sepsis treatment and research a critical domain in medical internet research and medical informatics, resulting in a current surge in the literature [30, 31, 32, 33]. Study shows improved and effective methods of early sepsis identification can substantially reduce the severity and epidemiological burden of sepsis in the U.S. [33, 34, 35, 36, 37, 38]. Besides, a number of studies recommended that identifying the prevalent risk factor(s), followed by the instant diagnosis, can reduce the cost in treatment workflow, and further scale down the mortality rate to some extent [35, 39, 40, 41, 42]. However, the literature manifests that the studies concentrated on one risk factor

<table>
<thead>
<tr>
<th>Condition</th>
<th>Hospitalization Cost (National cost, %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Septicemia</td>
<td>23,663 (6.2%)</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>16,520 (4.3%)</td>
</tr>
<tr>
<td>Liveborn</td>
<td>13,287 (3.5%)</td>
</tr>
<tr>
<td>Complication of device</td>
<td>12,431 (3.3%)</td>
</tr>
<tr>
<td>Acute myocardial infarction</td>
<td>12,092 (3.2%)</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>10,218 (2.7%)</td>
</tr>
<tr>
<td>Spondylosis</td>
<td>10,198 (2.7%)</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>9,501 (2.5%)</td>
</tr>
<tr>
<td>Coronary atherosclerosis</td>
<td>9,003 (2.4%)</td>
</tr>
<tr>
<td>Acute cerebrovascular disease</td>
<td>8,840 (2.3%)</td>
</tr>
</tbody>
</table>

Figure 1.2: Hospitalization cost in the United States in 2013.
at a time for the clinical assessment of sepsis limited the probability for sepsis detection as it requires complex reasonings and implications. Moreover, in many cases, it is apparent that results are sensitive to subtle variations in definition(s), as well as subjective suspicions of physicians [30, 43, 44, 45, 46].

MIMIC-III (Version 1.4), the recent major release of Medical Information Mart for Intensive Care, is an extensive, single-center, and comprehensive database comprising information, such as vital signs, laboratory measurements, observations and notes charted by care providers, imaging reports, fluid balance, medications, procedure codes, diagnostic codes, hospital length of stay, pertaining to patients admitted to the critical care units at Beth Israel Deaconess Medical Center in Boston [23, 30, 47, 48]. It is a multidisciplinary collaborative effort of Laboratory for Computational Physiology at MIT, Computer Science and Artificial Intelligence Laboratory at MIT, and Information Systems at Beth Israel Deaconess Medical Center in Boston. The underlying motivation behind this collaboration is assuring reproducibility and improving quality of the data-driven medical informatics research. The salient features of MIMIC-III (Version 1.4) include it is the only freely accessible critical care database in the United States of its kind that promotes analysis without any additional restriction after accepting the data use agreement. Furthermore, a critical care dataset with detailed individual patient care information spanning more than a decade empowers medical informatics research and pedagogy around the world. MIMIC-III (Version 1.4) contains data from 58,976 hospital admissions for the patients admitted to the critical care units from 2001 to 2012. Personal information is removed, and the original records are shifted and reformatted to warrant the data is not identifiable for human subjects. Besides, it comprises 26 tables linked by identifiers for the corresponding patient. Each of the tables is a spreadsheet including information on patient stays and the
physiological data collected in the ICU, with data dictionaries to explain the observational context. MIMIC-III (Version 1.4) allows a variety of data forms, ranging from text interpretations for radiology images to the time-stamped physiological measures [30, 49]. This open and unrestricted nature of the extensive healthcare data allows the clinical studies to be improved and reproduced in such ways that would not otherwise be possible [47]. Hence, MIMIC-III (Version 1.4) can facilitate exploratory and data-driven studies on sepsis, its diagnosis, and treatment in the ICU [30, 23].

Sepsis is first formally defined by a 1991 consensus conference as the host’s systemic inflammatory response syndrome (abbreviated as SIRS) to infection [7, 50]. According to the then-prevailing definition, sepsis associated with organ dysfunction was named as severe sepsis, and severe sepsis followed by the sepsis-induced persisting hypotension despite the adequate fluid resuscitation was termed as septic shock. Then, considering the limitations with 1991 consensus conference definitions, the 2001 task force extended the list of diagnostic criteria for sepsis. Albeit discrepancy in the 1991’s interpretation, the 2001 task force could not offer an alternative definition due to lack of supporting evidence, resulting in sepsis definition remained mostly unchanged since 1991 to 2016. In 2016, a task force comprising the experts of sepsis pathobiology, pathophysiology, epidemiology, and clinical trials convened by the Society of Critical Care Medicine, along with the European Society of Intensive Care Medicine, revised the definition of sepsis and septic shock. The substantial advances observed in pathobiology, epidemiology, immunology, and intervention management motivate to reexamine how we interpret sepsis. The definition by the 2016’s task force has been supported by 31 international sites [7]. [7] concluded that it is necessary to change the perception about sepsis to have a more reliable predictive indicator of mortality and impact in the survivability of
the patients. Hence, SIRS is replaced by the quick Sequential Organ Failure Assessment (qSOFA). qSOFA suggests three criteria to evaluate patients likely to have poor outcome due to sepsis: Hypotension, Altered Mental Status, and High Respiratory Rate [30]. Besides qSOFA, sepsis-3 (as it implies the third updated definition of sepsis) includes SOFA (Sepsis-related Organ Failure Assessment) for sepsis diagnosis. Albeit not substantial, SOFA provides better predictive accuracy with greater consistency compared to qSOFA. However, the telling intricacy and time-taking lab tests involving in SOFA left it not-well-known outside the critical care community since the definition had been updated in 2016.

This doctoral dissertation proposal endeavors several data-driven medical informatics solutions that impact the current dynamics of sepsis monitoring and intervention in the ICU. Besides, this research has important implications for sepsis treatment initiatives in the ICU and informing hospital resource allocation, and has the potential to improve the preventability of deaths from sepsis. After that, this dissertation will unravel the computational sustainability perspective of our medical informatics research. Computational Sustainability is a movement facilitated by CompSustNet—a virtual network led by Cornell University and supported by NSF—so that a novel scientific method, algorithm, or solution innovated to solve one particular problem of one domain can be repurposed for another distinct problem of another domain with a similar computational nature. Initially, we contextualized different perspectives and conceptual elements of Internet-of-Energy and Computational Sustainability implications. Then, we unravel how the Contextually-tailored Bayesian Online Change Point Detection Algorithm can be repurposed to address the Public Safety Power Shutoff (PSPS) issues impacting grid resiliency of IoE and Wildfire
threat in the West of the United States.

1.1 Research Questions

This doctoral dissertation proposal is centered around the following research questions.

Research Question (RQ1): How can we investigate the prevalence and underlying relation of the sepsis diagnosis criteria (qSOFA and SIRS)?

1.1 What is the most prevalent qSOFA and SIRS criterion?

1.2 What is the most prevalent sepsis-3 and sepsis-2 scenario?

1.3 Does there exist any multicollinearity among the qSOFA parameters and SIRS parameters?

Research Question (RQ2): How can we develop a more sustainable medical informatics assistive solution that will help make evidence-based judgment instead of flummoxing the caregivers in decision making?

2.1. What are the available machine learning-based solutions and existing research gaps?

2.2 Why change point detection-based assistive and practitioner-centric solution has potential to be more sustainable in sepsis monitoring?

2.3. How can we design the change point detection algorithm addressing contextual challenges for sepsis monitoring?

Research Question (RQ3): How can we develop a data-driven tool as a medical informatics solution that helps ICU practitioners and
researchers to monitor and intervene on the existing sepsis patients more efficiently and interactively and conduct retrospective studies to seek rationales to different sepsis scenarios in the ICU?

3.1. How can this tool assist the practitioners’ in data-driven decision making?

3.2. How can we address the recent shift in sepsis definition in this system?

3.3. How can several additional features be added to this data-driven software tool for efficient monitoring and intervention?

Research Question (RQ4): How can the Contextually-tailored Bayesian Online Change Point Detection Algorithm be repurposed to address the PSPS issues impacting grid resiliency of IoE and Wildfire threat in West US?

1.2 Dissertation Organization

Considering the breadth of this dissertation report, we summarize the organization of each of the chapters to help the reader navigate concepts and research issues discussed here.

Chapter 2, addressing the RQ1, answers a research gap critical to both hospital resource allocation and developing a predictive model using electronic medical record (EMR) data: understanding the prevalence and dichotomy in the SIRS (Systemic Inflammatory Response Syndrome) and qSOFA (quick-Sequential Organ Failure Assessment) criteria. The objective of this study is threefold. First, we aim to unpack the most prevalent criterion for sepsis (for both sepsis-3 and sepsis-2 predictors). Second, we intend to determine the most prevalent sepsis scenario in the ICU (among four possible scenarios for qSOFA and eleven possible scenarios for SIRS). Third, we investigate the
multicollinearity or dichotomy among qSOFA predictors and SIRS predictors. Besides, we discuss the results through the lens of sepsis pathophysiology. After that, we point out several research opportunities for the current and prospective researchers of this domain.

Chapter 3, 4, and 5 contextualize and address the RQ2. In chapter 3, we review the state-of-the-art solutions in understanding sepsis dynamics and developing tools that can help caregivers monitor sepsis patients in the ICU. This review explains the literature from four lenses: Automated Sepsis Risk Assessment System; Biomarker Discovery; Predicting Sepsis, Septic Shock, and Sepsis-associated Poor Outcomes; and Optimizing Therapeutics. Here, we discuss the existing research gaps in the current solutions and how their efficacy is affected in practice. After that, we particularly discuss the pitfalls and challenges confronted when developing predictive models using EMRs. Besides explaining the possible pitfalls and challenges, this chapter concentrates on the confounding medical intervention and illustrates how it affects predictive models’ real-world practice. It also infers how machine learning-based assistive solutions can be less susceptible to these pitfalls and challenges. It shows how assistive solutions help to make evidence-based judgment instead of flummoxing the caregivers in decision making. It implies that detecting structural changes in vital signs can guide caregivers’ monitoring and interventions.

Chapter 4 introduces the concept of change point detection and formulate the problem in a formal fashion. To identify the change point detection algorithm that we may use for our data-driven sepsis navigation tool, chapter 4 presents a comparative study among the state-of-art unsupervised change point detection algorithms using a combination of real-world and simulated time-series data. This study evaluates the algorithms from two lenses: clustering and classification. We used covering and F1-metric as evaluation metrics for
clustering and classification, respectively. Chapter 4 shows two rationals that helped us to decide bayesian online change point detection for our base algorithm from this comparative study. First, it outperforms all the algorithms as far as the hyperparameter tuning is concerned. Second, bayesian online change point detection gives us the provision to take the domain knowledge into consideration.

In chapter 5, we present the mathematical detail of the bayesian online change point detection algorithm. Besides, we discuss the related concepts to understand the complicated mathematics behind the algorithm. Chapter 5 also points out how this algorithm can differentiate between an anomaly and actual structural change in the data stream. However, the current hospitalization systems can provide more frequent data regarding bedside monitoring of vital signs. Being integrated into EMR, the more frequently captured data may induce noise in detecting structural changes in the physiological parameters. We discuss how the Contextually-tailored Bayesian Online Change Point Detection algorithm can address the contextual gap in the existing bayesian online change point detection.

In chapter 6, addressing RQ3, we develop Sepsis ICU Navigator (SepINav): a medical informatics endeavor that helps ICU practitioners and researchers to monitor and intervene on the existing sepsis patients more efficiently and interactively and conduct retrospective studies to seek rationales to different sepsis scenarios in the ICU. Besides, the Contextually-tailored Bayesian Online Changepoint Detection algorithm will help the practitioners understand the structural changes in patients’ vital sign regimes that may harbinger prior to septic shock. Besides, several additional features are added to this data-driven software tool to promise efficient monitoring and intervention and address confounding medical interventions in the ICU.

Chapter 7 and 8 contextualize and address the RQ4. In chapter 7, we
contextualize different perspectives and conceptual elements of Internet-of-Energy and Computational Sustainability implications. The burgeoning growth of Big Data not only matures and improves the data management efficiency and useful information extraction techniques, but also motivates computational science researchers to come up with a new method or solution that can be repurposed for problems across the domain. Computational Sustainability joins this movement for a transferrable computational technique for sustainable development and a better future. Internet-of-energy (IoE)—leveraging IoT to smart grids associated with advanced analytics—is one of the prominent efforts in this regard. This qualitative study includes the Grid Overview of the United States; Weather and Climate and its impact on the entire energy generation and consumption dynamics; Peak Load Forecasting and its techniques and burgeoning challenges; Variable Renewable Energy, its reliability challenges and how we can take advantage of this variability; Commodity Prices and its criticality; Energy Disaggregation and its impact on consumer awareness; and Generation Expansion and Decision Analysis. Besides, IoE integration, associated tradeoffs, challenges, research opportunities, and transferable computational techniques are addressed in this chapter.

In chapter 8, we answer how we can repurpose the Contextually-tailored Bayesian Online Change Point Detection Algorithm to address the PSPS issues impacting grid resiliency of IoE and Wildfire threat in West US. The escalating frequency and scale of wildfires in the recent past are increasing concerns worldwide, considering the alarming influence of climate change and associated socio-economic impacts. In particular, the hazards of wildfire in the western United States, especially California, are substantially destructive and alarming in recent years. Public Safety Power Shutoff (PSPS), introduced and practiced by PGE in the west after the 2017-18 wildfire, is a precautionary safety measure
that proactively turns off power lines in communities in the higher fire threat area. PSPS is implemented after the 2017-18 wildfire. PSPS decision-making process is complex, and still, further research is required to shorten the length and area of the power shutoff, assuring safety measures. We explain how the algorithm, as mentioned above, can help in monitoring the situations and making decisions efficiently to save people and lands from possible hazards.

Chapter 9 concludes with discussing the contribution to the body of knowledge, immediate and long-term impact of the work, future research opportunities and our plans to advance this research in two domains, as described, besides briefly summarize the entire dissertation report.

1.3 Related Publications

- Several parts of chapter 1 and 2 are published as follows:


- Several parts of chapter 3, 4, 5, and 6 are published as follows:

monitoring and intervention using Bayesian Online Change Point Detection. SoftwareX, 14, 100689.

• Several parts of chapter 7 are published as follows:


• Several parts of chapter 6 and 8 are published as follows:

Chapter 2
Unpacking Prevalence and Underlying Relation in qSOFA and SIRS

Considering morbidity, mortality, and annual treatment costs, the dramatically increasing incidents of sepsis and septic shock among ICU admissions in the United States hospitals are an increasing concern. The recent changes in sepsis definition (sepsis-3) motivate the medical informatics research community around the world to investigate score recalculation and information retrieval and study the intersection between Sepsis-3 and Sepsis-2. Besides, considering the stark contrasts evinced in the results (reflect by the evaluation metrics such as accuracy, sensitivity, precision, and G-mean) of predictive modeling using SIRS and qSOFA parameters; in this chapter, we decided to take a step back and have a more in-depth look at the qSOFA and SIRS parameters, and their underlying attributes and interrelations. Multicollinearity among parameters often intensifies the tension between optimization and generalizability, and eventually, leads to model overfitting. Model Overfitting hampers the generalizability of discriminant functions [51, 52]. Moreover, it indicates that a small deviation in the input data can result in considerable, sometimes aberrant, changes in the model, and even leading to change in the sign of parameter estimates [53, 51, 52].

As sepsis is still perceived as a spectrum disease that subsequently ends in organ dysfunction, septic shock is a crucial juncture to the multi-parameter intelligent sepsis prediction in ICU. However, this study is undividedly focused on sepsis based on SIRS and qSOFA. Quantifying the prevalence of the qSOFA criteria (so as SIRS) and understanding the underlying relation of the
parameters have important implications for sepsis treatment initiatives in the ICU and informing hospital resource allocation. Hence, this work has the potential to improve the preventability of deaths from sepsis. The unique data driven contributions of this study are as follows:

- **Prevalent Sepsis Criterion**: Unpack the most prevalent SIRS and qSOFA criterion.

- **Prevalent Sepsis Scenario in ICU**: Determine the most prevalent sepsis scenario based on SIRS criteria and qSOFA criteria.

- **Underlying Statistical Relation among Predictors**: Investigate the multicollinearity among SIRS and qSOFA criteria with design implication for predictive modeling.

2.1 Theoretical Background

2.1.1 Sepsis Pathophysiology

Sepsis—commonly interpreted as a spectrum disease—ranges from milder symptoms and ends in septic shock, and subsequently, followed by multiple organ dysfunction syndromes. This entire spectrum begins with the introduction of pathogens, such as gram-positive or gram-negative bacteria, fungus, viruses, and parasites, in the blood vessels. The appearance of pathogens in blood vessels makes them no longer sterile; when the white blood cells confront these infective materials (pathogens), they become activated. Consequently, more white blood cells are called in that infected location to eradicate the pathogens. Generally, these infective materials exist out in the interstitial tissue, not in the bloodstream. Therefore, to get into the infective materials there and eradicate them, the white blood cells release substances, such as nitric oxide. When these substances interact with blood vessels, there happen three things. First, the
diameter of the blood vessel expands, resulting in vasodilation. Vasodilation drops down the localized systemic vascular resistance; affects the speed of the blood flow, and the blood flow in the infected area slows down. Second, the permeability of the blood vessels increases so that the immune system can confront the peripheral infective material easily. It’s imperative in the context of this chapter that the blood pressure— in the mathematical sense— is the product of cardiac output and systemic vascular resistance, and so the tissue perfusion is. Hence, the lesser the systemic vascular resistance, the lesser the blood pressure, and the lesser the tissue perfusion [54, 55].

The decrease in tissue perfusion is even exacerbated by the increased permeability of the blood vessels since the fluid can reach out and build around the tissue, and eventually, makes it challenging for oxygen to diffuse through the fluids and get to the cells. This exacerbated tissue perfusion is the cardinal reason behind the shock. Third, when the white blood cells interact with the pathogens, they release lytic enzymes, as well as reactive oxygen species, to exterminate the infective materials. These enzymes damage not only the pathogens but also the blood vessels to some extent, resulting in serious complications. Now, when the blood vessels get ruptured, protein helps to cause clotting to patch them up due to the coagulation factor in the blood. It may initially preclude the blood from spilling into the extravascular space, but later, some of these clottings break off into the bloodstream and let the blood spill out of the blood vessels; resulting in disseminated intravascular coagulation. Since this complication is disseminated everywhere in the body, the damaging enzymes and cytokines, associated with different immune molecules, may damage the blood vessels in the lungs as well. Damage and rupture in all the blood vessels in the lungs seriously affect the oxygen absorption into the bloodstream and end up in acute respiratory distress syndrome. It causes severe respiratory distress since
the respiratory system can no longer pull in oxygen into their bloodstream from
the environment. On the note of cardiac output, the human body initially
pushes to increase the cardiac output to compensate for the decreased systemic
vascular resistance to keep the blood pressure the same. However, if remained
untreated for long enough, the septic shock will go on, and the cardiac output
will eventually start to be depressed; resulting in a serious decrease in cardiac
output [55, 54, 56, 57]. These pathophysiological incidents—caused by sepsis—are
reflected in the physiological parameters as clinical clues, hence commonly
named as symptom distributives. Though highly elusive in nature, the entire
purpose of the Sepsis-3 and Sepsis-2 is to capture the underlying symptom
distributives that matter the most.

2.1.2 Bedside Monitoring: qSOFA vs. SIRS

Sepsis, unlike most of the diseases, is not a particular disease, but rather
a syndrome consorted with the ambiguous pathobiology and the absence of gold
standard diagnostic tests for assessments [7, 30]. Therefore, numerous endeavors
have been reported to capture the pathobiology, pathophysiology, and
epidemiology of sepsis to explain the syndrome. An initial definition of sepsis
(Sepsis-I) was introduced at a 1991 Consensus Conference that described sepsis
as a systemic inflammatory response syndrome (and, abbreviated as SIRS).
Addressing the limitations of Sepsis-1, the 2001 task force extended the list of
diagnostic criteria for sepsis (Sepsis-2). Sepsis-2 (SIRS) identified four criteria,
namely; fever or hypothermia (temperature, T >100.4°F or <96.8°F), tachypnea
(respiratory rate, RR > 20 breaths/minute), tachycardia (heart rate, HR > 90
beats/minute), and white blood cell count, WBC >12000/mm3 or <4000/mm3
(or, > 10% immature bands) [58]. In particular, Sepsis-2 interpreted sepsis as a
cascaded disease primarily diagnosed as systemic inflammatory response
syndrome, then followed by sepsis, severe sepsis, and septic shock. At the very end of the spectrum, patients may experience multiple organ dysfunction syndromes, an incurable stage of sepsis. Table 2.1 illustrates the parameters and cascaded development of sepsis as per the systemic inflammatory response syndrome criteria. However, this definition failed to distinguish sepsis from the other uncomplicated infections and diseases that exhibit themselves with the identical criteria; and indispensably failed to define what sepsis really is [7].

The task force also coined definitions for severe sepsis and septic shock; interpreting severe sepsis as sepsis being complicated by the organ dysfunction and septic shock as sepsis-induced hypotension persisting despite sufficient fluid resuscitation [59].

With significant advancements in the understanding of sepsis pathophysiology and pathobiology, after nearly two decades, a new definition of sepsis was proposed at the Third International Consensus in 2016 [7]. Now, sepsis (Sepsis-3) is defined as a syndrome pertaining to a life-threatening organ dysfunction introduced by a dysregulated host response to microorganism. According to the definitions of Sepsis-3, SOFA score (criteria) is used in the ICU

<table>
<thead>
<tr>
<th>Parameters/Criteria</th>
<th>Phases of syndrome development</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Criterion 1</strong>: Body Temperature &gt;100.4°F or &lt;96.8°F</td>
<td><strong>Phase 1</strong>: SIRS ≥ 2 criteria</td>
</tr>
<tr>
<td><strong>Criterion 2</strong>: Respiratory Rate &gt;20 breaths/minute (or PaCO2 &lt;32 mmHg)</td>
<td><strong>Phase 2</strong>: Sepsis (SIRS + suspected or confirmed infection)</td>
</tr>
<tr>
<td><strong>Criterion 3</strong>: Heart Rate &gt; 90 beats/minute</td>
<td><strong>Phase 3</strong>: Severe sepsis (sepsis + organ dysfunction)</td>
</tr>
<tr>
<td><strong>Criterion 4</strong>: White blood cell count &gt;12,000/mm³ or &lt;4000/mm³ (or &gt;10% bands)</td>
<td><strong>Phase 4</strong>: Septic shock (severe sepsis + persistent hypotension)</td>
</tr>
</tbody>
</table>

**Final Phase**: Multiple Organ Dysfunction

Reported ≥ 2 organs failing
Table 2.2: Quick Sequential Organ Failure Assessment.

<table>
<thead>
<tr>
<th>qSOFA</th>
<th>Respiratory Rate</th>
<th>Altered Mental Status</th>
<th>Low Blood Pressure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>breaths per minute ≥22</td>
<td>Glasgow Coma Scale ≤13</td>
<td>Systolic Blood Pressure ≤100 mm Hg</td>
</tr>
<tr>
<td>Any 2 of 3 qSOFA criteria or more + suspected or confirmed infection can be diagnosed as sepsis</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To determine the extent of patients’ organ functions (dysfunction) [27]. Besides that, a person with a suspected infection can promptly be identified at the bedside using qSOFA (Sepsis-3) score. qSOFA requires to satisfy at least two of the following criteria to be diagnosed as patients likely to have poor outcome due to sepsis [53]. Table 2.2 delineates the qSOFA criteria in a nutshell.

Several studies have—intending to leverage the greater consistency of Sepsis-3 in clinical trials and epidemiologic studies—developed predictive machine learning models using the qSOFA parameters. [59] found that qSOFA score—compared to SIRS criteria—showed higher prognostic accuracy for mortality and organ failure. Moreover, in predicting mortality and ICU-free days, qSOFA rendered considerably better discrimination in comparison with SIRS [60]. [61, 62] observed substantial evidence to support employing SOFA and qSOFA in the ICU sepsis diagnosis and treatment workflow over SIRS criteria. However, numerous studies implied the contrary and asserted that qSOFA manifests inconsistent performance in the mortality prediction [53]. [60, 61, 62] reported that qSOFA showed poor sensitivity and inconsistent precision in the predictive models. Though counterintuitive to some extent, [60, 63] indicated that qSOFA took much longer time in the patients’ trajectory, in comparison with SIRS, to identify patients with sepsis, which further delayed the initiation of medical interventions in ICU, and thereby subjecting the patients at a higher risk of developing septic shock and multiple organ dysfunction. Considering these stark contrasts evinced in the results (reflect by the evaluation metrics such as
Table 2.3: Comparing Sepsis-2 and Sepsis-3 (SIRS vs. qSOFA)

<table>
<thead>
<tr>
<th>Sepsis-2</th>
<th>Sepsis-3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sepsis</strong></td>
<td>Suspected or Confirmed Infection + SIRS</td>
</tr>
<tr>
<td><strong>Severe Sepsis</strong></td>
<td>Suspected or Confirmed Infection + qSOFA ≥ 2</td>
</tr>
<tr>
<td>+ Organ Dysfunction (lab markers, i.e. hypoxia, hypotension, elevated lactate)</td>
<td><strong>Category Removed</strong></td>
</tr>
<tr>
<td><strong>Septic Shock</strong></td>
<td>Sepsis + Persistent Hypotension (after adequate fluid resuscitation)</td>
</tr>
<tr>
<td></td>
<td>Sepsis + Vasopressors to maintain MAP ≥ 65 mm Hg + Serum lactate level &gt; 2 mmol/L</td>
</tr>
</tbody>
</table>

accuracy, sensitivity, precision, and G-mean) of predictive modeling using SIRS and qSOFA parameters; in this chapter, we decided to take a step back and have a more in-depth look at the qSOFA and SIRS parameters, and their underlying attributes and interrelations. Multicollinearity among parameters often intensifies the tension between optimization and generalizability, and eventually, leads to model overfitting. Model Overfitting hampers the generalizability of discriminant functions [51]. Moreover, it indicates that a small deviation in the input data can result in considerable, sometimes aberrant, changes in the model, and even leading to change in the sign of parameter estimates [51, 53]. Table 2.3 compares the SIRS criteria and qSOFA criteria; and points out the changes brought in Sepsis-3 from Sepsis-2 throughout all the cascaded steps.

2.2 Data and Research Design

We use MIMIC-III (Version 1.4), a publicly available ICU patient database, for this study. The data, ranging from 2001 to 2012, involves 58,976 distinct hospital admissions. Next, for the purpose of our study, we use the parameters of the quick Sequential Organ Failure Assessment (qSOFA), as well as Systemic Inflammatory Response Syndrome (SIRS), to identify all the ICU
patients who had been diagnosed with sepsis or were most susceptible to the disease. Then, we analyze the qSOFA and SIRS parameters of these identified sepsis patients, or the patients gone through sepsis screening, to study the intra-relationship between the qSOFA parameters and SIRS parameters. In our population, we found that 1,994 hospital admissions are diagnosed for sepsis among 58,976 overall admissions from 2001 to 2012. We observed that among 1,994 patients, the mortality is 22% (421). For this research design, our selection criteria include identifying the unique key for the critical parameter records and omittable parameters that we deem as bias-free for the purpose of this study, such as patients’ gender, data store time, and deidentified date of birth in the case of sepsis. During research design and data wrangling, we confront missing data and outlier values that are not biologically reasonable, albeit not for a considerable amount of records. This modicum amount of unexpected data points opens up the possibility of two distinct research designs. First, we can ignore the observations that have such data point(s) because they are of negligible number compared to the total observations we have. Second, we can follow the conventional central-value imputation or multiple imputations by chained equations to handle the missing data. A multi-blind Delphi process, convened by Ubicomp Lab of the Department of Computer Science at Marquette University and Regenstrief Center for Healthcare Engineering at Purdue University, came into a decision that ignoring the observations that have such unexpected data point(s) will be more suitable for the purpose of this study which requires to avoid imputation-bias. Besides that, outlier values—which are not biologically reasonable—are excluded, considering them as mistaken data entries in the ICU.

To realize the prevalence and underlying relation of the qSOFA and SIRS parameters, our research design identifies 13,783,035 patient records
(Chartevent) from 330,712,483 records (Chartevent) available in MIMIC-III (Version 1.4), which are unique for each Hospital Admission Id and chart time and pertaining to the patients gone through sepsis diagnosis. Then, to identify the most prevalent qSOFA and SIRS criteria, we selected 540,953 patient records and 770,368 patient records respectively (RR is common in both cases). Figure 1 summarizes the research design in simple. We use MIMIC-III (Version 1.4), a publicly available ICU patient database, for this study. The data, ranging from 2001 to 2012, involves 58,976 distinct hospital admissions. Next, for the purpose of our study, we use the parameters of the quick Sequential Organ Failure Assessment (qSOFA), as well as Systemic Inflammatory Response Syndrome (SIRS), to identify all the ICU patients who had been diagnosed with sepsis or were most susceptible to the disease. Then, we analyze the qSOFA and SIRS parameters of these identified sepsis patients, or the patients gone through sepsis screening, to study the intra-relationship between the qSOFA parameters and SIRS parameters. In our population, we found that 1,994 hospital admissions are diagnosed for sepsis among 58,976 overall admissions from 2001 to 2012. We observed that among 1,994 patients, the mortality is 22% (421). For this research design, our selection criteria include identifying the unique key for the critical parameter records and omittable parameters that we deem as bias-free for the purpose of this study, such as patients’ gender, data store time, and deidentified date of birth in the case of sepsis. During research design and data wrangling, we confront missing data and outlier values that are not biologically reasonable, albeit not for a considerable amount of records. This modicum amount of unexpected data points opens up the possibility of two distinct research designs. First, we can ignore the observations that have such data point(s) because they are of negligible number compared to the total observations we have. Second, we can follow the conventional central-value
imputation or multiple imputations by chained equations to handle the missing data. A multi-blind Delphi process, convened by Ubicomp Lab of the Department of Computer Science at Marquette University and Regenstrief Center for Healthcare Engineering at Purdue University, came into a decision that ignoring the observations that have such unexpected data point(s) will be more suitable for the purpose of this study which requires to avoid imputation-bias. Besides that, outlier values– which are not biologically reasonable– are excluded, considering them as mistaken data entries in the ICU.

To realize the prevalence and underlying relation of the qSOFA and SIRS parameters, our research design identifies 13,783,035 patient records (Chartevent) from 330,712,483 records (Chartevent) available in MIMIC-III (Version 1.4), which are unique for each Hospital Admission Id and chart time and pertaining to the patients gone through sepsis diagnosis. Then, to identify the most prevalent qSOFA and SIRS criteria, we selected 540,953 patient records and 770,368 patient records respectively (RR is common in both cases). Figure 2.1 summarizes the research design in simple.

Then, to assure the consistency and interpretability of the result while determining the most prevalent sepsis scenario, our selection criteria only filter in those chart times for which we have observations for all the three qSOFA parameters since the observation frequency varies with the parameters based on the intricacy involved in measurement. For instance, observations for altered mental status (AMS-GCS) are less frequently recorded than that of the respiratory rate. More importantly, since sepsis is a spectrum disease, studying and comparing the observations for different parameters in different record-times for a particular patient can confound the result and its interpretability. For the same reason, studying the parameters that are observed in the same time can capture the patients’ disease trajectory more consistently. Then, while
Figure 2.1: A simple outline for research design.

determining the most prevalent sepsis scenario for SIRS, our selection criteria only filter in those chart times for which we have observations for all the four parameters (Temperature, Heart Rate, Respiratory Rate, and White Blood Cell Count). The white blood cell count observations are considerably less frequent compared to the other three parameters involved and that is why the observations considered for the SIRS criteria are substantially lesser than that of the qSOFA criteria. Then, our research design intended to address two possible selection biases. First, it is intuitive that the more the patient stays in the ICU, the more the observations we have from that particular patient. This may influence the result of our study to some extent if there is a considerable majority between the higher LOS and lower LOS patients. Second, when evaluating the respiratory rate for ICU patients, there may have a possible blend in the data between the intubated breathing and natural patient breathing. However, the possibility of these two selection biases comes with the opportunity to test the intra-generalizability of the result of this study (both for qSOFA and
SIRS). That is why, in the second phase of this study, we dissect our data for only the first observations of each hospital admission. This research design is grounded in statistical theory such that the result of this study can help in developing multiparameter-intelligent sepsis prediction or treatment models that require predictors exhibiting the least or no collinearity.

2.3 Result

2.3.1 Analyzing Statistical Distributions: qSOFA and SIRS

Our study— for the qSOFA criteria— demonstrates the distribution of Systolic Arterial Blood Pressure, Glasgow Coma Scale measurement, and Respiratory Rate with the mean of 116.4 mm Hg, 11.17 (unitless) (numerical mean, Glasgow Coma Scale measurement only takes discrete integer values), and 21.07 breaths/min respectively, and a standard deviation of 24.78 mm Hg, 3.66 (unitless), and 6.52 breaths/min, respectively, for the first phase of the study. The median for Systolic Arterial Blood Pressure, Glasgow Coma Scale measurement, and Respiratory Rate are identified as 114.0 mm Hg, 11.00 (unitless), 21.00 breaths/min respectively, and the interquartile ranges are 100-131 mm Hg, 9-15 (unitless), 17-25 breaths/min respectively. Then, the second phase of our study for the qSOFA criteria, when we only consider the first observation of each hospital admission, demonstrates the distribution of Systolic Arterial Blood Pressure, Glasgow Coma Scale measurement, and Respiratory Rate with the mean of 106.7 mm Hg, 11.53 (unitless), and 20.48 breaths/min respectively, and a standard deviation of 37.62 mm Hg, 4.32 (unitless), and 6.16 breaths/min, respectively. The median for Systolic Arterial Blood Pressure, Glasgow Coma Scale measurement, and Respiratory Rate are identified as 110.0 mm Hg, 14.00 (unitless), 20 breaths/min respectively, and the interquartile ranges are 96.0-126.0 mm Hg, 8-15 (unitless), 16-24 breaths/min respectively.
Our study— for the SIRS criteria— demonstrates the distribution of Heart Rate, Respiratory Rate, Temperature, and White Blood Cell Count with the mean of 89.1 beats/minute, 21.07 breaths/min, 98.37°F, and 13.14 K/mm$^3$ respectively, and a standard deviation of 18.61 beats/minute, 6.52 breaths/min, 1.57°F, and 7.30 K/mm$^3$, respectively, for the first phase of the study. The median for Heart Rate, Respiratory Rate, Temperature and White Blood Cell Count are identified as 87 beats/minute, 21.00 breaths/min, 98.30°F, and 11.70 K/mm$^3$ respectively, and the interquartile ranges are 76-100 beats/minute, 17-25 breaths/min, 97.30-99.30°F, and 8.10-16.70 K/mm$^3$ respectively. Then, the second phase of our study for the SIRS criteria, when we only consider the first observation of each hospital admission, demonstrates the distribution of Heart Rate, Respiratory Rate, Temperature, and White Blood Cell Count with the mean of 95.58 beats/minute, 20.48 breaths/min, 98.25°F, and 14.34 K/mm$^3$ respectively, and a standard deviation of 20.76 beats/minute, 6.16 breaths/min, 2.01°F, and 8.28 K/mm$^3$, respectively. The median for Heart Rate, Respiratory Rate, Temperature and White Blood Cell Count are identified as 94 beats/minute, 20 breaths/min, 98.20°F, and 12.80 K/mm$^3$ respectively, and the interquartile ranges are 80-109 beats/minute, 16-24 breaths/min, 97.00-99.50°F, and 8.50-18.90 K/mm$^3$ respectively. Statistical Distributions for qSOFA and SIRS are summarized in Table 2.4.

Kernel density estimation distribution for the qSOFA criteria (systolic arterial blood pressure, altered mental status in Glasgow Coma Scale, and respiratory rate) and SIRS criteria (heart rate, respiratory rate, temperature, and white blood cell count) are depicted in Figure 2.2 to investigate the prevalent sepsis parameter. Visual statistics evince that most of the patients’ observations did not meet the qSOFA criterion for systolic arterial blood pressure Figure 2.2(a). The distribution for systolic arterial blood pressure
Table 2.4: Statistical Distributions for qSOFA and SIRS.

<table>
<thead>
<tr>
<th>Statistical Distributions for qSOFA (for entire patients’ trajectory)</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic Arterial Blood Pressure (mm Hg)</td>
<td>116.4</td>
<td>24.78</td>
<td>114.0</td>
<td>100-131</td>
</tr>
<tr>
<td>Glasgow Coma Scale (unitless)</td>
<td>11.17</td>
<td>3.66</td>
<td>11.00</td>
<td>9-15</td>
</tr>
<tr>
<td>Respiratory Rate (bpm)</td>
<td>21.07</td>
<td>6.52</td>
<td>21.00</td>
<td>17-25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistical Distributions for qSOFA (for patients’ first observation only)</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systolic Arterial Blood Pressure (mm Hg)</td>
<td>106.7</td>
<td>37.62</td>
<td>110.0</td>
<td>96.0-126.0</td>
</tr>
<tr>
<td>Glasgow Coma Scale (unitless)</td>
<td>11.53</td>
<td>4.32</td>
<td>14.00</td>
<td>8- 15</td>
</tr>
<tr>
<td>Respiratory Rate (bpm)</td>
<td>20.48</td>
<td>6.16</td>
<td>20.00</td>
<td>16.00-24.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistical Distributions for SIRS (for entire patients’ trajectory)</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Rate (bpm)</td>
<td>89.1</td>
<td>18.61</td>
<td>87</td>
<td>76-100</td>
</tr>
<tr>
<td>Respiratory Rate (bpm)</td>
<td>21.07</td>
<td>6.52</td>
<td>21.00</td>
<td>17-25</td>
</tr>
<tr>
<td>Temperature (° F)</td>
<td>98.37</td>
<td>1.57</td>
<td>98.30</td>
<td>97.30-99.30</td>
</tr>
<tr>
<td>White Blood Cell Count (K/mm³)</td>
<td>13.14</td>
<td>7.30</td>
<td>11.70</td>
<td>8.10-16.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistical Distributions for SIRS (for patients’ first observation only)</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Rate (bpm)</td>
<td>95.58</td>
<td>20.76</td>
<td>94.00</td>
<td>80.00-109.00</td>
</tr>
<tr>
<td>Respiratory Rate (bpm)</td>
<td>20.48</td>
<td>6.16</td>
<td>20.00</td>
<td>16.00-24.00</td>
</tr>
<tr>
<td>Temperature (° F)</td>
<td>98.25</td>
<td>2.01</td>
<td>98.20</td>
<td>97.00-99.50</td>
</tr>
<tr>
<td>White Blood Cell Count (K/mm³)</td>
<td>14.34</td>
<td>8.28</td>
<td>12.80</td>
<td>8.50-18.90</td>
</tr>
</tbody>
</table>

implies that most of the observations were in the range of 100-125 mmHg. From the clinical point of view, a systolic arterial blood pressure observation in a range of 100-125 indicates healthy blood pressure. Then, in Figure 2.2(a), Glasgow Coma Scale measurement distribution indicates a significant portion of these observations were in the safe zone (15 and 14). However, as the Glasgow Coma Scale ranges from 1 to 15, and the domain of consideration for the not-safe zone (qSOFA, 1-13) and the domain of consideration for the safe zone (14-15) are significantly disproportionate; and the visual analytics may be confusing to interpret this. In the next subsection, the explicit numerical interpretation is provided to understand the prevalence. Now, in the case of respiratory rate in 2.2(a), it is critical to interpreting whether the majority of the observation was in the qSOFA criterion or not, though it is evident that most of the data ranges
Figure 2.2: Kernel density estimation distribution of (a) qSOFA and (b) SIRS parameters to understand the prevalence.

between 15 and 24 breaths per minute. From the clinical point of view, at a rest state, a respiratory rate observation of 12-20 breaths per minute considers a healthy respiratory rate.

Moving to the SIRS criteria in Figure 2.2(b), we observe that the distribution for heart rate observations was less confounding for visual analytics in inferring prevalence—more of the kernel density was below the criterion margin (90 beats per minute), which indicates more healthy observations. In the case of respiratory rate measurement, it is worth mentioning to understand the difference that the cut-off for the SIRS criteria is different than that of the qSOFA criteria. For SIRS criteria, the criterion cut-off is 20 breaths per minute, and anything above that is considered as tachypnea. It is visually discernible that as the cut-off is shifted left (from 22 to 20) for SIRS, the more patients’ observations now met the sepsis criteria. For body temperature, it can be interpreted as a band: the observations inside two temperature cut-offs indicate the density of the healthy observations, and they represent a significant portion of the distribution. After that, in the case of white blood cell count, as the
domain of consideration for the not-safe zone and the domain of consideration for the safe zone are significantly disproportionate; the visual analytics may be confusing to imply prevalence. However, we can infer that the majority of observations met the SIRS criteria. In the next subsections, the specific numerical interpretation is presented to understand the prevalence and underlying statistical relation between the predictors. Understanding the prevalence of qSOFA criteria (so as SIRS) and the underlying relation of the parameters have important implications for sepsis treatment initiatives in the ICU and informing hospital resource allocation.

2.3.2 Analyzing Patients’ Entire Trajectory (qSOFA)

Kernel density estimation distribution of qSOFA parameters for both safe and qSOFA criterion met observations are presented in Figure 2.3 to understand the prevalent qSOFA parameters. Our study reveals that 25.12% of the Systolic Arterial Blood Pressure observations, 59.28% of the Glasgow Coma Scale measurements, and 45.11% of the respiratory rate observations met the qSOFA criterion. It is intuitive from the qSOSA criteria if we can have an understanding of the most prevalent criterion from observational studies, it can help practitioners and researchers in further factorial experiments. This observational study entirely relies on passive retrospective observation without assigning any further treatment. This study evinces that altered mental status is the most prevalent qSOFA criterion experienced in ICU. After that, this study addresses a nearly double-barrelled question: determining the most prevalent sepsis scenario in the ICU. The study reports that 28.19% of the observations (when three measurements available at the same time) have the two-factored qSOFA of high respiratory rate and altered mental status (among $3C3 + 3C2 = 4$ possibilities), resulting in this pair as the most prevalent qSOFA (Sepsis-3) scenario in the
Figure 2.3: Kernel density estimation distribution of qSOFA parameters for both safe and qSOFA criterion met observations to understand the prevalent qSOFA parameters.

ICU. It is to note that no-sepsis is another possible scenario besides these four possible qSOFA scenarios in the ICU (in our observations as well).

After that, a facet grid plot is illustrated on qSOFA parameters in Figure 2.4 to capture the most prevalent sepsis scenario and the underlying relation among the parameters. Figure 2.4 has multiple implications. However, the most obvious one is to compare the Pearson correlation coefficients (absolute) of each of the qSOFA parameters’ pairs. Our findings include the absolute Pearson correlation coefficients for respiratory rate-Glasgow Coma Scale measurement, Glasgow Coma Scale measurement-systolic arterial blood pressure, and respiratory rate-systolic arterial blood pressure are 0.09, 0.07, and 0.04, respectively. These insignificant correlation coefficients nullify the possibility of any linear correlation among the qSOFA parameters; and thereby ensuring that multicollinearity does not exist between the parameters. Understanding this relationship can help in developing predictive models as it implies that the overdetermined system involved in the modeling is a full-ranked matrix (not rank-deficient). However, the lack of multicollinearity cannot guarantee that two random variables are statistically independent. Besides, it is also evident from the pathophysiology of sepsis that sepsis is a spectrum disease, and one predictor may influence another with the development of sepsis and septic shock.
Figure 2.4: Facet grid illustration of Sepsis-3 (qSOFA) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario in the ICU.
2.3.3 Analyzing Patients’ Entire Trajectory (SIRS)

After that, kernel density estimation distribution of SIRS parameters for both safe and SIRS criterion met observations are presented in Figure 2.5 to understand the prevalent SIRS parameters. Our study reveals that 43.30% of the heart rate observations, 50.89% of the respiratory rate observations, 23.08% of the body temperature observations, and 53.12% of the white blood cell count observations met the SIRS criterion. Though both the white blood cell count and respiratory rate have a significant prevalence in the observations undergone through the sepsis screening, white blood cell count is the most prevalent SIRS criterion experienced in ICU. Besides, the study reports that 12.32% of the observations (when four measurements available at the same time) have the three-factored SIRS of tachypnea-high heart rate-high white blood cell count. It is critical to consider that there exist six possible pairs of combinations, 4 possible trios of combination, and one combination considering all the parameters as the possible sepsis scenario in the ICU. It is to note that no-sepsis is another possible scenario besides these eleven possible SIRS scenarios in the ICU (in our observations as well). Knowing the most prevalent criterion and sepsis scenario in the ICU for SIRS can help practitioners and researchers in diagnosis, treatment, and further factorial experiments.

Then, a facet grid plot is illustrated on SIRS (Sepsis-2) parameters in
Figure 2.6: Facet grid illustration of Sepsis-2 (SIRS) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario in the ICU.

Figure 2.6 to capture the most prevalent SIRS scenario and the underlying relation among the parameters. Figure 2.6 has multiple implications. However, the most obvious one is to compare the absolute Pearson correlation coefficients of each of the SIRS parameters. Our findings include the absolute Pearson correlation coefficients for heart rate-respiratory rate, heart rate-temperature, heart rate-white blood cell count, respiratory rate-temperature, respiratory rate-white blood cell count, and temperature-white blood cell count are 0.32, 0.34, 0.13, 0.11, 0.05 and 0.03, respectively. The insignificant absolute correlation coefficients invalidate the possibility of any correlation among the critical parameters; and thereby ensuring that multicollinearity does not exist between the parameters. However, the absolute correlation coefficients are not negligible at all in the case of heart rate-respiratory rate and heart rate-temperature pairs. Understanding this relationship can help in developing predictive models as it
implies that the overdetermined system involved in the modeling is a full-ranked matrix (not rank-deficient). However, the lack of multicollinearity cannot guarantee that two random variables are statistically independent.

### 2.3.4 Analyzing Patients’ First Observation Only (qSOFA)

In the second phase of this study, we dissect our data for only the first observations of each hospital admission. This may address two possible selection biases, including the opportunity to test the intra-generalizability of the result of this observational study. First, it is intuitive that the more the patient stays in the ICU, the more the observations we have from that particular patient. This may influence the result of our study to some extent if there is a considerable disproportion between the higher LOS and lower LOS patients. Second, when evaluating the respiratory rate for ICU patients, there may have a possible blend in the data between the intubated breathing and natural patient breathing.

Kernel density estimation distribution of qSOFA parameters for both safe and qSOFA criterion met observations are presented in Figure 2.7 to understand the prevalent qSOFA parameters. Our study reveals that 32.58% of the Systolic Arterial Blood Pressure observations, 44.54% of the Glasgow Coma Scale measurements, and 40.53% of the respiratory rate observations met the qSOFA criterion. This observational study entirely relies on passive retrospective observation without assigning any further treatment. This study evinces that altered mental status is the most prevalent qSOFA criterion experienced in the ICU. Besides, The study reports that 18.25% of the observations have the two-factored qSOFA of high respiratory rate and altered mental status (among $3C3 + 3C2= 4$ possibilities), resulting in this pair as the most prevalent qSOFA (Sepsis-3) scenario in the ICU. It is to note that no-sepsis is another possible scenario besides these four possible qSOFA scenarios in the ICU (in our
observations as well).

Then, a facet grid illustration is depicted on qSOFA parameters in Figure 2.8 to understand the most prevalent qSOFA scenario and the underlying relation among the parameters. Figure 2.8 has multiple implications. However, the most

Figure 2.7: Kernel density estimation distribution of qSOFA parameters for both safe and qSOFA criterion met patients at first observations to understand the prevalent qSOFA parameters.

Figure 2.8: Facet grid illustration of Sepsis-3 (qSOFA) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario of patients at first observations in the ICU.
obvious one is to compare the absolute Pearson correlation coefficients of each of the qSOFA parameters. Our findings include the absolute Pearson correlation coefficient for respiratory rate-Glasgow Coma Scale measurement, Glasgow Coma Scale measurement-systolic arterial blood pressure, and respiratory rate-systolic arterial blood pressure are 0.15, 0.01, and 0.02 respectively. These insignificant correlation coefficients invalidate the possibility of any correlation among the critical parameters; and thereby ensuring that multicollinearity does not exist between the parameters. However, the lack of multicollinearity cannot guarantee that two random variables are statistically independent.

2.3.5 Analyzing Patients’ First Observation Only (SIRS)

After that, kernel density estimation distribution of SIRS parameters for both safe and SIRS criterion met observations are presented in Figure 2.9 to understand the prevalent SIRS parameters. Our study reveals that 57.03% of the heart rate observations, 45.89% of the respiratory rate observations, 33.93% of the body temperature observations, and 60.57 of the white blood cell count observations met the SIRS criterion. This observational study entirely relies on passive retrospective observation without assigning any further treatment. This study evinces that white blood cell count is the most prevalent criterion experienced in the ICU, albeit considering that both the white blood cell count and respiratory rate have a significant prevalence. Besides, the study reports that 11.38% of the SIRS criteria met sepsis patients have the three-factored SIRS of tachypnea-high heart rate-high white blood cell count (among $4C4 + 4C3 + 4C2= 11$ possibilities), resulting in the pair as the most prevalent sepsis (SIRS) scenario in ICU. It is incumbent to take into account that there exist six possible pairs of combinations, four possible trios of combination, and one combination considering all the parameters as the possible sepsis scenario in the
Figure 2.9: Kernel density estimation distribution of SIRS parameters for both safe and sepsis criterion met patients at first observations to understand the prevalent SIRS parameters.

Figure 2.10: Facet grid illustration of Sepsis-2 (SIRS) parameters to capture the underlying relationship between parameters and the most prevalent sepsis scenario of patients at first observations in the ICU.

It is to note that no-sepsis is another possible scenario besides these eleven possible qSOFA scenarios in the ICU (in our observations as well). Knowing the most prevalent criterion and sepsis scenario at the first observation for SIRS can help practitioners and researchers in diagnosis, treatment, and further factorial experiments.

Then, a facet grid illustration is depicted on SIRS parameters in Figure
2.10 to understand the most prevalent sepsis (SIRS) scenario and the underlying relation among the parameters. Figure 2.10 has multiple implications. However, the most obvious one is to compare the absolute Pearson correlation coefficients of each of the SIRS parameters. The insignificant absolute correlation coefficients invalidate the possibility of any correlation among the critical parameters; thereby ensuring that multicollinearity does not exist between the parameters. However, the absolute correlation coefficients are not negligible at all in the case of heart rate-respiratory rate and heart rate-temperature pairs.

2.4 Discussion

2.4.1 Theoretical Reasoning

This study reveals that the altered mental status and systolic arterial blood pressure are the most and least prevalent qSOFA criteria, respectively. Mathematically, blood pressure is the product of systemic vascular resistance and cardiac output. Hence, seemingly with the decrease in systemic vascular resistance due to the vasodilation, blood pressure will drop down if the cardiac output remains the same. However, in practice, when the systemic vascular resistance drops down, the human body immediately tries to maintain the equilibrium for a few moments, and compensate it with the cardiac output. Cardiac output depends nonlinearly and proportionately on the respiratory rate; hence, the increase in the respiratory rate increases the cardiac output, and eventually maintains the equilibrium of the blood pressure for initial moments. After a particular time, that equilibrium breaks down, albeit the cardiac output (so as the respiratory rate) continually tries to meet the stability again. This fact advocates the possibility of respiratory rate to be more prevalent compared to systolic arterial blood pressure as a symptom. From the aspect of SIRS criteria, it is apparently intuitive why white blood cell count is the most
prevalent criteria. When the microorganism invades, the body’s immune response is triggered, and white blood cells appear immediately to rescue. Heart rate, respiratory rate, and temperature are the consequential symptoms associated with it. As sepsis is a spectrum disease, and one predictor may influence another with the development of sepsis and septic shock, though they are not linearly correlated. The findings of this observational study support the established pathophysiology of sepsis described in the background section. This observational study entirely relies on passive retrospective observation without assigning any treatment.

2.4.2 Medical Informatics Implications

Quantifying the prevalence of the qSOFA criteria of sepsis-3 in comparison with the SIRS criteria of sepsis-2 and understanding the underlying relationships among these parameters provides significant inferences for sepsis treatment initiatives in the ICU and informing hospital resource allocation. Understanding prevalent sepsis criteria and associated distributions can help planning triage the patients and put a graded urgency category. It may help plan to sort out the suggested geographic location the patients need to be in, starting from the emergency department pod to the geographic location inside the hospital—general ward, intermediate care unit, high dependent unit, intensive care unit. It can also help determine the intensity of nursing care, nurse-patient ratio, frequency of mandated nursing follow-up monitoring. The study result has the potential to guide the clinicians to follow specific laboratory testing and pharmacological interventions, including the decision-making process in antibiotic use. We know that there is strong evidence in support of the role of early antibiotics in reducing mortality in sepsis.
2.4.3 Research Opportunities

Though MIMIC-III is an extensive critical care database, it is a single-center database comprising of critical-care-unit electronic health record data of Beth Israel Deaconess Medical Center in Boston. Regardless of the myriad number of patients’ data, the findings that are valid for the Beth Israel Deaconess Medical Center in Boston may not be useful for the other medical centers and critical care units. The epidemiology and treatment facility varies with the hospitals, states, and infrastructures of countries. Epidemiology and treatment facility has a significant impact on the patients’ outcome, so on the patients’ symptom distributives. On the flip side, this observational study entirely relies on passive retrospective observation, and the dynamics of the treatment and medicine advances with time and research. Besides the prevalence of the physiological parameters, this time-variability and resource-variability may also affect the interrelation nature among parameters. Also, it may vary if we study from the individual aspect. Though the collective analysis infers no-multicollinearity among parameters, there may be a possibility that one patient shows strong multicollinearity. Again, the parameters measured may vary according to the therapeutics undertaken in the intensive care unit. For instance, the Glasgow Coma Scale score may become low due to sedation, or catecholamines may be responsible for healthy blood pressure, or mechanical ventilation may affect the respiratory rate. Any predictive modeling and treatment plan should take this variability and uncertainty into account.

This uncertainty around its generalizability opens up the research opportunity for the researchers in the health informatics domain in three possible directions.

- Does this finding hold its generalizability while integrating data from
multiple electronic health records?

- How can we study confounding variables induced by numerous groups of people with different characteristics?

- How can these findings address the confounding medical interventions in sepsis treatment?

Apart from that, the comparison between qSOFA and SIRS can be extended to comparing SOFA and qSOFA, or SIRS and SOFA, or all the three criteria available, to understand the underlying interrelation between the parameters.

2.5 Summary

2.5.1 Objective

The objective of this study is threefold. First, we aim to unpack the most prevalent criterion for sepsis (for both sepsis-3 and sepsis-2 predictors). Second, we intend to determine the most prevalent sepsis scenario in the ICU (among four possible scenarios for qSOFA and eleven possible scenarios for SIRS). Third, we investigate the underlying relationship among qSOFA predictors and SIRS predictors.

2.5.2 Methods

We conducted this observational study using MIMIC-III (Version 1.4): the critical care database of MIT. We took the qSOFA (sepsis-3) and SIRS (sepsis-2) parameters into account for patients admitted to the critical care units (2001-2012) in Beth Israel Deaconess Medical Center in Boston to understand the prevalence and underlying relation between these parameters among the patients undergone sepsis screening. We adopted a multi blind Delphi method to
seek a rationale for decisions in several stages of research design regarding handling missing data and outlier values, statistical imputations and biases, and generalizability of the study.

2.5.3 Results

This study reveals that the altered mental status in Glasgow Coma Scale (59.28%, 38854 of 65545 observations) is the most prevalent Sepsis-3 (qSOFA) criterion, and the white blood cell count (53.12%, 17163 of 32311 observations) is the most prevalent Sepsis-2 (SIRS) criterion confronted in the ICU. Besides, the two-factored sepsis of high respiratory rate (RR $\geq$ 22 bpm) and altered mental status (28.19%, among four possible qSOFA scenarios besides no-sepsis) is the most prevalent sepsis-3 (qSOFA) scenario, and the three-factored sepsis of tachypnea, high heart rate and high white blood cell count (12.32%, among eleven possible scenarios besides no-sepsis) is the most prevalent sepsis-2 (SIRS) scenario in the ICU. Apart from that, the insignificant absolute Pearson correlation coefficients nullify the likelihood of any linear correlation among the critical parameters, thereby assuring that multicollinearity does not exist between the parameters. However, the absence of multicollinearity cannot guarantee that two random variables are statistically independent.

2.5.4 Conclusions

Quantifying the prevalence of the qSOFA criteria of sepsis-3 (so as SIRS of sepsis-2) and understanding the underlying relation among the parameters have significant inferences for sepsis treatment initiatives in ICU and informing hospital resource allocation. Besides, these data-driven results offer design implications for multi-parameter intelligent sepsis prediction in the ICU.
Chapter 3

State-of-the-art Medical Informatics Solutions Related to Sepsis: Existing Pitfalls and Potential Challenges

3.1 State-of-the-art

In this chapter, we provide an overview of current published literature relevant to medical informatics solutions, existing research gaps, potential pitfalls, and challenges in machine learning-based predictive models developed in an effort to address sepsis, including diagnosis, monitoring, and intervention. Here, we discuss the state-of-the-art from four aspects of the problem: Automated Sepsis Risk Assessment System, Predicting Sepsis, Septic Shock, and Sepsis-associated Poor Outcomes, Biomarker Discovery, and Optimizing Therapeutics.

3.1.1 Automated Sepsis Risk Assessment System:

With the advent of Electronic Medical Record, an increasing number of reports are available in the literature that endeavored to use information systems for sepsis management. In 2009, for early identification of older emergency department patients with possible life-threatening infection such as systemic inflammatory response syndrome, Meurer et al. developed and characterized an automated syndromic surveillance mechanism [64].

Then, in 2010, Berger et al. proposed a computerized alert screening for severe sepsis (According to Sepsis-2) in emergency department patients. This findings of this study shows an increase in lactate testing as a result of adopting this system. However, this report does not indicate any improvement in inpatient mortality [65].
As early therapy involving antibiotic administration and fluid resuscitation evinced improved patient outcomes in the case of sepsis and severe sepsis, Sawyer et al. came up with an idea of a tool in 2011 that may proactively identify patients at risk for developing sepsis, with the motivation to lessen intervention time and improve the patient outcomes. They implemented a real-time computerized sepsis alert system in nonintensive care unit patients. This report also evaluate if an automated sepsis screening and alert system can facilitate early appropriate interventions for sepsis patients [66].

After that, in 2011, Nelson, Smith, Jared, and Younger published a paper on prospective trial of real-time electronic surveillance to facilitate early care of severe sepsis (Sepsis-2). This paper hypothesize that their algorithm, followed by a real-time electronic medical record query as well as caregiver notification system, can increase the rate and timeliness of blood lactate sampling and blood cultures, facilitate the provision of antibiotics, and improve the performance of chest radiography in an effort to improving sepsis care [67].

Similar effort is seen in 2012 from Hooper et al. and they proposed a randomized trial of automated electronic monitoring system to facilitate early detection of sepsis (Sepsis-2) in the ICU. They concluded that the real-time alerts to physicians in one tertiary care ICU were feasible and reported as safe. However, it did not affect the measured therapeutic interventions or considerably change the clinical outcomes in the case of sepsis[68].

McRee, Thanavaro, Moore, Goldsmith, and Pasvogel studied the impact of an electronic medical record surveillance program in 2014. This study was sepsis specific and it evaluated primarily based on the outcomes for patients with sepsis. Their promising evidence advocated decrease in exacerbating effects of sepsis, severe sepsis and septic shock (Sepsis-2) by early identification of the problem and adequate treatment [69].
In 2014, Alsolamy et al. evaluated the diagnostic accuracy of a screening electronic alert tool for severe sepsis and septic shock. In this paper, the research design included patients from the emergency department for the consideration of this tool’s diagnostic efficacy. This paper bolsters its efficacy in early recognition and management as, in recognizing severe sepsis and septic shock (Sepsis-2), this electronic sepsis alert tool shows high sensitivity and specificity [70].

Nguyen et al. proposed an automated EMR sepsis detection in 2014, concentrating on emergency department cases. This system leverages the EMR system—which continually collects vital signs and laboratory test information on all patients admitted to the emergency department—and triggers a sepsis alert for those who met SIRS criteria and then for those who experienced more than one major organ dysfunction. The diagnostic accuracy of this EMR-based sepsis identification tool was evaluated at a renowned academic urban emergency department with 64,000 annual visits. The manual review of physicians, laboratory records, and the nursing staff’s output notes was used in this study. The study concluded this automated EMR-based detection has the potential to provide a viable strategy in the emergency department for identifying sepsis [71].

The comprehensive study by Umscheid et al. on the development, implementation, and impact of an automated early warning and response system for sepsis was a major contribution in 2015. In the first phase of the study, the research team presented the tool derivation and validation. After that, in the second phase, a pre-implementation/post-implementation study was presented with multivariable-adjustment-measured impact. It also included the screen positive, sensitivity, specificity, and positive and negative predictive values and likelihood ratios for the composite of ICU transfer, rapid response team call, and death in the derivation cohort. Authors affirmed that this tool identified at-risk patients with significant accuracy, and thus, prompted a bedside evaluation.
This paper advocates that this effort may result in more timely sepsis care, improved documentation, and hence, reduced mortality due to sepsis [72].

Development and validation of an automated sepsis risk assessment system published in research in nursing and health journal in 2016 is another major study. Promoting early detection of sepsis, followed by aggressive resuscitation, to reduce sepsis mortality, the authors developed and verified a tool named Auto-SepRAS. This tool for automated sepsis risk assessment can assess the sepsis risk of inpatients automatically by applying advanced data mining techniques leveraging EMR data. It provides daily updates to the caregivers. The algorithm assesses the patient condition compounding seven predictors, such as heart rate, respiratory rate, patient’s age, diastolic blood pressure, length of stay, etc. Then, based on the predictive values from this risk-scoring algorithm, Auto-SepRAS can categories the inpatients’ conditions into three risk levels (low, high, and moderate) [73].

All of these efforts aimed to prompt first-line interventions to treat sepsis. However, two possible reasons challenged their efficacy in practice. First, there may be a delay in the time between a caregiver following an appropriate procedure after receiving an alert. Moreover, this delay may be compounded by the time it takes to enter/update sepsis progression into EMR, as these systems give an alarm message only after vital signs (Sepsis-2) have changed. This scenario is evinced in the report published by Nelson et al. in 2011. It points out the fact that a sepsis screening system used in an emergency department sent most of the alerts when patients were beyond the early stages [67]. Second, defining "what sepsis really is" still remained an unanswered question in the medical fraternity. Most of the early warning techniques presented above were based on the second major modification in sepsis definition in 2001 (also known as Sepsis-2). The recent major change in definition (Sepsis-3) will question these
previously practiced techniques and tools’ efficacy [27].

Based on Sepsis-3 clinical criteria, John Karlsson Valik and his colleagues presented an observational study using EMR data in 2020. In this study, they endeavored to validate automated sepsis surveillance against physician record review. They demonstrated a retrospective study on the general hospital population, excluding data from the ICU stay. This automated sepsis surveillance is based on a rule-based algorithm, and to validate its performance, they classified a stratified random sample of one thousand hospital admissions according to the Sepsis-3 clinical criteria. The findings of this observational study advocate this algorithm’s efficacy while compared to physician medical record reviews in non-intensive care wards. This study also reveals variations in in-hospital-sepsis-onset incidences between wards [74].

The literature discussed on risk assessment tools developed for sepsis indicates a conspicuous trend of adopting data-driven algorithms. However, a discernable research gap is found since different host factors and pathogen factors were not addressed. Again, in many cases, methods presented for evaluating the performance of these assessment algorithms in routine clinical settings do not meet the current standard of scientific work. Furthermore, though we can carve out the differences between various approaches, this wide variety of studies may have very different study designs, recruiting approaches, sample sizes, and evaluation approaches, thereby precluding statistically analyzing the study outcomes. Besides, in several cases, study designs may be predisposed to biased results [75].

In [76], Ginestra et al. surveyed clinicians’ and caregivers’ perceptions of the Machine Learning-based early warning system’s overall efficacy, including the alert’s helpfulness and impact on patient care. This six-week study conducted in a tertiary teaching hospital in Philadelphia interviewed nurses and providers
after five months since the Early Warning System 2.0 alert was implemented. The results and discussions imply unsatisfactory clinical perceptions of Early Warning System 2.0. More importantly, there observed widely varied perceptions among providers and nurses about alerts’ benefits, challenging the acceptance of machine learning-based sepsis alerts.

3.1.2 Predicting Sepsis, Septic Shock, and Sepsis-associated Poor Outcomes:

In 2017, [77] developed a machine learning-based algorithm for severe sepsis prediction. In this study, the researchers presented this algorithm’s impact on patient survival and hospital length of stay through a randomized clinical trial. This randomized controlled clinical trial included two medical-surgical intensive care units at the University of California, San Francisco Medical Center. In this factorial open-label study, the care team evaluated the patient and initiated the severe sepsis bundle on receiving an alert, if appropriate. This study advocated the algorithm’s efficacy as it revealed that the average LOS was decreased by 2.7 days (from 13.0 days in the control to 10.3 days in the experimental group). As far as survivability is concerned, in-hospital mortality was reduced by 12.4 points in percentage. Though this study was the first randomized controlled trial of a sepsis surveillance system predicting severe-sepsis that demonstrated statistically significant differences in in-hospital mortality and average LOS, the recent shift in sepsis definition eliminated the term severe sepsis and may question the acceptance of its scientific contribution to some extent.

In 2017, CRMC (Cape Regional Medical Center), collaborating with Dascena, published their research endeavor on improving sepsis-related patient outcomes through a revised sepsis management approach. This endeavor
included forming a quality improvement team to implement a machine learning-based sepsis prediction algorithm to identify sepsis in patients earlier. The team ran their comparative study between two groups. First, the team evaluated all patients using SIRS screenings twice-daily to estimate sepsis-related outcomes. The team then implemented a machine learning algorithm-based predictive system tailored, incorporating feedback from the user (caregivers and healthcare experts) and integrated it into the current hospital workflow. They reported astonishing results in patients' sepsis-related outcome: 60.24% decrease in in-hospital mortality rate, 9.55% decrease in average sepsis-related LOS, and 50.14% decrease in sepsis-related 30-day readmission rate between pre-implementation cohort and post-implementation cohort [78].

After that, [79] proposed and validated a machine learning-based algorithm named InSight for predicting sepsis. This algorithm adopted the gradient tree boosting method and used a combination of six vital signs of patients to detect and predict three sepsis-related gold standards. The data used for this retrospective study is mixed-ward and multicentered to address the generalizability issue. The authors indicated that this algorithm outperformed existing sepsis scoring systems to identify and predict sepsis, severe sepsis, and septic shock and claimed it as the first sepsis screening system exceeding an AUROC of 0.90 using patients' vital sign inputs only. It also addressed the model robustness issues, missing data issues, and customization issues in new hospital data.

In 2018, [80] developed the AISE algorithm (Artificial Intelligence Sepsis Expert) for predicting sepsis early. An extensive literature study showed that, albeit early aggressive intervention with antibiotics is critical to avoid poor sepsis-related outcomes and improve survivability, there is no clinically validated system available for the real-time sepsis onset prediction. For their development
cohort, they considered more than thirty-one thousand ICU admissions from Emory University hospitals (taking two units into account for their study). Besides, for the validation cohort, they used the patients’ data from the MIMIC-III ICU database. For this study, they used the sepsis-3 definition and excluded any observations from the patients who met sepsis-3 criteria before or within 4 hours of ICU admission. A collection of 65 features, including vital signs time series and EMR data, were passed to the AISE algorithm and calculated on an hourly basis to predict sepsis onset in the proceeding 4, 6, 8 and 12 hours. This observational study also provides a list of significant factors that contribute the most. However, determining the clinical utility of this artificial intelligence-based sepsis prognostication model requires further investigation.

[81] summarized a particular segment of the PhysioNet/Computing in Cardiology Challenge 2019: Early Prediction of Sepsis from the Clinical Data. Participants were given 40,336 patient records for training their model from two distinct hospital systems. To test the model using independent data, 22,761 patient records from three distinct hospital systems were secluded. For feature selection, 40 measurements were available for each patient, including laboratory measurements, vital sign data, and demographics information. Participants were suggested to use sepsis-3 clinical criteria to define sepsis onset and asked to design an automated algorithm to predict sepsis onset six hours before clinical recognition of sepsis in the hospital. PhysioNet/Computing in Cardiology Challenge 2019 developed a clinical utility-based evaluation metric to evaluate each of the algorithms. This evaluation metric had provision for rewards for early sepsis predictions and penalty for late or missed predictions as well as false alarms. [81] presented analysis and implications from the outcome of this challenge. It also pointed out how late or missed predictions represented by false alarms can flummox the practitioners in the ICU intervention.
To quantify the performance of the machine-learning-based predictive models for sepsis, [82] conducted a meta-analysis combining a number of observational studies in 2019. Compared to the pooled AUROC for SIRS, MEWS, and SOFA of 0.70, 0.50, and 0.78, respectively, for predicting sepsis onset prior to 3 or 4 hours, their data revealed that machine learning models demonstrated the pooled AUROC 0.89 (with the sensitivity of 0.81 and specificity of 0.72). Though their findings advocated that the machine-learning-based approach exhibits better performance compared to the existing scoring systems in the case of predicting sepsis, it implied that a feasible tool for predicting sepsis remains elusive to the medical informatics researchers.

Concentrating on the pediatric sepsis (patients between 2–17 years of age) and sepsis-related poor outcomes, [83] endeavored to predict pediatric severe sepsis using machine learning algorithm and retrospective EMR data (of University of California San Francisco Medical Center) to facilitate early detection and optimize effective pediatric treatment. Their result is quite promising as a four-fold cross-validation evaluation revealed an AUROC of 0.916. It outperformed the pediatric SIRS and Pediatric Logistic Organ Dysfunction score in predicting severe sepsis four h prior to onset. Their algorithm claimed that it had enormous potential for high-performance severe sepsis detection and prediction if integrated into automated monitoring of EHR data.

In 2019, Shigehiko Schamoni et al. proposed non-circular machine learning for sepsis prediction by leveraging implicit expert knowledge. The authors addressed that since only a few hours delay of antibiotic treatment of patients with severe sepsis may be associated with increased mortality by a considerable margin, the insight that makes models for early sepsis prediction needs to be as effective as possible. Previous approaches found in the literature achieved high AUROC by learning from EMRs. However, in those cases, sepsis
labels were automatically defined by following established clinical criteria. This paper argued that incorporating clinical criteria to automatically define ground truth sepsis labels as severity scoring models’ features sometimes compromises the proposed approaches’ validity. Rather than entirely depending on the clinical criteria, the authors proposed to create independent ground truth for sepsis by exploiting implicit knowledge of clinical practitioners via an electronic questionnaire. That helped them to capture physicians’ daily judgments of patients’ sepsis status. Albeit in small data size (considering the stark contrast in data labeling capacity between this process and automated process), they achieved an improved AUROC score. Integrating physicians’ judgment facilitates the predictive model with learned weights, inferring potentially surprising feature contributions and interpreting seemingly counterintuitive findings [84].

A research outcome from the PhysioNet/Computing in Cardiology Challenge 2019 is published in [85] in 2020, aiming to develop and validate a machine learning algorithm that would be clinically interpretable and able to predict sepsis onset during critical care in a real-time fashion based on a retrospective observational cohort study. This time-phased model first estimates the likelihood of sepsis onset for each hour of an ICU stays in the following 6 hours. After that, it will make a binary prediction with three time-phased cutoff values. According to the guide of the challenge, the team used utility score to evaluate the model performance. The models achieved 0.430 and 0.354 in utility score on the internal validation set and test set, respectively. Though not having considerable accuracy, the contribution of this proposed time-phased machine learning model for Sepsis Prediction is achieving interpretability for real-time prediction of sepsis onset in critical care.
3.1.3 Biomarker Discovery:

Here, we will briefly discuss the works on discovering sepsis biomarkers. This discussion will capture the problem from different lenses.

In [86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99], cytokine/chemokine biomarkers, such as GRO-alpha, Osteopontin, IL-1 receptor antagonist, High Mobility Group-Box 1 protein, Macrophage inflammatory protein (MIP)-1 and- 2, IL-1b, Macrophage migration inhibitory factor (MIF), and Monocyte chemotactic protein (MCP)-1 and 2, are discussed for prognosticating sepsis and sepsis-related complicacy. [100] compared these sepsis markers by assessing whether they were evaluated in experimental studies, clinical studies, and how they are as a prognostic factor. The authors also included brief critical comments, such as whether it is higher in septic shock than in sepsis, whether it is strongly correlated with SOFA score, if it is increased in septic compared with non-septic patients, whether it is distinguished between survivors and non-survivors at Twenty-eight days, if it is increased in sepsis compared with healthy controls, and whether it is predictive of lethal outcome from postoperative sepsis, based on their research.

After that, the cell marker biomarkers for sepsis and sepsis-related complications are identified in [101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111]. The cell marker biomarkers include CD10, CD11b, CD11c, CD14 (cellular and soluble), CD18, CD25 (cellular and soluble), CD28 (soluble), CD163 (soluble), and mHLA-DR (soluble). [100] compared these cell marker sepsis biomarkers by assessing whether they were evaluated in experimental studies, clinical studies, and how they are as a prognostic factor. The authors also included brief critical comments, such as whether it is decreased in septic shock compared with healthy
controls, if it has a correlation with APACHE II and SOFA scores, if it Predicted
development of septic shock, and whether it is distinguished between survivors
and non-survivors at 28 days in patients with septic shock, based on their
research.

Then, [112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125,
126, 127, 128, 129] evinced works on receptor biomarkers to prognosticate
sepsis-related complications. These literature include CC chemokine receptor
(CCR) 2, CCR 3, C5L2, CRTh2, Fas receptor (soluble), Fc-gamma RIII, FLT-1
(soluble), GP130, IL-2 receptor (soluble), RAGE (soluble), ST2 (soluble, IL-1
receptor), Toll-like receptor (TLR) 2 and 4, Transient receptor potential vanilloid
(TRPV)1, TREM-1 (soluble), TNF-receptor (soluble), and Urokinase-type
plasminogen activator receptor (uPAR) (soluble). [100] compared these sepsis
markers by assessing whether they were evaluated in experimental studies,
clinical studies, and how they are as a prognostic factor. The authors also
included brief critical comments, such as whether it can predict the development
of MOF, if it is increased in sepsis compared with healthy controls, and strongly
correlated with APACHE II score, whether it is increased in septic compared
with non-septic critically ill patients, if it is distinguished between survivors and
non-survivors at 28 days, based on their research.

In [130, 131, 132, 133, 134, 135, 136, 137], coagulation biomarkers are
discussed for prognosticating sepsis and sepsis-related complicacy. These studies
covered Antithrombin, Activated Partial Thromboplastin Time, Fibrin,
Plasminogen Activator Inhibitor-1, Protein C and S, and Thrombomodulin as
the prognosticating factor for sepsis. [100, 137, 134] compared these sepsis
markers by assessing whether they were evaluated in experimental studies,
clinical studies, and how they are as a prognostic factor. The researchers also
included brief critical comments, such as whether it is higher in septic shock
than in sepsis, whether it is correlated with MOF score in patients with sepsis and DIC, whether it has a high negative predictive value, whether it is distinguished between survivors and non-survivors at Twenty-eight days, if it is increased in patients with Gram-negative bacteremia, and whether it predicted the development of MOF, DIC, and response to therapy, based on their research. [138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149] presented the sepsis biomarker pertaining to the vascular endothelial damage, including Angiopoietin, Endothelial leukocyte adhesion molecule, Endothelial progenitor cells, Intracellular adhesion molecule, Neopterin, Vascular endothelial growth factor, Von Willebrand factor and antigen, Endocan, ADAMTS-13, Intracellular adhesion molecule, Laminin, P-Selectin, L-Selectin (soluble), and Vascular cell adhesion molecule. [100, 144, 147] compared these sepsis markers prognosticating sepsis and sepsis-related complicity by assessing whether they were evaluated in experimental studies, clinical studies, and how they are as a prognostic factor. The authors also included brief critical comments, such as whether it is decreased in septic patients with DIC compared with no DIC, if it is distinguished between survivors and non-survivors at 28 days, if it predicted the development of MOF, whether it is correlated with SAPS score, and if it predicted the development of the acute lung injury, based on their evaluative studies.

After that, [150, 151, 152, 153, 154, 155, 156] discussed biomarkers pertaining to vasodilation, these discussions included Anandamide, Copeptin, Elastin, Cycling nucleotides, cGRP, Adrenomedullin, Proadrenomedullin, Angiotensin-converting enzyme, and C-type natriuretic peptide. [100, 156, 154] compared these sepsis markers by assessing whether they were evaluated in experimental studies, clinical studies, and how they are as a prognostic factor. The authors also included brief critical comments, such as whether it predicted the development of septic shock, if it correlated with APACHE-II score, whether
it is distinguished between survivors and non-survivors at twenty-eight days, and if it is decreased in sepsis compared with healthy controls, based on their critical study.

In [157, 158, 159, 160, 161], organ dysfunction biomarkers, such as Brain natriuretic peptide, Filterable cardiodepressant substance, Gc-globulin, Atrial natriuretic peptide, Carbomyl phosphate synthase, Troponin, MEGX test, Hepatocyte growth factor, Pancreatitis-associated protein-I, Glial fibrillary acidic protein, Surfactant protein, Myocardial angiotensin II, and Alpha glutathione S-transferase are discussed for prognosticating sepsis and sepsis-related complicacy. [100, 159] compared these sepsis markers by assessing whether they were evaluated in experimental studies, clinical studies, and how they are as a prognostic factor. The authors also included brief critical comments, such as whether it is higher in septic shock than in sepsis, whether it is strongly correlated with SOFA score, if it is increased in septic compared with non-septic patients, whether it is distinguished between survivors and non-survivors at Twenty-eight days, if it is increased in sepsis compared with healthy controls, and whether it is predictive of lethal outcome from postoperative sepsis, based on their research.

After that, [162, 163, 164, 165, 166, 167] considered acute phase protein biomarkers. It included Ceruloplasmin, Ferritin, Lipopolysaccharide binding protein, C-reactive protein, Serum amyloid A, Hepcidin, Pentraxin 3, and Procalcitonin. [100, 167, 162] compared these sepsis markers by assessing whether they were evaluated in experimental studies, clinical studies, and how they are as a prognostic factor. The authors also included brief critical comments, such as whether it is distinguished between survivors and non-survivors at Twenty-eight days, if it is correlated with CRP in patients with septic shock, if it predicted liver dysfunction in patients with sepsis, if it is higher
in sepsis compared with no sepsis, whether it is increased in infected compared with non-infected patients, and if it is increased in sepsis compared with healthy controls and patients with chronic renal failure, based on their research.

We found more than 170 different biomarkers for sepsis prognosis. A very few of them are acceptably used in sepsis diagnosis at present. Again, none of them showed enough specificity or sensitivity, so that we cannot employ them in routine clinical practice. Furthermore, in some cases, complicated combinations of multiple biomarkers showed more efficacy compared to that of a single biomarker. This necessarily makes the model more complicated and susceptible to different biases and requires further evaluation. Again, the third amendment in sepsis definition took place in 2016. This definition shift challenged most of these studies as far as the diagnosis is concerned.

3.1.4 Optimizing Therapeutics:

Sepsis causes injuries to the host’s tissues and organs, leaving critically ill patients at an increased risk of death. Hence, infection and appropriate anti-infective treatment are the cardinal purpose of sepsis treatment and management since successful infection control is crucial for resuscitation in the case of sepsis and septic shock. Therefore, early diagnosis followed by aggressive interventions within the initial hours of triage is essential. However, ambiguity in definition and diagnosis and a distinction from non-infectious causes frequently creates identification uncertainties and potential time delays. Again, the appropriate use of antibiotics remains a significant challenge. A wide variety of available therapeutic interventions and overly complicated factors associated with sepsis (such as host factors and pathogen factors) made the antimicrobial management for sepsis patients a practical challenge for hospitalization. Failure in antimicrobial management may lead to increased risk for opportunistic
infections, resistance to multiple antimicrobial agents, and toxic side effects, resulting in higher mortality and healthcare costs. Hence, optimizing therapeutics for sepsis is a major challenge in medical informatics, necessitating interdisciplinary coordination in efforts to overcome these challenges and ensure optimal care.

Literature manifested clinical experiments in an effort to reveal rules and flowcharts of anti-infection therapy for sepsis. These scientific inquiries covered timely selection and administration of appropriate antimicrobials, significant physiological alterations affecting antimicrobial pharmacokinetics, and interpatient variability of the antimicrobial concentrations using regular dosing methods. Understanding these factors and their possible impact on the probability of achieving pharmacokinetic-pharmacodynamic target goals is crucial to optimize the therapeutics. [1] revealed the physiological alterations and effects on antimicrobial pharmacokinetics as illustrated in Figure 3.1 and Figure 3.2.

Here, in these figures, the 1-acid glycoprotein is abbreviated as AAG, while ECMO, MV, and RRT stand for extracorporeal membrane oxygenation, mechanical ventilation, and renal replacement therapy. Vd means the volume of distribution.

In recent, [168] published a set of rules (commonly named as 6Rs rule) for sepsis patient management, concentrating on optimizing therapeutics. This study reflected a thorough literature review combined with comments from clinical experts. This set of rules include (R)ight patients, (R)ight time, (R)ight target, (R)ight antibiotics, (R)ight dose, and finally, (R)ight source control. The authors illustrated how this set of rules could help make rational decisions regarding the timing of treatment, identify the pathogen correctly, select antibiotics appropriately, formulating antibiotic dosage regimens based on scientifically
Figure 3.1: Visualizing physiological alterations and their effects on the antimicrobial pharmacokinetics [1].

- Altered fluid balance
  - Aggressive fluid resuscitation third spacing
  - ↓Vd
  - ↓Plasma concentrations of hydrophilic drugs

- Altered protein binding
  - ↓Albumin
  - ↑AAG
  - ↓Vd
  - ↓Free drug concentration
  - ↓Elimination of drugs highly bound to albumin
  - ↓Elimination of drugs highly bound to AAG

- Extracorporeal organ support
  - RRT, ECMO, MV
  - ↑Vd
  - ↓Clearance of hydrophilic drugs
  - ↑Sequestration of lipophilic drugs
  - ↓Plasma concentration

Figure 3.2: Visualizing physiological functional alterations and their effects on the antimicrobial pharmacokinetics [1].

- Altered hepatic function
  - Hypoperfusion to liver
  - ↓Hepatic metabolism and clearance
  - ↑Plasma concentration

- Altered renal function
  - Acute kidney injury
  - Augmented renal clearance
  - ↓Vd
  - ↓Clearance of renally eliminated drugs
  - ↑Plasma concentration
  - ↓Clearance of renally eliminated drugs
  - ↓Plasma concentration
proven rationals, and eventually achieve adequate control of infectious foci.

Then, [169] suggested that the first and the most critical step towards optimizing therapeutics is reliable diagnosis. When there is a lack of gold standard for infection diagnosis, different biomarkers—in different clinical studies—showed potential in early risk stratification. Besides, they indicated that different prognostic information might help in optimizing therapeutic decisions. The authors put stress on the blood infection marker procalcitonin (PCT) as it exhibited reliable discriminatory properties with rapidly available results to differentiate between viral and bacterial inflammations. Again, PCT serves in managing antibiotic use in the case of respiratory infection. It also helps by limiting initiation and shortening treatment duration. It is compelling as there is still no gold standard for sepsis detection and additional tests are required to early diagnosis. Again, [170] bolstered continuous renal replacement therapy as a major pathway to optimize therapy.

To address the sequential decision-making issues in sepsis treatment, in 2018, Matthieu Komorowski et al. published their research on Artificial Intelligence (AI) Clinician [171]. Here, they developed a reinforcement learning agent, extracting implicit knowledge from an amount of patient data using MIMIC-III electronic medical record that may exceed the life-time experience of human clinicians. The authors claimed that the AI clinician model learned optimal treatment by analyzing a number of suboptimal treatment decisions made by the human clinicians in the ICU. They further asserted that the value of the AI Clinician’s selected treatment is higher than human clinicians. They used the eRI database for their validation cohort, which has a substantial amount of patient data and independent of the training data of MIMIC-III. They showed that the number of cases, when human clinicians followed the treatment regime provided by the AI clinicians and experienced the lowest
mortality, is quite substantial. Literature shows that the wide variety of pathogen factors (compounded by the different strains of certain pathogens) and host factors cannot be addressed by a model developed on single-center data, challenging the generalizability of the regime suggested by the model in the administration of intravenous fluids and vasopressors.

Considering the breadth of the literature discussed here, we summarize the key works presented in this literature review in Figure 3.3 for the readers’ convenience.

Next, we will discuss the existing pitfalls and potential challenges in the
solutions developed using Electronic Medical Records.

3.2 Existing Pitfalls and Potential Challenges

There observed a recent surge in hospitals integrating Electronic Health Records (EHR) in their systems in an exponential manner since the HITECH Act was enacted in 2009. EHRs are primarily designed to advance healthcare providers’ administrative and operational efficacy, leveraging data mining power, and advanced analytics. Now, the secondary use of EHR data has been extended to data-driven scientific studies and medical informatics applications. Several studies reported that EHR or EMR (Electronic Medical Record) contributed to concept extraction, disease inference, biomarker discovery, risk stratification, early prediction and optimizing resource allocation, cohort stratification for recruiting in the clinical trial, noninvasive prediction, patient trajectory modeling, and hospital decision support system, resulting in a recent landscape shift healthcare research. Recently, compared to traditional machine learning and statistical techniques, deep learning techniques—through capturing long-range dependencies in data—have achieved remarkable success in different avenues of medical informatics research and further advocate the utility of the EHRs in scientific studies. EHRs preserve a wide range of patient data and granular measurements, including demographics and prior issues, diagnoses, physical exams, physiological information and sensor measurements, laboratory test results, prescriptions, interventions, and clinical notes over the course of patient’s stay. The heterogenic nature of the EHR data can primarily be classified into five groups: numerical quantities, DateTime objects, categorical values, natural language free text, and derived time-series data.

There are several benefits of using observational EMR data over data collected from controlled studies to train and evaluate machine learning and deep
learning models. First, the abundance of data required for deep learning models is available in the EMR data. Second, collecting data in the case of EMR is of virtually zero cost. Third, data is collected when treating patients and then deidentified when releasing it for the research purposes; hence EMR data is less susceptible to privacy issues compared to controlled studies. Forth, EMR data is collected near real-time while providing care; hence it has enormous potential to be useful in real-time prediction and patient management.

In an Intensive Care Unit (ICU) setting, recognizing potential complications early may create opportunities to intervene in the patients early and more meticulously, resulting in a cost-effective allocation of resources and better patient outcomes. It is evinced from the literature that Trauma patients, who were diagnosed early and intervened rapidly in treating shock and organ dysfunction, required lesser hospital resource consumption and resulted in better outcomes. On the other hand, several studies regarding ICU treatment indicated that delays in identification, intervention, and management is significantly correlated with higher mortality rates and increased hospital resource consumption. In particular, as most of the sepsis cases are confronted and dealt with in the ICU, these scientific findings advocate in favor of stand-alone monitoring or diagnostic system and assisted monitoring supported by the machine learning models developed using EMRs. The practitioners and caregivers use stand-alone monitoring systems to decide whether it requires to intervene or not. On the other hand, assisted monitoring systems facilitate monitoring-in-parallel with the practitioners to detect patients at risk and function as a second eye in monitoring so that any prospective critical situation can not be missed. However, evidence shows that early prediction models (stand-alone monitoring systems) of septic shock using Electronic Health Records (EHR) are less useful in practice as the EHR data often comprise
systematic biases that may violate the conceptual assumptions made by state-of-the-art machine learning algorithms. Moreover, confounding medical interventions during the Intensive Care Unit (ICU) stay may strongly affect the predictive outcomes. In recapitulation, learning and evaluating stand-alone systems, in particular, early prediction systems, using EMR data alone confront three following significant challenges: Incomplete Observations, Selection Bias, Confounding Medical Interventions [172].

3.2.1 Incomplete Observation

Sometimes, there may have missing values in the middle of the patient’s trajectory. Treating it using statistical imputation is not ethically correct to some extend. Again, excluding the observation may create a hole in understanding the patient’s trajectory. Then, EMR data, albeit rare, experience undeliberate error in data entry, which is not physically reasonable. Sometimes, data can be right or left censored as some patients can be discharged too early, and some can be transferred too late. Especially in the case of training predictive models, there may be a requirement to have data for at least certain hours. Samples need to exclude if it does not meet this criterion.

3.2.2 Selection Bias

EMR Data may not represent a random sample from the population, and this gap primarily varies with the geographical location of the medical center and nature of the particular practice. This inevitable gap may question the trained model’s generalizability and impose restrictions where the predictive model can be deployed. For example, the adaptive immune responses of patients, which is modulated by vitamin D intake, may vary from the case of Wisconsin or Minnesota to California or Florida, considering their exposure to the sun
throughout the year. In this case, the bias induced in the host factor may hamper the generalizability, and eventually the utility, of the model. Again, most of the off-the-shelves machine learning models are based on some statistical assumptions. The bias induced by the data selection process may result in an invalid assumption in practice.

### 3.2.3 CMIs

Confounding Medical Interventions are the caregivers’ interventions in practice, which may influence the risk of the outcome concerned. For example, in the case of sepsis intervention and management, broad-spectrum antibiotics, and administration of pressors relevant interventions. From the aspect of training predictive models, Confounding Medical Interventions sometimes mask the true labels of the patients’ trajectory. After these interventions, it is very difficult to distinguish between a subject who was treated unnecessarily due to the conservative treatment management policy and judgment by the caregiver, and a subject who was actually at risk but then treated due to effective antibiotic treatment. Understanding and handling the data containing CMIs require special attention before feeding them into the machine learning models.

The following table considers four categories of cohort generally present in the observational data. Among them, categories A and B contain no confounding medical interventions. For these cohorts, the absence and presence of an adverse condition inside a defined time frame can truly reveal whether a patient was at risk during his stay in the hospital. Data associated with categories A and B can be considered as clear data samples. For category D, subjects with CMIs intervened, the absence and presence of an adverse condition (outcome such as septic shock) inside a defined time frame cannot appropriately unpack whether the patient was truly at risk. Data collected from this cohort can be considered
as confounded data samples. In the case of the category C cohort, even though CMIs were intervened, adverse conditions (such as septic shock) were still observed as an unintended consequence of the interventions itself. Category C may also be counted as the confounded data sample. For sepsis, the appearance of an adverse condition indicates the patient was truly at risk, albeit it is due to CMIs were applied too late or insufficient. Considering the context of the study, we still recognize these data as clear samples. Stratifying the patients into different cohorts also depends on caregivers’ treatment judgment and may vary patient to patient and practice to practice [172].

Apart from addressing these challenges, the literature discussed three aspects of the learning application: learning algorithms and their comparative efficacy in practice; choice of features, their correlations (multicollinearity and dichotomy) and causal relations; guidelines on developing and evaluating predictive models under the uncertainties of CMIs. In particular, to address the possible class mismatch problem, researchers suggest using a transductive approach to infer true labels rather than ignoring adhoc labeling for the confounded data.

Studies show that training and evaluating a stand-alone system using EMR is more susceptible to the pitfalls discussed here, reflected by its less usability in practice. On the other hand, assisted monitoring systems developed using EMRs are less susceptible to biases and challenges due to the nature of the problem dynamics. Therefore, literature touts more about the assisted monitoring-based solutions for sepsis monitoring and interventions in the ICU. However, there found a contradictory perception in the literature as well. It implies that practitioners, in practice, make their decisions independently from the system. Interventions mostly follow the caregivers’ suspicion that the patient is in a risky condition. Early prediction systems can add an extra layer of the
supportive tool in their decision-making process. Hence, the findings end in a note that the early prediction system confounded by systematic biases may not be very detrimental in practice. However, as far as the system’s utility is concerned, both of the perception held in the machine learning and medical informatics scientific community advocate EMR-based assistive monitoring system as a decision support system for the caregivers in the ICU.

A multi-blind Delphi process, convened by Ubicomp Lab of the Department of Computer Science at Marquette University, Regenstrief Center for Healthcare Engineering at Purdue University, College of Medicine at University of Central Florida, and RB Annis School of Engineering at University of Indianapolis, addressing the uncertainty around predictions on septic shock onset and multiple organs dysfunction syndrome in the ICU monitoring, came into a decision that detecting the regime-switching in univariate and multivariate physiological signals can help the clinicians understand the transition dynamics and the behavior of sequential events. Biological systems are, by nature, susceptible to spontaneous as well as recurrent transitions, which can be gradual or abrupt between healthy and pathological states. Bedside monitoring of physiological parameters allows the macroscopic collection of episodic events to capture the pathophysiology and delineates it as a time-series [173]. By identifying the abrupt variation between physiological signals’ structure of individual patients, we can assist practitioners in focusing more on the critical transitions. These transitions may imply either detecting changes in conditions resulting from treatments or suggesting to treat adverse events preemptively. Accurate detection of these vital transitions can prompt early diagnoses and interventions, leading to improved clinical outcomes [174, 173, 3].
Chapter 4

Change Point Detection: Comparative Analysis on Methods and Utility

In time-series data analytics, we often observe a moment(s) when an abrupt change(s) happens in the behavior of the time-series data stream. It may infer a significant switch in the data generating process, from the pragmatic aspect, in the overall dynamics of a system, thereby seeking more attention and further research for monitoring, controlling, and interventions. Hence, detecting changepoints in the time-series data stream is very significant in data analytics and machine learning. The application of change point detection includes finance, business, climate change detection, medical condition monitoring, human activity monitoring, network traffic analysis, image analysis, speech recognition, quality control, and so on. In particular, to the interest of our problem, change point detection has potential, in a few cases, have already been implemented to facilitate medical condition monitoring and human activity analysis. Continuous and automated monitoring of the patient’s health involves trend detection (or detecting regime-switching) in the patient’s physiological parameters, such as heart rate, respiratory rate, and EEG. Change point detection methods have already been implemented for specific medical issues, such as epilepsy, sleep problems, and MRI interpretation, to facilitate caregivers. Besides, it can also help in health activity analysis through mobile devices, and contribute to offering activity aware services by detecting behavioral changes that provide useful insights on individuals’ health status.

Figure 4.1 illustrates examples of change points (breakpoints) and states (regimes) in a sample time series problem containing several change points. This
time series represents the long term (1899–2010) mean annual temperature trends of Spitsbergen, an island in northern Norway [2]. This graph, highlighting the climate of Spitsbergen, can be used for formulating a climate change detection problem as the plot depicts it went through six distinct regimes in these 110 years. We surmise that these states represent parameters governing the process that do not change in this time period; a change point distinguishes two consecutive distinct states. The aim of detecting change points is to recognize these partitions by discovering the switches in generative parameters in the data stream.

4.1 Defining and Formulating the Problem

Here, we define the key terms and formulate the mathematical nature of the problem we will use throughout our study.

Definition 1: A time-series data stream can be defined as an infinite sequence of elements $S$, where $S$ is represented as \{ $x_1, \ldots, x_i, \ldots$ \} and $x_i$ as a d-dimensional vector appearing at time stamp $i$ [175]. A finite variance process
whose statistical properties are all constant over time can be defined as a stationary time series, assuming the mean value function is time-independent and \( \mu_t = E(x_t) \) is constant, and the autocovariance function \( \gamma(s, t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)] \) is dependant on time stamps \( s \) and \( t \) through \( |s - t| \), their time difference.

**Definition 2**: An independent and identically distributed-time series is a specific case of the stationary time series. Here, the variables are identically distributed suchly that it is perceived they are drawn from the same probability distribution. Besides, the variables are mutually independent of each other.

**Definition 3**: A point representing the transition between different states (regime) in a process (system/parameters) that generates a time series data-stream can be defined as a change point \([176, 177]\).

**Definition 4**: Considering the stream \( \{x_m, x_{m+1}, \ldots, x_n\} \) as a sequence of time series variables, change point detection can be regarded as the hypothesis testing problem between two alternatives as follows: the null hypothesis \( H_0 \): “No change occurs” and the alternative hypothesis \( H_A \): “A change occurs” \([178, 179]\).

- \( H_0 \): \( P_{X_m} = \ldots = P_{X_k} = \ldots = P_{X_n} \)
- \( H_A \): There exists \( m < k^* < n \) such that \( P_{X_m} = \ldots = P_{X_{k^*}} \neq P_{X_{k^*+1}} = \ldots = P_{X_n} \); given that \( P_{X_i} \) and \( k^* \) represent the probability density function of the sliding window start at point \( x_i \) and a change point, respectively.

Though literature shows a large body of work exists on change-point detection, particularly its theoretical and applied aspects, minimal work has been found on the evaluation of change-point detection algorithms. In most
cases, when a new change point detection algorithm is introduced, the evaluation follows simple and straightforward ways: either deploying on a set of simulated data streams or a very few real-world data, or in some cases, both. Simulated data seldom represents the complexity, uncertainty, and existence of confounding factors that influence the real-world time series data describing a physical system. Moreover, authors may take an unfair advantage as the simulated data has known change points. Sometimes, the generative parameters of different partitions are set so distinctly that they advocate both the model fit and detection accuracy during evaluation. It is also noticed that a tiny number of real-world data sets are used (reused) in evaluating change point detection for a long time. These data sets may often suffer from the absence of unambiguous ground truth, reflected by the post hoc analysis and arguments presented in the literature. The lack of evaluation using a considerable amount of real-world data across multiple domains indicates a gap in the body of knowledge of change point detection. This gap may result in a less effective algorithm when deploying real-world exposure.

An extensive set of high-quality realistic benchmark data sets across multiple domains can bridge this gap for change point detection algorithms, their applications, and efficacy. These data sets will let the data scientists evaluate change point detection algorithms systematically and realistically. As change point detection, by nature, is an unsupervised learning problem, using real-world data sets is a more reliable approach to evaluate and compare detection accuracy and efficiency, which may not be possible if evaluated using simulated data. Again, change point detection can be formulated as a segmentation problem as it seeks partition(s) in the data stream. Hence, its evaluation process can be compared with image segmentation problems as they partition images into distinct segments in terms of different textures or objects. At the beginning of
image segmentation theories and applications, it also suffered from a similar problem that time series change point detection is facing now: the lack of an extensive and high-quality data set having objective ground truth. That gap was addressed and bridged by BSDS (Berkley Segmentation Data Set) by allowing a realistic quantitative comparison of segmentation algorithms [5, 180].

In like manner, TCPDBench Data Set by Alan Turing Institute has the potential to contribute similarly in the case of change point detection algorithms by facilitating realistic quantitative evaluation and comparison. TCPDBench Data Set by Alan Turing Institute includes 37 real-world time series data specifically designed, covering a broad spectrum of real-world scenarios and having annotations collected from data scientists and domain experts. It is freely accessible for researchers, academicians, and data scientists to facilitate change point detection and time series analysis research. Using TCPDBench Data Set by Alan Turing Institute, we will study on evaluating present change point detection algorithms, unpacking failure cases of existing methods, discussing evaluation metrics that take multiple annotations per data set into the account, and addressing various potential improvements from the application point of view [5].

A total of 37 real time-series data include 33 univariate and 4 multivariate data sets (See Appendices). This effort gathered data from various online sources and few readily available data sets frequently used for change point detection. Online sources, such as EuroStat, World Bank, GapMinder, US Census Bureau, and Wikipedia, were helpful in finding data with potential change point attributes. Most of the data sets capture different aspects of the financial crisis of 2007-2008. Some of them incorporate the GDP series of different countries reflecting different policies and legislation introduced in that time and their effects. TCPDBench Data Set also includes the existing data sets used for evaluating change point detection algorithms, such as well-log data,
bee-waggle data, and Nile data. The cardinal criterion behind incorporating these data sets is that the data needs to manifest interesting behavior that may include an abrupt change in generative parameters. This study considers a large number of frequently used and recently-developed change point detection algorithms [181, 5].

For TCPDBench Data Set, the annotation collection was performed by the machine learning researchers and data scientists, perceiving that there is a distinct difference between an outlier and a change point, and there may also have multiple change points or no change point in the time series data stream. To avoid any possible bias induced from the annotators’ knowledge of historical events to interpret the data, dates were intentionally removed on the time axis. Besides, values are also removed on the vertical axis. Furthermore, names of the series—while requesting to annotate—are also not given to the annotators to avoid bias induced by domain knowledge. The annotation rubric was inspired by the BSDS for image segmentation considering the similarity in the nature of the problem. The detail regarding the annotation tool is presented in [181, 5, 182].

4.2 Evaluation Metrics

Before comparing the frequently used or recently proposed change point detection algorithms against multiple ground truth annotations, we will discuss the standard evaluation metrics available in the literature and their perspectives on the consistency of annotations. This discussion has a fundamental similarity with Martin et al. and Arbelaez et al. for the BSDS, and TCPDBench for change point detection, considering the underlying similarities in the nature of the image segmentation and change point detection problem. Evaluation metrics of change point detection algorithms can broadly be categorized into clustering and classification metrics. Both of these two kinds aim to capture different
aspects of the detection problem. Here, for our comparative study, we will discuss both aspects.

4.2.1 Evaluating the Change Point Detection Problem in the lens of Clustering:

The locations where the change points occur are denoted by an ordered set $T_k = \{\tau_1, ..., \tau_{nk}\}$ where $k \in \{1, ..., K\}$ represents annotator and $\tau_i \in [1, T]$ for $i = 1, ..., n_k$, where $\tau_i < \tau_j$ for $i < j$. Here, $T_k$ implies a partition $G_k$ in an interval of $[1, T]$ into disjoint sets $A_j$, representing the segment from $\tau_{j-1}$ to $\tau_j - 1$ for $j = 1, ..., n_k + 1$.

Evaluating change point detection algorithms in the lens of clustering metrics aims to assess the switching points dividing time series into distinct regions with a constant data-generating process. Classically, the Adjusted Rand Index (ARI), Variation of information (VI), and Hausdorff distance are used to evaluate clustering problems. However, literature discusses the intricacy involved when VI is used in the case of addressing multiple ground truth partitions. Again, the small dynamic range of ARI may not be appropriate to evaluate change point detection. Then, Hausdorff distance is necessarily a maximum discrepancy metric; therefore, it may obscure the true performance of methods. This may result in reporting many false negatives [183, 184]. For these reasons, we disregard these three classical evaluation metrics in this study and decide to choose the covering metric as the clustering metric [183, 184, 5].

For sets $A, A' \subseteq [1, T]$, the Jaccard index is as follows. Jaccard index represents the intersection over the union.

$$J(A, A') = \frac{|A \cap A'|}{|A \cup A'|} \quad (4.1)$$

The covering metric of partition $G$ by partition $G'$ is defined as follows.
\[ C(G', G) = \frac{1}{T} \sum_{A \in G} |A|.max_{A' \in G'} J(A, A') \] (4.2)

Here, the average of \( C(S, G_k) \) for all annotators represents a single measure of performance, given that \( \{G_k\}_{k=1}^{K} \) is the collection of ground truth partitions by the annotators (data scientists) and \( S \) is the partition from the algorithm.

4.2.2 Evaluating the Change Point Detection Problem in the lens of Classification:

Another lens to evaluate change point detection algorithms is to consider it a classification problem with "non-change point" and "changepoint" classes. For the change point detection problem, as the number of "changepoint" classes is negligible compared to the "non-change point" classes, the standard classification evaluation metrics' score may be highly skewed, particularly accuracy. Literature shows– among the classification metrics– precision and Recall are promising in the case of evaluating change point detection algorithms. Precision represents the ratio of the number of change points detected accurately over the algorithm's overall number of change points detected. On the other hand, Recall is the ratio of the number of change points identified correctly over the number of true change points in the data stream. To incorporate both Recall (R) and precision (P) in terms of a single measure, we use \( F_\beta \) in this study. \( F_\beta \) is expressed as follows, given that \( \beta \) for the standard F1-score.

\[ F_\beta = \frac{(1 + \beta^2)PR}{\beta^2P + R} \] (4.3)

Another critical issue is when we evaluate change point detection algorithms from the classification aspect, it is required to define an error margin
around the true regime-switching point. Though it allows for the minor discrepancy, special attention is necessary to avoid double counting within the error margin. Multiple detections within an error margin need to be considered as one true positive only. In this effort, addressing similarity in the nature of the problem, we adopt the Precision-Recall framework used in BSDS image segmentation.

Here, \( \chi \) means set of regime-switching locations obtained from the detection algorithm, \( T^* = U_k T_k \) indicates a combined set representing all annotations by the data scientists. \( TP(T, \chi) \) is the set of true positives of \( \chi \). The Precision-Recall is expressed as follows.

\[
P = \frac{|TP(T^*, \chi)|}{|\chi|} \tag{4.4}
\]

\[
R = \frac{1}{K} \sum_{k=1}^{K} \frac{|TP(T_k, \chi)|}{|T_k|} \tag{4.5}
\]

According to Precision’s definition, we recognize only those points detected by the algorithm but not annotated by the data scientists as false positive. In an effort to explain all human annotations, Recall addresses the combined set of annotations and avoids any possible favor to individuals’ annotation. Thus, by adopting both segmentation covering metric and F1-score, this study captures both aspects (clustering and classification) of the change point detection problem [5].

4.3 Overviewing the Experimental Design

The cardinal objective of this comparative study is to understand how the state-of-the-arts changepoint detection algorithms perform when deployed for a number of real-world data. This study will help us figure out which algorithm we
Table 4.1: State-of-the-arts change point detection algorithms for this comparative study.

<table>
<thead>
<tr>
<th>Method</th>
<th>Algorithm</th>
<th>Author</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>At Most One Change</td>
<td>AMOC</td>
<td>Hinkley et al.</td>
<td>1970</td>
</tr>
<tr>
<td>Binary Segmentation</td>
<td>BINSEG</td>
<td>Scott et al.</td>
<td>1974</td>
</tr>
<tr>
<td>Segment Neighborhoods</td>
<td>SEGNEIGH</td>
<td>Auger et al.</td>
<td>1989</td>
</tr>
<tr>
<td>Bayesian Online Change Point Detection</td>
<td>BOCPD</td>
<td>Adams et al.</td>
<td>2007</td>
</tr>
<tr>
<td>Kernel Change Point Analysis</td>
<td>KCPA</td>
<td>Harchaoui et al.</td>
<td>2009</td>
</tr>
<tr>
<td>Pruned Exact Linear Time</td>
<td>PELT</td>
<td>Killick et al.</td>
<td>2012</td>
</tr>
<tr>
<td>Wild Binary Segmentation</td>
<td>WBS</td>
<td>Fryzlewicz</td>
<td>2014</td>
</tr>
<tr>
<td>Energy Change Point</td>
<td>ECP</td>
<td>Matteson</td>
<td>2014</td>
</tr>
<tr>
<td>Nonparametric Change Point Detection</td>
<td>CPNP</td>
<td>Haynes et al.</td>
<td>2017</td>
</tr>
<tr>
<td>BOCPD with Model Selection</td>
<td>BOCPDMS</td>
<td>Knoblauch et al.</td>
<td>2018</td>
</tr>
<tr>
<td>Prophet</td>
<td>PROPHET</td>
<td>Taylor et al.</td>
<td>2018</td>
</tr>
<tr>
<td>Robust BOCPDMS</td>
<td>RBOCPDMS</td>
<td>Knoblauch et al.</td>
<td>2018</td>
</tr>
<tr>
<td>Robust Functional Pruning Optimal Partitioning</td>
<td>RFPOP</td>
<td>Fearnhead et al.</td>
<td>2019</td>
</tr>
</tbody>
</table>

should deploy as the base change point detection algorithm in our assistive tool, considering the higher restraints of accuracy for the medical informatics deployment in the ICU. It motivates us to include a large number of time-series data, combining many real-life and simulated data, and state-of-the-arts changepoint detection algorithms for our consideration. This large number of time-series data ensures that no particular algorithm may get any unfair favor in the evaluation. Table 4.1 lists the change point detection methods considered for this study. They are presented in ascending order of the date. These unsupervised methods are either frequently used across multiple domains of scientific studies or have been recently developed. For this comparative study, we used off-the-shelves software packages for all methods to avoid any possible implementation issues.

These methods have different parameters and hyperparameters, and they may influence the predicted change point locations. In this study, we intend to evaluate how accurate these methods are in practice. To facilitate a realistic understanding of change point detection algorithms’ performance, two separate evaluations are presented. First, we evaluate the performances while using the
default settings for each of the methods. These default settings are as specified in the documentation of the packages developed for deploying these algorithms. Second, aiming to identify the highest possible performance of each of the algorithms, we evaluate the performances based on the maximum score reporting over a grid search of parameter configuration. To get the best possible performances, we run a full grid search over its hyperparameters. On some of the time-series data, issues with numerical precision are crucial for some of the methods. To address these possible issues, all of the time-series are standardized to zero mean and unit variance.

4.4 Results

In Table 4.2, we listed the default and maximum score on covering metrics for univariate series. Then, Table 4.3 illustrates the default and maximum score on the F1 metric for univariate series. Here, both the table listed the methods in the ascending order of the date. For both tables, the value corresponding to the default experiment represented the average of the single run of the corresponding algorithm. In like manner, the maximum score experiment value in both tables represented the average of the maximum score achieved for all series for the corresponding algorithm. Table 4.4 and Table 4.5 illustrated the default and maximum score on the covering and F1 metrics for multivariate series, respectively.

It is evinced from Table 4.2 and Table 4.3 that Binary Segmentation shows the best performance both in covering and F1 score, as far as the default experiment is concerned. For the default experiment, covering and F1 scores obtained for At Most One Change and Pruned Exact Linear Time algorithm also showed promising results. Table 4.4 and Table 4.4 showed that Bayesian Online Change Point Detection with Model Selection performed the best for multivariate
Table 4.2: Default and maximum score on covering metric for univariate series.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cover (Default)</th>
<th>Cover (Maximum Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>At Most One Change</td>
<td>0.702</td>
<td>0.746</td>
</tr>
<tr>
<td>Binary Segmentation</td>
<td>0.706</td>
<td>0.780</td>
</tr>
<tr>
<td>Segment Neighborhoods</td>
<td>0.676</td>
<td>0.784</td>
</tr>
<tr>
<td>Bayesian Online Change Point Detection</td>
<td>0.636</td>
<td>0.789</td>
</tr>
<tr>
<td>Kernel Change Point Analysis</td>
<td>0.062</td>
<td>0.626</td>
</tr>
<tr>
<td>Pruned Exact Linear Time</td>
<td>0.689</td>
<td>0.725</td>
</tr>
<tr>
<td>Wild Binary Segmentation</td>
<td>0.330</td>
<td>0.428</td>
</tr>
<tr>
<td>Energy Change Point</td>
<td>0.523</td>
<td>0.720</td>
</tr>
<tr>
<td>Nonparametric Change Point Detection</td>
<td>0.535</td>
<td>0.552</td>
</tr>
<tr>
<td>BOCPD with Model Selection</td>
<td>0.633</td>
<td>0.744</td>
</tr>
<tr>
<td>Prophet</td>
<td>0.540</td>
<td>0.576</td>
</tr>
<tr>
<td>Robust BOCPDMS</td>
<td>0.629</td>
<td>N/A</td>
</tr>
<tr>
<td>Robust Functional Pruning Optimal Partitioning</td>
<td>0.392</td>
<td>0.414</td>
</tr>
</tbody>
</table>

Table 4.3: Default and maximum score on F1 metric for univariate series.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1 (Default)</th>
<th>F1 (Maximum Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>At Most One Change</td>
<td>0.664</td>
<td>0.783</td>
</tr>
<tr>
<td>Binary Segmentation</td>
<td>0.680</td>
<td>0.803</td>
</tr>
<tr>
<td>Segment Neighborhoods</td>
<td>0.626</td>
<td>0.806</td>
</tr>
<tr>
<td>Bayesian Online Change Point Detection</td>
<td>0.611</td>
<td>0.831</td>
</tr>
<tr>
<td>Kernel Change Point Analysis</td>
<td>0.057</td>
<td>0.589</td>
</tr>
<tr>
<td>Pruned Exact Linear Time</td>
<td>0.653</td>
<td>0.703</td>
</tr>
<tr>
<td>Wild Binary Segmentation</td>
<td>0.300</td>
<td>0.390</td>
</tr>
<tr>
<td>Energy Change Point</td>
<td>0.551</td>
<td>0.706</td>
</tr>
<tr>
<td>Nonparametric Change Point Detection</td>
<td>0.534</td>
<td>0.567</td>
</tr>
<tr>
<td>BOCPD with Model Selection</td>
<td>0.476</td>
<td>0.569</td>
</tr>
<tr>
<td>Prophet</td>
<td>0.456</td>
<td>0.506</td>
</tr>
<tr>
<td>Robust BOCPDMS</td>
<td>0.418</td>
<td>N/A</td>
</tr>
<tr>
<td>Robust Functional Pruning Optimal Partitioning</td>
<td>0.372</td>
<td>0.392</td>
</tr>
</tbody>
</table>

Table 4.4: Default and maximum score on covering metric for multivariate series.

<table>
<thead>
<tr>
<th>Method</th>
<th>Cover (Default)</th>
<th>Cover (Maximum Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Online Change Point Detection</td>
<td>0.188</td>
<td>0.718</td>
</tr>
<tr>
<td>Kernel Change Point Analysis</td>
<td>0.009</td>
<td>0.456</td>
</tr>
<tr>
<td>Energy Change Point</td>
<td>0.238</td>
<td>0.388</td>
</tr>
<tr>
<td>BOCPD with Model Selection</td>
<td>0.286</td>
<td>0.386</td>
</tr>
<tr>
<td>Robust BOCPDMS</td>
<td>0.402</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 4.5: Default and maximum score on F1 metric for Multivariate series.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1 (Default)</th>
<th>F1 (Maximum Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Online Change Point Detection</td>
<td>0.455</td>
<td>0.801</td>
</tr>
<tr>
<td>Kernel Change Point Analysis</td>
<td>0.047</td>
<td>0.626</td>
</tr>
<tr>
<td>Energy Change Point</td>
<td>0.402</td>
<td>0.590</td>
</tr>
<tr>
<td>BOCPD with Model Selection</td>
<td>0.486</td>
<td>0.623</td>
</tr>
<tr>
<td>Robust BOCPDMS</td>
<td>0.250</td>
<td>N/A</td>
</tr>
</tbody>
</table>
series for the default settings. However, tuning the hyperparameters, we found that Bayesian Online Change Point Detection Algorithm outperformed all other algorithms listed here convincingly for both covering and F1 metrics, as evinced from Table 4.2 and Table 4.3. Furthermore, Table 4.4 and Table 4.5 showed that the Bayesian Online Change Point Detection algorithm performed the best for multivariate series, considering the maximum scores.

Theoretically, covering metric and F1 metric addressed two distinct aspects of the problem. A high score in covering metrics implies a high degree of agreement with each annotator. On the other hand, a high value on the F1 score means a strong agreement with a combined set of annotations. It is an exciting finding that the best performing algorithms maintained the same order (or ranking) regardless of the evaluation metric (covering/F1) we used, reflecting a higher degree of stability in these methods’ relative performance. All of these four tables also inferred how the impact of hyperparameter tuning varied with the algorithms. Theoretically, it mostly depends on the number of hyperparameters involved with the particular algorithm. It is also interesting—and worth mentioning as well—that, assuming a simple gaussian model with the constant mean for each of the segments, bayesian online change point detection can outperform most of the recent and mathematically more complicated models in addressing complex time series behaviors, such as seasonality and autoregressive structures [5].

Two rationals helped us to decide bayesian online change point detection for our base algorithm from this comparative study. First, it outperforms all the algorithms as far as the hyperparameter tuning is concerned. Second, bayesian online change point detection gives us the provision to take the domain knowledge into consideration.

In the next chapter, we will discuss how Bayesian Online Change Point
Detection algorithm works, the assumptions and mathematics behind it, and how it can be affected by the context while detecting the regime-switching in univariate and multivariate physiological signals to help the clinicians understand the transition dynamics and the behavior of sequential events, and how that may induce noise in detection. We will also discuss how we can address this challenge and come up with a contextually-tailored Bayesian Online Change Point Detection algorithm for ICU monitoring of sepsis patients.
Chapter 5

Contextually-tailored Online Change Point Detection for Sepsis Vital Sign Monitoring

Here, in this section, first, we explain the change point detection algorithm inclusion criteria for sepsis vital sign monitoring. Then, we describe the mathematical detail of the Bayesian Online Change Point Detection Algorithm. After that, we discuss how we can address its potential susceptibility to the noise confronted because of the nature of the problem. Understanding the mathematical explanation and how integrating context-awareness into this algorithm helps in noise reduction may require some knowledge and experience in Bayesian inference for the Gaussian, conjugacy, and the exponential family. Here, we consider a datum in the stream as an observation from a generative process.

5.1 Algorithm Inclusion Criteria

Our study presents algorithm inclusion criteria from two aspects: Algorithmic Perspective and Medical Informatics Perspective.

5.1.1 Algorithmic Perspective

Three rationals helped us decide Bayesian Online Change Point Detection as the primary algorithm to monitor structural changes in the sepsis vital signs data regime:

- The algorithm outperforms all the algorithms as far as hyperparameter tuning is concerned.
- The algorithm gives us the provision to take the domain knowledge into consideration.
• Bayesian paradigm offers a more personalized or particular problem-centric approach while considering the probability of events.

5.1.2 Medical Informatics Perspective

To select the change point detection algorithm for sepsis vital signs monitoring, we have the following algorithm inclusion criteria.

• The algorithm has to detect the changepoint in the physiological parameters online; its accuracy needs to be competitive with the available offline change point detection algorithms.

• The algorithm should have the provision to take in the domain knowledge information as input so that considering this domain information can reflect in leveraging with the change point detected.

• It has to be robust to noise so that it experiences occasional false positives and outliers.

• It has to be interpretable and explainable so that understanding the inner functionalities of the algorithm can help in extending the algorithm as per the requirements of the problem or context. For instance, an intuitive understanding of the algorithm allows tuning hyperparameters efficiently.

• The change point detection algorithm has to detect any kind of change in time-series data, not specific to mean only or variance only.

Our comprehensive literature study finds that only Bayesian Online Change Point Detection algorithm meets all the criteria [3, 185, 5]. Unlike the frequentist approach, Bayesian inference is a data-driven approach to estimate how the target variable behaves as the data drives the prior and posterior distributions. This updating is crucial, particularly for the dynamic analysis of a data stream [3].
5.2 Bayesian Online Change Point Detection

5.2.1 Model

Let, the \( t \)-th observation in a data stream is denoted as \( x_t \in \mathbb{R}^d \) and \( X_{s:t} \) represents the sequence \( x_s, x_{s+1}, ..., x_{t-1}, x_t \) where \( s \leq t \). Further, we are assuming that \( T \) data points, which can be denoted as \( X_{1:T} \), can be partitioned in such a way that observations within each partition represents independent and identically distributed samples from some distribution, demonstrating product partition model [186]. In a mathematical fashion, \( \theta^{(p)} \) represents partition \( P \)'s generative parameters and \( \theta_0 \) represents hyperprior. Therefore, \( \theta^{(p)} \sim p(\theta_0) \) are independent and identically distributed for \( p = 1, 2, 3, ..., P \) partitions. Figure 5.1 illustrates the conceptual sketch representing change point detection and run-length of Bayesian Online Change Point Detection [4, 3].

Bayesian online change point detection functions by modeling the run length, representing the time since the last change point. \( r_t \) denotes the run length at a given time \( t \). According to the Bayesian online change point detection algorithm, \( r_t \) can only take either of the two values.

\[
 r_t = \begin{cases} 
 0 & \text{if change point occurs at time } t \\ 
 r_{t-1} + 1 & \text{else} 
\end{cases} \quad (5.1)
\]

This implies run length can only be dropped down to zero or increased by one. Then, assuming that \( r_t \) is conditionally independent of every other thing but \( r_{t-1} \), the transition probabilities can be modeled by change point prior \( p(r_t | r_{t-1}) \). In order to predict the next data point, given all the data we have seen so far, we need to infer two things. First, the run length posterior distribution \( p(r_t | X_{1:t}) \). And then, the posterior predictive distribution \( p(x_{t+1} | X_{1:t}) \) [187, 3].
5.2.2 Recursive run length posterior estimation

If $D$ symbolizes a set of $N$ independent and identically distributed observations $\{X_n\}_{n=1}^{N}$ and $\theta$ are model (generative) parameters, Bayesian statistics infers that the posterior predictive distribution over a new observation $\hat{X}$ can be denoted as,

$$p(\hat{X} | D) = \int p(\hat{X} | \theta)p(\theta | D)d\theta$$ \hspace{1cm} (5.2)$$

Here, we estimate the prediction of the probabilistic model by the uncertainty of the parameters of the model. Marginal probability indicates evaluating this for all possible parameter values. This posterior gets a different
form, though slightly, in the case of Bayesian online change point detection. In this case, the posterior predictive is evaluated by the probability of the next observational point $x_{t+1}$, given the data seen so far $X_{1:t}$. This is calculated by marginalizing $r_t$ (run length) out.

$$p(x_{t+1} | X_{1:t}) = \sum_{r_t} p(x_{t+1}, r_t | X_{1:t})$$  \hspace{1cm} (5.3)

Here, $p(x_{t+1} | r_t, X(l))$ and $p(r_t | X_{1:t})$ denote underlying probabilistic model predictive and run length posterior respectively. $X(l)$ denotes all observations associated with $r_t = l$. For a high level of understanding, $l$ can be compared with $r_t$, and in that way, we can hypothesize a change point occurred $l$ time steps ago. Hence, the predictive distribution over $X_{t+1}$ is only influenced by the observations within the past $l$ time. This lets us to rewrite $p(x_{t+1} | X_{1:t})$ in a more explicit way as follows.

$$p(x_{t+1} | X_{1:t}) = \sum_{l=0}^{t} p(x_{t+1} | r_t = l, X(t-l); t)p(r_t = l | X_{1:t})$$  \hspace{1cm} (5.4)

Now, the idea is if we can estimate the underlying probabilistic model predictive, we may only require to calculate the run length posterior. The run length posterior is proportional to the joint distribution and can be represented as follows.

$$p(r_t | X_{1:t}) = \frac{p(r_t, X_{1:t})}{\sum_{r_{t'}} p(r_{t'}, X_{1:t})}$$  \hspace{1cm} (5.5)

Now, we can write this joint distribution $p(r_t, X_{1:t})$ recursively as follows.
\[ p(r_t, X_{1:t}) = \sum_{r_{t-1}} p(r_t, r_{t-1}, x_t, X_{1:t-1}) \]

\[ = \sum_{r_{t-1}} p(r_t, x_t | r_{t-1}, X_{1:t-1}) p(r_{t-1}, X_{1:t-1}) \] (5.6)

\[ = \sum_{r_{t-1}} p(x_t | r_t, r_{t-1}, X^{(l)}) p(r_t | r_{t-1}, X_{1:t-1}) p(r_{t-1}, X_{1:t-1}) \]

The first two steps of this equation are merely marginalization and implementing chain rule. The third step, however, is crucial as it holds if and only if the modeling assumptions are followed. The equations that fall out of the modeling assumptions are as follows.

\[ p(r_t | r_{t-1}, X_{1:t-1}) = p(r_t | r_{t-1}) \] (5.7)

\[ p(x_t | r_t, r_{t-1}, X_{1:t-1}) = p(x_t | r_t, X^{(l)}) \] (5.8)

Now, after summarizing and annotating the derivation, we get

\[ p(r_t, X_{1:t}) = \sum_{r_{t-1}} p(x_t | r_t, X^{(l)}) p(r_t | r_{t-1}) p(r_{t-1}, X_{1:t-1}) \] (5.9)

where, \( p(x_t | r_t, X^{(l)}) \), \( p(r_t | r_{t-1}) \), and \( p(r_{t-1}, X_{1:t-1}) \) denote underlying probabilistic model predictive, change point prior, and message respectively.

Now, we can compute the run length posterior with the help of this recursive algorithm. After estimating the joint distribution \( p(r_{t-1}, X_{1:t-1}) \), the message is forwarded: pass the mass of this distribution to the estimation of \( p(r_t, X_{1:t}) \). This message passing opens up a question as a recursive algorithm requires to define its initial conditions. This question is answered later in this chapter. Contextually, this recursive algorithm is comparable to hidden Markov
models’ forward algorithm: \( r_t \) is the latent variable with no generative parameters. Figure 5.2 envisages the recursive algorithm by visualizing the message-passing algorithm [187, 3].

![Message-passing Algorithm](image)

Figure 5.2: Illustrating the message-passing algorithm where each of the nodes has mass associated with it [3].

At a particular time point, this mass is either passed upward or downward. Passing upward implies the run length is incremented or growth probabilities. Passing downward indicates mass is truncated to zero. At \( t \), the discrete run-length posterior has support over \( t + 1 \) values. While computing growth probabilities, the summation disappears in (5.9) as each term in the summation comes out as zero, except when \( r_t = r_{t-1} + 1 \).

5.2.3 Underlying Probabilistic Model:

We can leverage conjugacy and exponential family to compute the underlying probabilistic model predictive. Assuming the model is from an
exponential family member with parameters $\eta$, we can calculate the underlying probabilistic model predictive as follows.

$$p(x_{t+1} \mid r_t, X^{(l)}) = \int p(x_{t+1} \mid \eta)p(\eta \mid r_t, X_{1:t})d\eta$$  \hspace{1cm} (5.10)

Where, $p(x_{t+1} \mid r_t, X^{(l)})$ represents underlying probabilistic model predictive. $p(x_{t+1} \mid \eta)$ and $p(\eta \mid r_t, X_{1:t})$ are Exponential Family model and Exponential Family posterior respectively. Here, computing exponential family posterior is crucial. Interestingly, we do not need to directly compute these integrals– a useful fact about exponential family models. We will discuss the posterior predictive for the exponential family models after reviewing the prior predictive for any conjugate model [187, 3].

### 5.2.4 Posterior Predictive for Conjugate Models:

Assuming the observations as $D$, a new observation as $\hat{X}$, parameters as $\theta$, and hyperparameters as $\alpha$, the prior predictive distribution can be expressed as follows, the distribution on $\hat{X}$ given $\alpha$ marginalized over $\theta$.

$$p(\hat{X} \mid \alpha) = \int p(\hat{X} \mid \theta)p(\theta \mid \alpha)d\theta$$  \hspace{1cm} (5.11)

Here, $p(\hat{X} \mid \theta)$ represents Model and $p(\theta \mid \alpha)$ represents Prior.

As the prior on $\theta$ is conjugate, for some different hyperparameters $\alpha'$, we have $p(\theta \mid D, \alpha) = p(\theta \mid \alpha')$

The prior has a similar functional form as the posterior if we use the conjugate prior. This can be implied by the equation presented as follows.
\[ p(\hat{X} \mid D, \alpha) = \int p(\hat{X} \mid \theta)p(\theta \mid D, \alpha)d\theta \]
\[ = \int p(\hat{X} \mid \theta)p(\theta \mid \alpha')d\theta \]
\[ = p(\hat{X} \mid \alpha') \quad (5.12) \]

In this way, the posterior predictive is matched with the prior predictive, having a hyperparameter \( \alpha' \) instead of \( \alpha' \). Now, we need to compute \( \alpha' \). The conjugate exponential model permits us to compute \( \alpha' \) efficiently. In that case, we can calculate the posterior predictive without integrating and computing the posterior \( p(\theta \mid D, \alpha) \) [187, 3].

### 5.2.5 Posterior Predictive for Exponential Family Models:

Now we will explain the prior predictive for the exponential family. Considering the natural parameter as \( \eta \), underlying measure as \( h(x) \), sufficient statistic of data as \( u(x) \), and normalizer as \( g(n) \), the general form of an exponential family member is as follows.

\[ p(X \mid \eta) = h(X)g(\eta)exp\{\eta^Tu(X)\} \quad (5.13) \]

With hyperparameters \( v \) and \( \chi \), the conjugate prior is as follows.

\[ p(\eta \mid \chi, v) = f(\chi, v)g(\eta)^vexp\{\eta^T\chi\} \quad (5.14) \]

Here, \( \eta \) and \( g(\eta) \) are same as before and \( f(\chi, v) \) is based on form of the exponential family member. Now, in an effort to have the posterior, we multiply the likelihood times the prior as follows.
\[
p(\eta|X, \chi, v) = \left(\prod_{i=1}^{N} h(X_n)\right) g(\eta)^n \exp \left\{ \eta^T \sum_{n=1}^{N} u(X_n) \right\} \left( f(\chi, v) g(\eta)^\nu \exp \left\{ \eta^T \chi \right\} \right) \\
= \left(\prod_{i=1}^{N} h(X_n)\right) f(\chi, v) g(\eta)^{N+\nu} \exp \left\{ \eta^T \sum_{n=1}^{N} u(X_n) + \eta^T \chi \right\} \\
\text{(5.15)}
\]

where \( \left(\prod_{i=1}^{N} h(X_n)\right) g(\eta)^n \exp \left\{ \eta^T \sum_{n=1}^{N} u(X_n) \right\} \) represents the Likelihood and \( f(\chi, v) g(\eta)^\nu \exp \left\{ \eta^T \chi \right\} \) is the Prior.

Considering that the first \( N + 1 \) terms are constant to \( n \), we can re-write the equation as follows.

\[
p(\eta|X, \chi, v) \propto g(\eta)^{N+\nu} \exp \left\{ \eta^T \left( \sum_{n=1}^{N} u(X_n) + \chi \right) \right\} \\
\text{(5.16)}
\]

This is the same exponential family form as the prior with parameters \( \nu' \) and \( \chi' \).

\[
\nu' = \nu_{\text{prior}} + N \\
\chi' = \chi_{\text{prior}} + \sum_{n=1}^{N} u(X_n) \\
\text{(5.17)}
\]

This lets us compute the updated hyperparameters \( \nu' \) and \( \eta' \) efficiently, resulting in facilitating the computation of the posterior predictive in (5.2) without integrating and computing the posterior over \( \eta \). It is to mention that these sufficient statistics are additive, and hence is computed sequentially [187, 3].

### 5.2.6 Message-passing Parameters:

The takeaway of the previous two discussions is that it is only required to keep track of the exponential family parameters by \( t - 1 \) to make a prediction at time \( t \) to compute the underlying probabilistic model predictive, rather than by
calculating the exponential family posterior and then integrating out the parameters. As depicted in Figure 5.2, \( r_{t-1} \) takes \( t \) possible values at a time point \( t \). \( p(x_t \mid r_{t-1} = l, X^{(l)}) \) can be denoted by \( p(x_t \mid \nu_{t-1}^{(l)}, \chi_{t-1}^{(l)}) \), when \( r_{t-1} \) takes value \( l \) and exponential family parameters at time \( t \) is denoted as \( \nu_{t-1}^{(l)} \) and \( \chi_{t-1}^{(l)} \).

Similar to what is illustrated in Figure 5.2, we can visualize the parameter updates as another message-passing algorithm in Figure 5.3 that is living on a trellis of possible run-lengths.

![Diagram](image)

**Figure 5.3:** Illustrating parameter updates that is living on a trellis of possible values.

At each time point \( t \), in the case of modeling a new possible run-length, each of the parameters’ value is just its prior. Alternatively, for each run length value \( r_t = l \), we can compute \( \nu_{t}^{(l)} \) and \( \chi_{t}^{(l)} \) by message-passing up the trellis [187, 3]. These can be presented in the following fashion.
\[v_t^{(0)} = v_{\text{prior}}\]
\[\chi_t^{(0)} = \chi_{\text{prior}}\]
\[v_t^{(l)} = v_{t-1}^{(l-1)} + 1\]
\[\chi_t^{(l)} = \chi_{t-1}^{(l-1)} + u(x_t)\]

(5.18)

### 5.2.7 Change Point Prior:

The purpose of the change point prior is to let us encode the domain knowledge about the data into model. For the given current run length \(\tau\), assume \(T\) is the discrete, nonnegative random variable and \(f(t)\) is the corresponding probability, \(S(\tau)\) is the survival function at \(\tau\). Survival function, conceptually, represents the probability that \(T \geq \tau\).

\[S(\tau) = \mathbb{P}\{T \geq \tau\} = \sum_{\tau' = \tau}^{\infty} f(\tau')\]  
(5.19)

After that, the hazard function \(H(\tau)\) can be defined as follows.

\[H(\tau) = \frac{f(\tau)}{S(\tau)}\]  
(5.20)

The original form of \(H(\tau)\) depends on the distribution of \(T\). \(H(\tau)\) estimates the probability that a change point will occur at \(\tau\), given that a change point has not occurred by the run length \(\tau\).

This modeling is assuming that our change point prior is as follows.

\[p(r_t \mid r_{t-1}) = \begin{cases} 
H(r_{t-1} + 1) & \text{if } r_t = 0 \\
1 - H(r_{t-1} + 1) & \text{if } r_t = r_{t-1} + 1 \\
0 & \text{Otherwise}
\end{cases}\]  
(5.21)

Here, (5.21) requires to estimate two cases: first, the probability of a
changepoint (representing by the associated hazard function), and second, the probability that a changepoint does not occur (representing by one minus the associated hazard function). Since it only requires compute $H(r_{t-1} + 1)$ once for both cases, estimating prior is computationally an efficient process. This process is memoryless as $H(r_{t-1} + 1)$ is time-independent.

5.2.8 Initial Conditions

Previously, we discussed the initial conditions for the recursive algorithm as follows.

$$p(r_0 = 0, X_{1:t} = \emptyset) = p(r_0 = 0)$$ (5.22)

Now, for the initial conditions, we may confront two scenarios. First, a changepoint has occurred immediately before the first data point is observed. This is represented as $p(r_0 = 0) = 1$. As far as the modeling assumption is concerned, this scenario may not be a realistic one. However, it is the simplest one from the implementation aspect. Second, we can hypothesize that a changepoint will occur in the future, considering we have access to the recent data. For the second case, prior over the initial run length can be defined as the normalized survival function as follows.

$$p(r_0 = \tau) = \frac{1}{Z} \sum_{\tau' = \tau + 1}^{\infty} f(\tau')$$ (5.23)

In this equation, $Z$ is the appropriate normalizing constant. It is estimated by dividing all the mass in the future by all the mass of the distribution $f(\cdot)$. Algorithmically, the greater the value of $\tau$, the more mass the sum accumulates. Contextually, the probability that a changepoint will occur is higher for a large value of $\tau$. 
5.3 Sequential Inference Algorithm

Based on the theoretical and mathematical explanation provided above, the sequential inference of the Bayesian online change point detection is illustrated in Figure 5.4.

5.4 Potential Challenges

There observed a recent surge in hospitals integrating Electronic Health Records (EHR) in their systems in an exponential manner since the HITECH
Act was enacted in 2009. EHRs are primarily designed to advance healthcare providers' administrative and operational efficacy, leverage data mining power, and advance analytics. Now, the secondary use of EHR data has been extended to data-driven scientific studies and medical informatics applications. To facilitate both objectives, the modern hospitalization systems now provide more frequent data regarding bedside monitoring of vital signs. For instance, in the case of MIMIC-III, which captured hospitalization data about ten years ago, we observe vital sign data every thirty-minute (in some cases, an hour). But, current hospitalization facilitates vital sign data to integrate to EMR every minute with a time-phased lagging. The time-phased lagging is not considerable. Being integrated into EMR, the more frequently captured data may induce noise in detecting structural changes in the physiological parameters. For some reason, suppose a patient may experience a spike in any of his vital sign data stream, which does not reflect a medical reason. And this spike, reflecting an external reason, may propagate, resulting in subsequent higher values in further four-five minutes. Suppose the Bayesian Online Change Point Detection Algorithm is employed as it is described in the last section. In that case, it may trigger and detect a structural change in the data stream, which may not be a useful detection for the practitioners.

5.5 Contextually-tailored Bayesian Online Change Point Detection

To address this issue, we– after initializing– rather propose to take the average of the data in ten minutes intervals and consider the average value as the new datum confronted by the algorithm. This algorithm has the potential to address the challenges discussed above. Algorithm 1 illustrates the contextually-tailored Bayesian online change point detection method for a data stream with prediction.
This section manifests an illustration and implementation of the algorithm in a proving manner. Figure 5.5 presents a data stream of 130 observations, supposing, as the context demands, each of the observational values represents a mean of ten values in a ten-minute time frame (data collected in the one-minute interval). We generate the simulated data such as the first 30 observations follow a Gaussian distribution with mean=0 and sigma=1; subsequent 50 observations follow a Gaussian distribution with mean=3 and sigma=1, successive 20 observations follow a Gaussian distribution with mean=1 and sigma=1, and then, the last 30 observations maintain a Gaussian distribution with mean=4 and sigma=1. Figure 5.5 depicts the three ground-truth change points with vertical bold red dashed lines. As the algorithm treats the first observation as a change point (intuitively understandable), we use a vertical bold dash line for the first observation for the sake of simplicity and clarity. As the observed (after taking the average) data behaves roughly Gaussian, having a variable mean and reasonably constant variance, we can model the data representing as Gaussian model with mean, \( \mu \) (unknown) and precision, \( \lambda = 1/\sigma^2 \) (constant).
The posterior predictive (the UPM predictive) for this model is

$$p(\hat{x}|D) = N(\hat{x}|\mu_T, \frac{1}{\lambda} + \frac{1}{\lambda_T}) \quad (5.24)$$

where $\hat{x} \in \mathbb{R}$ is a new observation and $\mu_T$ and $\lambda_T$ are defined as follows.

$$\mu_T = \frac{\mu_{prior}}{\lambda} + \frac{\sum_{t=1}^{T} x_t}{\lambda_{prior} + \frac{\lambda}{\lambda_T}} \quad (5.25)$$

$$\lambda_T = \lambda_{prior} + \lambda T \quad (5.26)$$

To make the algebra easier, assume that $\lambda = \lambda_{prior} = 1$. Now, we have

$$\mu_T = \frac{\mu_{prior} + \sum_{t=1}^{T} x_t}{T + 1} \quad (5.27)$$

$$\lambda_T = 1 + T \quad (5.28)$$

In change point detection, the number of samples $T$ is just the current time point $t$. So, the parameter updates for $\mu_t^{(l)}$ and $\lambda_t^{(l)}$ as $t$ increases are as follows.

$$\mu_0^{(l)} = \mu_{prior} \quad (5.29)$$
Figure 5.6: Implementing proposed algorithm realizing scenario presented in 5.5.

\[
\begin{align*}
\mu_1^{(l)} &= \frac{\mu_0 + x_1}{2} \\
\mu_2^{(l)} &= \frac{2\mu_1 + x_2}{3} \\
\mu_3^{(l)} &= \frac{3\mu_2 + x_3}{4} \\
\lambda_0^{(l)} &= \lambda_{prior}^{(l)} \\
\lambda_1^{(l)} &= 1 + \lambda_1^{(l)} \\
\lambda_2^{(l)} &= 1 + \lambda_2^{(l)} \\
\lambda_3^{(l)} &= 1 + \lambda_3^{(l)}
\end{align*}
\]

In both instances, the parameter update is a simple function of the previous parameters and the next data point.

In Figure 5.6, we implement the proposed Bayesian online change point
detection algorithm to realize the scenario presented in figure 5.5. The algorithm identifies structural changes in data stream in 1,31, 81, 102 which exactly captures the actual scenario.
Chapter 6

SepIINav (Sepsis ICU Navigator): A Data-driven Software Tool for Sepsis Monitoring and Intervention using Contextually-tailored Bayesian Online Change Point Detection

6.1 Motivation and Significance

Sepsis is a syndrome induced by the human body’s overreaction to an infection. This unregulated response results in biochemical, physiological, and pathological abnormalities in the human body [27, 9]. An epidemiological study shows that more than 1.7 million adults are affected by sepsis in the United States each year. A study reports that, annually, over 970,000 cases associated with sepsis are admitted to U.S. hospitals. Annually, more than 250,000 deaths (representing approximately 50% of all hospital deaths) are reported due to sepsis [8, 10, 11, 14]. Moreover, recent statistics reflect an 8.7% annual increase in sepsis incidence among hospitalized patients [11, 16]. Apart from the high mortality-rate and alarmingly increasing sepsis incidences, the average Length of Stay (LOS) for sepsis in hospitals is approximately 75% higher than that of most other conditions [17, 16]. In 2013, sepsis caused more than $24 billion of hospital expenses, reflecting 13% of total U.S. hospital costs. Not surprisingly, the cost associated with sepsis management and intervention ranked highest among the admissions for all diseases and medical conditions [20, 21]. In the global spectrum, a study shows that globally, 19.7% of all deaths (1 in 5) are sepsis-related [188]. With the increase of scientific data available on COVID-19, now it is evident that COVID-19 does engender sepsis as it is associated with direct viral invasion. With the notable increase in mortality rate and annual health care expenditure, the recent upsurge in the COVID-19 pandemic made
sepsis treatment and research a critical domain in medical and biological sciences [189, 190].

These profound clinical significance and motivation reflect in literature as there is considerable research found on sepsis diagnosis, treatment, and event prediction [30, 31, 32, 33]. However, the studies concentrated on one risk factor at a time for the clinical assessment of sepsis limited the probability for sepsis detection since it requires complex reasoning and implications [44, 45]. Besides, it– in many cases– is apparent that results are sensitive to subtle variations in definition(s), as well as subjective suspicions of physicians. Furthermore, evidence shows that early prediction models of septic shock using Electronic Health Records (EHR) are less useful in practice as the EHR data often comprise systematic biases that may violate the conceptual assumptions made by State-of-the-art machine learning algorithms. Moreover, confounding medical interventions during the Intensive Care Unit (ICU) stay may affect the predictive outcomes strongly [172].

Addressing the uncertainty around predictions on septic shock onset and multiple organs dysfunction syndrome in the ICU monitoring, detecting the regime-switching in univariate and multivariate physiological signals can help the clinicians understand the transition dynamics and the behavior of sequential events [185, 173]. Biological systems are, by nature, susceptible to spontaneous as well as recurrent transitions, which can be gradual or abrupt between healthy and pathological states. Bedside monitoring of physiological parameters allows the macroscopic collection of episodic events to capture the pathophysiology and delineates it as a time-series [173]. By identifying the abrupt variation between physiological signals’ structure of individual patients, we can assist practitioners in focusing more on the critical transitions. These transitions may imply either detecting changes in conditions resulting from treatments or suggesting to treat
adverse events preemptively. Accurate detection of these vital transitions can prompt early diagnoses and interventions, leading to improved clinical outcomes [174, 173, 3].

In an effort to help clinicians to navigate the patients’ trajectories, investigate the treatments’ effectiveness on patients, and review (if needed) the treatment plan in the ICU for sepsis monitoring and intervention, this chapter proposes an ICU sepsis navigation application. This application adopts the Bayesian Online Changepoint Detection algorithm to detect the univariate structural change in vital signs. Here, to facilitate navigating patients’ trajectory, we follow both of the standard criteria—Sepsis-3 (qSOFA and SOFA) and Sepsis-2 (SIRS). Apart from that, clinicians can navigate patient’s condition and transition in sepsis spectrum and the corresponding pathological transitions, if requires. Besides, practitioners can review their interventions in detail. Furthermore, as respiratory rate is the only common parameter between Sepsis-2 and Sepsis-3, this application will facilitate practitioners to understand how respiratory rate is correlated with other frequently observed physiological parameters.

6.2 Software Description

As sepsis is a broad term applied to capture a process that is not entirely understood yet, the task force of 2016 convened by the Society of Critical Care Medicine and European Society of Intensive Care Medicine addressed the pragmatic compromise and emphasized on generalizability and readily measurable identifiers to define sepsis-3. Hence, both the qSOFA and SOFA reflect the current conceptualization of underlying dynamics using physiologic and biochemical tests offered in routine patient care [191, 27].

SepINav identifies a patient by its hospital admission id. This
application, plugged into hospital EMR, can extract patients’ information (to get a deidentified idea about host factor), vital signs, lab tests, interventions, and prescription records for the purpose of monitoring. While monitoring conditions, several patients’ information, such as ethnicity, age, religion, marital status, language, and admission type, are available for the practitioners. Then, the practitioners can observe detailed information regarding their interventions in a timeline, which includes item, category, parameter type, amount, unit of measurement of the amount, rate, unit of measurement of rates, order component type, order category, patient weight, dose amount, unit of measurement of doses, and status (continued/discontinued/stopped). Multiple interventions at the same time are represented separately to avoid any confusion. Immediately below the intervention visualization, practitioners can observe the patients’ condition in the sepsis spectrum. As some hospitalizations and practitioners have not adopted the recent theoretical advances in the sepsis diagnostic criteria yet, this application offers practitioners to observe patients’ conditions in both Sepsis-2 and Sepsis-3. Practitioners and researchers will have an option on the top-left side to choose whether they want to monitor the condition in Sepsis-2 and Sepsis-3 [191, 27].

For Sepsis-3, the spectrum has three clinical states: Non-sepsis, Sepsis, and Septic Shock. On the other hand, the Sepsis-2 spectrum involves four clinical states: Non-sepsis, Sepsis, Severe Sepsis, and Septic Shock. This application strictly follows the criteria provided for Sepsis-2 and Sepsis-3 in the consensus conference in 2001 and 2016, respectively [191, 27]. Figure 6.1 illustrates the algorithm used to detect different clinical states for Sepsis-3 [27].

Besides, this application facilitates practitioners and research by identifying points where structural changes occur in the data stream for all the vital signs and critical parameters associated with sepsis-2 and sepsis-3. In the previous chapter, we discuss the rationales behind selecting the primary change
Figure 6.1: Sepsis-3 clinical criteria to diagnose patients with sepsis and septic shock.

point detection algorithm for this application and its contextual tailoring.

Here, in this Contextually-tailored Online Change Point Detection, each node includes five state values—first, the probability density function (pdf) value of the new datum; second, the new distribution values computed, factoring the new datum in underlying distribution; third, Hazard value; forth, growth probability of run-length (of a data stream with change point) represented as the multiplication of three probabilities: the probability of the continuation of the run-length till "r-1", the probability that no Hazard happened till arriving the run length to "r," pdf for the new datum in terms of the underlying probability
Figure 6.2: Contextually-tailored bayesian online change point detection algorithm with prediction.

distribution; and fifth, the change probability value for that particular node.

Figure 6.2 illustrates the contextually-tailored bayesian change point detection algorithm to detect change points in the physiological parameters reflected by the interventions provided by the practitioners [3, 185, 5].

6.3 Illustrative Examples

Figure 6.8 shows a segment of the interface that facilitates monitoring the patient’s condition in Sepsis-2 spectrum and practitioner’s interventions in detail.
throughout the entire trajectory. The input sequence helps to understand how many individual interventions are performed. Visualizing the patient’s condition in the Sepsis-2 spectrum (Non-sepsis, sepsis, severe sepsis, and septic shock) integrated with the intervention timeline will promote a better understanding of the impact of an intervention on the individual patient’s condition. In the top left corner, there is an option to select a patient with his hospital admission id and an option for navigation, such as Sepsis-2, Sepsis-3, and correlating vital signs. In the top right corner, one can observe the preliminary information about a patient collected during admission.

Figure 6.7 shows a segment of the interface that facilitates monitoring the patient’s condition in the Sepsis-3 spectrum (Non-sepsis, sepsis, and septic shock) and practitioner’s interventions in detail throughout the entire trajectory.

Figure 6.3: Illustration of monitoring practitioner’s interventions and patient’s trajectory in Sepsis-2.
Figure 6.4: Illustration of monitoring practitioner’s interventions and patient’s trajectory in Sepsis-3.

In both interfaces offering Sepsis-2 and Sepsis-3 navigation, this application provides visualizing the individual vital signs’ and critical parameters’ trajectories associated with Sepsis-2 and Sepsis-3, respectively, integrated with the intervention timeline. It will help to study the impact of an intervention on the particular physiological parameters. Besides, the contextually-tailored bayesian online change point detection algorithm will provide the points where structural changes occur in the data stream for all the vital signs and critical parameters associated with sepsis-2 and sepsis-3. Figure 6.5(a) shows a snap of change points in a heart rate data stream (representing a critical parameter in SIRS of sepsis-2) of a patient visualizing in the tool and associated scientific computation (Bayesian probabilities and run lengths) in the background. Figure 6.5(b) shows change points in the respiratory rate data
Figure 6.5: Structural change detected in the patient’s vital signs using bayesian online change point detection algorithm.

Apart from that, this tool will let the practitioners know how other frequent vital signs, such as heart rate, Glasgow Coma Scale score, systolic arterial blood pressure, temperature, white blood cell count, mean arterial pressure, bilirubin, creatinine, and P/F ratio, are correlating in terms of the respiratory rate. We select respiratory rate as the standard parameter to compare correlation since it is the only common bed-side vital sign in both Sepsis-2 and Sepsis-3. The underlying relationship between the parameters may vary patient to patient (some may show linearity, some other may offer strong nonlinearity), or it may be affected by the interventions provided in the ICU.
Figure 6 depicts the interface facilitating demonstrating the correlation between parameters (for example, with heart rate, Glasgow Coma Scale score, systolic arterial blood pressure, and temperature). Here, one can see correlations for the entire data (both under the safe zone and sepsis) or just select data under a safe zone or data that met sepsis criteria.

Figure 6.6: Illustrating the option to observe correlations between vital signs crucial for Sepsis.
6.4 Usability, Impact, and sustainability

To understand and interpret the utility of the change point detection in patients’ vital sign regimes, we adopt MIMIC-III (version 1.4) datasets. MIMIC-III (version 1.4) offers openly available datasets developed by MIT Lab for Computational Physiology. It integrates de-identified comprehensive clinical data of approximately 40,000 distinct critical care patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts. MIMIC-III (version 1.4) is widely accessible to researchers internationally under a data use agreement. It includes datasets regarding demographics, vital signs (bedside monitoring), laboratory tests, medications (practitioners’ input), and more. Our study primarily concentrates on CHARTEVENTS (containing 330,712,483 observations) and ADMISSIONS (containing 58976 admissions) datasets. The filter-in process considers the hospital admission ids of patients diagnosed for sepsis and then churns out vital signs data crucial for both Sepsis-2 and Sepsis-3 for these particular patients. To further understand results and implications, we consider the following datasets of MIMIC-III (version 1.4): INPUTEVENTS, LABEVENTS, MICROBIOLOGYEVENTS, PROCEDUREEVENTS, and NOTEEVENTS. It is crucial to note that MIMIC-III captured patients’ data between 2001 and 2012 and hence does not reflect the data frequency offered by the modern hospitalization. This study—considering this limitation—circumscribes to illustrating the potential utility of the change point detection in early risk assessment dynamics.

Sepsis-3 reformulates the term "Sepsis" and "Septic Shock" from the perception we had for Sepsis-2. The definition of sepsis undergoes two-step assessments: first, qSOFA for patients with suspected infection, and second, SOFA seeks organ dysfunction evidence. qSOFA incorporates three criteria: high
respiratory rate (22 breaths/minutes), altered mental status (13 on the Glasgow Coma Scale), and low blood pressure (Systolic Arterial Blood Pressure 100 mm Hg). If it meets any two or more criteria, qSOFA assessment prognosticates sepsis-related poor outcomes. qSOFA vitals signs are collected at bedside monitoring and well-reflected in frequent MIMIC-III CHARTEVENTS (version 1.4) observations. We study data streams for respiratory rate, GCS (Glasgow Coma Scale), and Systolic Arterial Blood Pressure for a particular hospital id (de-identified data). Then, leveraging the sepsis definition criteria proposed in JAMA and INPUTEVENTS, LABEVENTS, MICROBIOLOGYEVENTS, and PROCEDUREEVENTS from MIMIC-III (version 1.4), we identify patient’s condition in the sepsis-3 spectrum.

Since biological systems are, by nature, susceptible to spontaneous and recurrent transitions– which can be gradual or abrupt between healthy and pathological states, bedside monitoring of physiological parameters allows the macroscopic collection of episodic events to capture the pathophysiology and represents it as a time series. By identifying abrupt variation between physiological signals’ structure of individual patients, we can assist practitioners in focusing more on critical transitions. These transitions may imply either detecting changes in conditions resulting from treatments or suggesting to treat adverse events preemptively. Hence, the primary objective of this study is to identify structural changes in patient’s relevant vital signs regimes. Sepsis pathophysiology illustrates that the host’s unregulated responses to pathogen or infection reflect in vasodilation, capillary leak, blood clotting, and metabolic acidosis, resulting in structural changes of the vital signs, such as respiratory rate and systolic arterial blood pressure. Sepsis pathophysiology also indicates that when the host first experiences vasodilation, cardiac output initially increases, balancing out the drop in systemic vascular resistance and keeping the
blood pressure the same—since blood pressure is proportionate to cardiac output times systemic vascular resistance. This scenario is well-reflected in the change point detection paradigm as we observe more change points detected in respiratory rate compared to systolic arterial blood pressure (Figure 6.7).

Figure 6.7 illustrates a particular patient scenario. Figure 6.7(a) depicts the patients’ scenario in the sepsis-3 spectrum using SepINav. The patient was diagnosed with sepsis. In SepINav, if the cursor is put on the patient’s status in the sepsis spectrum, it will show the specific time and date of the observation along with the patient’s status. We notice that the patient initially experienced
multiple switches (or shifts) in status from sepsis to no-sepsis condition in a smaller time segment from 4 PM, June 09 to 7 PM, June 10 because of medications, compounded with host factors and pathogen factors. MIMIC-III deliberately represents an unrealistic year, albeit keeping month, day, and time the actual one, to provide de-identified patient data for research purposes. The structural changes in Respiratory Rate (6/10-4:00 PM, 6/10 9:00 PM) well-captured these pathological incidences. The absence of structural changes in Systolic Arterial Blood Pressure and Glasgow Coma Scale score is understood by sepsis pathophysiology.

Then, that patient’s status got exacerbated, from sepsis to septic shock, at 10:00 PM, June 12, got intervened, and returned to the previous condition, sepsis, at 5:00 AM, June 13. We also observed structural changes in Respiratory Rate (6/12-3:30 PM, 6/12-11:00 PM, 6/13-8:00 PM, and 6/14-12:00 PM), Systolic Arterial Blood Pressure (6/11-10:14 AM, 6/14-5:00 PM), and Glasgow Coma Scale score (6/12-8:00 AM) in that synchronal time frame.

After that, we observed the patient’s condition got deteriorated to septic shock at 08:06 AM, June 24, and then turned back to sepsis at 11:00 AM, June 25, and again aggravated to septic shock at 8:00 PM, June 26, and then got back to the condition, sepsis, at 1:00 PM, June 27. These incidences observed structural changes (or change point detected) in vital signs: Respiratory Rate (6/23-4:00 PM, 6/26-3:00 PM, 6/27- 1:00 AM, and 6/27-4:00 PM), Systolic Arterial Blood Pressure (6/23-11:00 AM, 6/26-4:00 AM, 6/27-11:57 AM), and Glasgow Coma Scale score (6/24- 0:30 AM). And after maintaining the sepsis condition for a particular time, we again observed that the patient’s condition got aggravated to septic shock at 4:00 AM, July 2. We also observed structural changes in Respiratory Rate (6/30- 0:00 AM, 7/2- 9:00 AM) in a similar time frame. Overall, in a 24 day-stay of this patient, we observed 17 change points in
his/her Respiratory Rate data stream, 8 change points in his/her Systolic Arterial Blood Pressure data stream, and 8 change points in his/her Glasgow Coma Scale score. To avoid any perceptual confusion, it is to mention that the mathematical framework of Bayesian online change point detection assumes the first data point as a change point. According to MIMIC-III, the patient expired at 9:35 PM, July 3.

It is to reiterate that the primary purpose of this endeavor is to assure efficient monitoring and investigate the impact of the interventions in patient’s vital signs and disease pathophysiology from a more subjective lens and practitioner’s perspective. The secondary purpose is to understand the state (or structures) of the vital signs and sepsis parameters towards the patient’s status in the sepsis spectrum.

To scrutinize the utility of the change point detection algorithm and this assistive risk assessment for Sepsis-2, we revisit the Sepsis-2 spectrum. It includes sepsis, severe sepsis, septic shock, and MODS. It defines sepsis using SIRS criteria with suspected infection. This study investigates SIRS vital signs data collected at bedside monitoring. Body temperature, respiratory rate, heart rate, white blood cell count are available in frequent MIMIC-III CHARTEVENTS (version 1.4) observations. We study these data streams for a particular hospital id (de-identified data) using change point detection. Then, leveraging INPUTEVENTS, LABEVENTS, MICROBIOLOGYEVENTS, and PROCEDUREEVENTS data from MIMIC-III (version 1.4), we identify and explain the patient’s condition in the sepsis-2 spectrum. Figure 6.8 illustrates a sepsis-2 scenario of the same patient using SepINav, besides R illustrations. The patient was diagnosed with sepsis. As the sepsis-2 spectrum has more states, compared to the recent version, we observed transitions from one state to another more frequently in this case. Besides, in a 24 day-stay of this patient, we
observed structural changes in vital signs or SIRS parameters: 17 change points in his/her Heart Rate data stream, 17 change points in his/her Respiratory Rate data stream, 2 change points in his/her Body Temperature, and 4 change points in his/her White Blood Cell counts.

Like all the supervised machine learning-based stand-alone solutions, predicting sepsis-related poor outcomes leveraging EMRs suffers from generalizability issues due to bias-variance trade-off. There are distinct
differences in hospitalization facilities, hosts’ responses (patients), and pathogen variants across the globe. Models developed using one EMR may not be helpful and generalizable in serving a wide variety of populations. Furthermore, it is required to scrutinize if the data represents the synthetic population. On the other hand, the interpretability of the deep learning models using EMRs can be challenged; the hidden layers leave out the model’s functionalities as a black box. This proposed change point detection-based assistive solution may bridge the perceptual gaps. First, the unsupervised nature of the solution makes it less susceptible to bias-variance trade-offs, hence more generalizable. This solution concentrates on the distinct generative parameters representing different partitions in the data stream. Second, this solution is interpretable and explainable. Furthermore, if the data stream aligned with the intervention trajectory, practitioners can interpret the consequences of the interventions and explain the changes. Hence, this change point detection-based assistive solution is a more sustainable solution toward data-driven medical informatics solution for sepsis early risk assessment, monitoring, and interventions.

SepINav embodies a step towards bridging the perceptual gap between explainability and computational intricacy centered around machine learning-based medical informatics solutions. It is represented by the facility to navigate both patients’ trajectory and the interventions made by the practitioners and to detect the changepoints in vital signs that may harbinger prior to septic shock onset. This decease-specific and custom-tailored tool has a considerable impact on ICU monitoring and interventions, as sepsis is shaped by both host factors and pathogen factors in a convoluted manner and resulted in aberrant host response. These host factors include age, sex, race, comorbidities, and different genetic determinants. As far as the treatments and interventions are concerned, the host factors—besides the pathogen factors—affect the result of
an intervention, such as an effort to improve the patient’s condition from sepsis to no sepsis may result in exacerbating the patient’s condition and leading to septic shock. This application addresses this intraspecific variability in clinical response, bolstering not one cure (intervention) for all scenarios (taking the host factor into account). This effort will help practitioners navigate how their clinical interventions reflect in patients’ responses in vital signs and sepsis spectrum. This application lets the practitioners observe patients’ conditions in terms of both Sepsis-2 and Sepsis-3 since some hospitalizations and practitioners have not adopted the recent theoretical advances in the sepsis diagnostic criteria yet. Besides Sepsis-2 and Sepsis-3, it will help practitioners to navigate patients’ situations in terms of the common vital signs available in bedside monitoring and to notice whether there is any structural change in patients’ trajectory. Again, to have a meticulous view regarding individualized treatment dynamics, ICU experts often suggest that it is crucial to understand how different vital signs correlate with each other for an individual patient and how these inter-relations are affected by interventions and therapeutics. This application adds this scope for the typical vital signs at the individual level to facilitate ICU monitoring and intervention. Apart from that, since this tool is fueled by EMR data, it opens up two scopes of functioning: first, helping practitioners in monitoring and intervening on the existing sepsis patients, and second, conducting retrospective studies on the previous patients for research purpose and seeking rationales to different sepsis scenarios confronted in the ICU.

From the aspect of sustainability of the solution, changepoint detection has the potential to address challenges and pitfalls confronted while developing machine learning-based predictive models using EMRs. The granular measurements available in the EMRs are used for early prediction and risk stratification, resource optimization, and biomarker discovery. On the
bright-side, training and evaluating early detection systems using EMRs has several benefits over using data from controlled studies, namely the abundance of available data, zero data collection cost, and addressing privacy issues. However, while blindly relying on symptoms and vital signs data available in EMRs, confounding medical interventions may mask the ground truth labels, though required to address in training and evaluating phase of a prediction system. Confounding Medical Interventions (CMIs) are the interventions that the practitioners execute that may affect the risk associated with the outcome of interest, such as administration of pressors and broad-spectrum antibiotics for sepsis and septic shock. After a CMI, it is not easy to distinguish between a patient who was at risk but is now treated due to antibiotic treatment, and a patient received unnecessary treatment due to conservative judgment by the caregiver. Again, a predictive model, considering N hours of data to predict patients’ situations after M hours, can lead to being less useful in practice, if not taking this masking into account. CMIs from N+1 hour to M-1 hour can mislead the predicted situation in a significant margin. In the like-manner, a model trained on data from one system might not generalize to another system. We stress that change point detection introduced in SepINav is a more sustainable solution in patient monitoring, early-sign detection, and intervention, as far as the CMIs are concerned. Besides, it offers flexibility to the users considering the variability in different practices. In a certain way, this application can inspire new and better software to address different challenges and pitfalls in patient monitoring and risk stratification.

6.5 Summary

We presented SepINav, a data-driven software tool that allows ICU practitioners to navigate patients’ trajectory for sepsis, considering both sepsis-2
and sepsis-3 for representation. It facilitates navigation as the practitioners can observe the impact of their interventions in patients’ conditions in the sepsis spectrum. We accentuated that the structural changes in patients’ vital signs often warn the practitioners before septic shock onset. This tool helps to capture these changepoints using Bayesian Online Change Point Detection. This application demonstrated its effectiveness (a) in monitoring and intervening on the existing sepsis patients and (b) in conducting retrospective studies for research purposes and seeking rationales to different sepsis scenarios confronted in the ICU. Figure 6.9 summarizes the functionalities of SepINav.

Figure 6.9: A Step Towards Evidence-based and Data-driven Decision Making in Sepsis Monitoring and Intervention
Chapter 7

Computational Sustainability and Internet-of-Energy

7.1 Introduction

Computational Sustainability, a massively interdisciplinary field of study, lies in the intersection of the multiple domains, such as applied mathematics, statistics, computer and information science, electrical and electronic engineering, economics, environmental science, operational research, and policymaking [192, 193]. The overarching goal of this field of study is leveraging the knowledge of these multiple domains to meet the essentials and demands of the current generation without compromising the future generation’s potentiality to confront their known needs and prosper [194, 195, 196]. Computational Sustainability joins the movement of sustainable development through developing data-driven and robust computational models and adopting scientific methods to optimize decisions regarding resource allocation and management with the motivation to solve the most challenging sustainability problems [197, 198, 199, 200]. The rise of Big Data and advanced analytics have contributed to the recent surge in this effort [201, 202].

The advent of the Big Data era brings scopes and opportunities for computational sustainability research regarding multi-dimensional challenges, complexities of the problems, scalability issues, computational efficacy, and impact towards overarching motivation [203, 204, 205]. This abundance of data not only comes with ample information and potential knowledge but also offers a scientific approach driven by multi-source data and enhances the efficiency and accuracy of problem-solving. That is why the growth of extensive multi-dimensional data and computational sustainability are crucial to meet the
sustainability challenges [203, 206]. They contribute to addressing tradeoffs in
scientific decision making, understanding complicated systems, and explaining
uncertainties with complex reasoning [207, 208]. CompSustNet, a unique virtual
network led by Carla Gomez at Cornell University and supported by the National
Science Foundation (NSF) of the United States, establishes on the research,
results, and achievements of the ICS (Institute of Computational Sustainability)
[192, 193, 209]. It unites and helps more and more scholars, across the domain,
use data mining techniques to solve the most complex and pressing problems of
this time, such as efficient and reliable energy management [210, 211], healthcare
[212, 213, 214], biodiversity loss protection, addressing issues regarding climate
change and environmental collapse [215, 216], poverty eradication [217, 218],
meteorology [219, 220, 221], disaster management [222, 223], and material
discovery for renewables sources [224, 225, 226, 227]. The most compelling
aspect of this virtual network—besides making a platform for computational
science researchers to put their muscle towards making the world a more
sustainable and livable place—is that a new method or solution created to solve
one particular problem can be repurposed for another distinct problem.

One of the major attention of computational sustainability research is
centered around the question of how we can leverage Big Data accumulated from
the smart grid components and raise collective awareness and proactive
demeanor towards smart and sustainable energy management [228, 229]. Besides
Big Data, the recent advancement of information and communication
technologies allows the regime switch from a traditional "predict (forecast) and
provide" approach to a more flexible and responsive demand-based approach of
power system management. The purpose of this approach is to reach several
policy targets regarding sustainability, such as reducing carbon emissions,
generating power from renewable resources to a certain percentage, smoothing
peak demand, assuring a better rate of return on investments, and preventing network overprovisioning [230, 206, 231].

Smart Grid technology facilitates more accurate energy-loss monitoring and more precise control and adaptive techniques by escalating the intelligence and capacity of the energy distribution, as well as the control system, from the central cores to numerous peripheral nodes [232, 233, 204]. On a different note, recent studies showed that IoT is looming as a significant trendsetter in realizing the advancement of information and communication technologies, and analytics at a considerable dimension. IoT enables connecting, monitoring, and controlling the physical objects used in our day-to-day life by extending the web paradigm. It engenders more frequent and impactful human-to-machine and machine-to-machine interactions in everyday life. Smart Grid is one of the recent inclusions in this avenue, realizing the concept of the Internet of Energy (IoE) [234, 235, 236, 237].

The overarching motivation behind the Internet of Energy (IoE) is assuring a flexible but highly reliable and resilient, cost-effective, and efficient power supply network in the combination of large-scale centralized generators and small-scale renewable sources. IoE can convincingly be defined as a network infrastructure that enables a real-time balance between the local and global generation and storage capability based on the energy demand of the consumer [238]. It allows for a high level of consumer awareness and involvement with the help of advanced analytics. From the functional point of view, IoE, de facto, integrates power distribution, energy storage, grid monitoring, and synchronous and asynchronous communication, as illustrated in Figure 7.1. This network infrastructure is built on the standard and interoperable communication transceivers, gateways, and protocols. Besides, by taking advantage of widely accepted security and privacy frameworks, it can assure seamless interoperability
and broad connectivity. And, by leveraging the power of cloud computing systems, it can promote service virtualization and distribution [239].

In this chapter, we primarily discuss the Grid Overview of the United States; Weather and Climate and its impact on the entire energy generation and consumption dynamics; Peak Load Forecasting and its techniques and burgeoning challenges; Variable Renewable Energy, its reliability challenges and how we can take advantage of this variability; Commodity Prices and its criticality; Energy Disaggregation and its impact in consumption awareness; and Generation Expansion and Decision Analysis and trade-offs before addressing IoE integration, and challenges. This chapter manifests how these conceptual elements are correlated with each other in the energy internet. This qualitative study pursues a process of scientific inquiry that seeks an in-depth understanding
of scientific phenomena and its cause and effect in respective contextual settings. It primarily concentrates on answering "why" rather than "what" of the scientific phenomena and entrusts on the evidence manifested in the literature, in addition to comments and suggestions by the domain experts at Oregon Renewable Energy Center (OREC) at Oregon, United States.

7.2 Grid Overview of the United States

With the evolution of the energy industry, utility companies- for the very first time in the United States- adopted the joint operations in order to share the peak coverage and backup power in the 1920s after a more than fifty years adherence to the conception electric energy needs to be produced near the device or service requiring that particular energy. Then, in this regard, the Public Utility Holding Company Act was passed in 1934, realizing the electric grid of the United States with outlined restrictions and regulatory oversight of operations. Later on, to date, the Energy Policy Act of 1992 and the Energy Policy Act of 2005 are considered as the stepping stone of the modern electric grid of the United States [240, 241, 242]. The first one granted the electric

![United States Transmission Grid](image)

Figure 7.2: United States synchronous grid of 186, 411 miles.
generation companies open access to the transmission line network and initiated competition in power generation as opposed to vertical monopolies, where generation, transmission, and distribution were administered by a single authority [243]. The later one promoted the alternative energy production and greenhouse emission free cutting-edge technologies with incentives and loan guarantees [244, 245].

United States interconnects are synchronized at 60 Hz, unlike those of Asia and Europe operate at 50 Hz. The interconnects in the United States are tied to each other either via DC ties (HVDC power transmission lines) or with VFTs (variable frequency transformer), allowing a controlled flow of energy, and at the same time, isolating the each side’s independent AC frequencies functionally. The advantages of having synchronous zones consolidated by the utility grid include pooling of generation, pooling of load, common provisioning of reserves, opening of the markets, and collective assistance in the event of interruptions [246]. On the contrary, the possibility of repercussions (like a chain reaction) across the entire grid, if any problem happens in one part, is a certain threat in the case of synchronous grid [247]. The United States synchronous grid is presented in Figure 7.2, consisting of about 186,411 mi operated by more than 500 companies [247].

The United States utility electric grid is expected to confront certain challenges in the years to come posed by the modern power generation and distribution systems. [248] identified the challenges and the reasons behind it, and contemplated the possible solutions of them in a comprehensive fashion. In a nutshell, the challenges- categorized there based on the severity and impact- are cyber threats and attacks in utility, challenges in transmission system, from aging infrastructure, regulatory challenges, challenges in workforce, challenges from distributed generation and mixed sources of generation, challenges from the
intermittent nature of renewable energy sources, challenges from microgrid and smart grid, challenges from communication, challenges from energy storage systems and evolving technologies, and challenges from system complexity and cost issues. Here, their feasible solutions— with detailed explanation and depiction— include cyber security measures, upgrading the system infrastructure, new business strategies, compensating for the intermittency of renewables (concentrated on the law of large number, power of prediction, incentivizing energy production at the right time and place), and the proper use of energy storages. Later, they presented the severity and frequency analysis for each of the challenges in the United States context.

Super Grid, commonly known as Mega Grid, aims to considerably advance the transmission capacity with a particular policy that include effectively enabling the renewable energy industry to sell electricity to distant markets, increasing intermittent energy source usage by distributing them across the extensive geological region, and trimming the congestion that averts the electricity markets from succeeding. Then, to promote the concept of integrating localized generation into the centralized generation-based distribution, microgrid technology has been introduced. In 2010, Office of Electricity Delivery and Energy Reliability of the US Department of Energy (DOE)— incorporating the final amendment made in 2017— proposed the definition of Microgrid (MG) as it is a group of interconnected loads and distributed energy resources within clearly defined electrical boundaries that acts as a single controllable entity with respect to the grid, can connect and disconnect from the grid to enable it to operate in both grid-connected or island-mode, considering that a remote MG is a variation of an MG that operates in islanded condition [249]. DOE started their major MG program in 2008, initiating with nine RDSIs (Renewable and Distributed Systems Integration) depicted as green points in Figure 7.3. These projects—
with $100M budget equally financed by the DOE and co-funders—aimed to achieve a minimum 15 peak load reductions. The red points illustrate the projects under SPIDERS program (Smart Power Infrastructure Demonstration for Energy Reliability and Security) launched in 2010. The SPIDER program was introduced to meet the fact that almost all the military bases are located in the resource-limited setups inadequately served by the utility grid where highly reliable power is often required. The first three among this effort are Hickam Air Force Base and Camp Smith in Hawaii, and Fort Carson in Colorado. Though the federal programs are the cardinal efforts to the United States MG research in the early stage, private sector activities in the recent years are noteworthy. The large commercial organizations, such as educational campus, medical institutes, and industrial sites, focused on building self-generation projects. In a dramatic fashion, these efforts made 2011-2012 a pivotal year in MG development for the United States. DOE defined the next-generation MG system with certain specific goals expected to achieve by 2020. The goals are to establish MG systems of a capacity <10 MW in commercial-scale capable of curtailing outage time of required loads by more than 98% at a cost comparable to the nonintegrated-baseline solutions, while offering more than 20% improvement both in emission reduction and energy efficiency. Research shows control and protection are the significant challenges to meet this goal [249].

According to North Carolina Clean Energy Technology Center, in 2017, 37 states—well-reflected by 82 relevant bills introduced in the different regions of the United States—endeavored to modernize electric grid to make it more interactive and resilient. These endeavors include deploying advanced metering infrastructure, smart grid, and offering time-varying rates for the residential consumers. Recently, in August 2018, a policy paper has been published with five major recommendations to modernize the United States electric power grid.
Figure 7.3: Geographical distribution of Federal MG assessment and demonstration projects in the United States.

It points out making the federal permit process more efficient and effective for advanced energy projects, inspiring grid planners look at the alternatives to making investment in transmission, promoting energy efficiency and allowing energy storage to compete with the additional generation, allowing big consumers to adopt their own source for electricity, and letting both the consumers and utilities to take advantage leveraging the cloud computing facilities [250, 251].

21st Century’s electrical grid in the United States is blessed with smart grid technology that leverages the power of two-way communication and distributed-intelligent devices, assuring improved delivery network. With the objective to enable utilities predict their demand efficiently and involve customers in smart-time-of-use-tariff, smart grid development was facilitated in the United States by Energy Policy Act of 2005 and Energy Independence and Security Act of 2007. A recent surge has been observed in the literature regarding different systems and aspects of smart grid. We can categorize these research into three clusters: infrastructure system research, research on the management system, and research on the protection system. The infrastructure
system research are aiming to meet advanced electricity generation, uninterrupted delivery, and intelligent consumption; smart information metering, monitoring, and management; and last but not the least, advanced interactive communication technology. Research on the management system—leveraging advanced machine learning, optimization, game-theoretical approaches—include improving energy efficiency, demand profile, cost, utility, and carbon emission. Most of the research on protection systems focus on grid reliability, failure, and privacy protection, security services [252, 253].

In 2017, Utility-scale facilities generated about 4.03 trillion KWh of electricity in the United States. Among them, majority (about 67%) of this generation was from fossil fuel, 19% was from nuclear energy, and roughly 14% was from renewable energy sources. Apart from that, United States Energy Information Administration reckoned an additional generation of 24 billion kWh from the small-scale solar photovoltaic systems, such as small-scale solar photovoltaic systems that are installed on building rooftops, in 2017 calendar year [254]. Figure 7.4 illustrates the distribution of generation from different sources. Then, Figure 7.5 shows the annual share trend in United States for electricity generation by source from 1950 to date and Figure 7.6 depicts the evolution of the generation mix contributing to the United States electricity generation over the time. The generation mix is highly affected by the resource availability of the particular state. The following figure (Figure 7.7) illustrates the net generation distribution of electricity by type and states [254].

This varying-nature of resources with time and region, along with other commercial factors, have a predominant influence on the tariff. In 2006-2007, average electricity tariff in the United States—though it varies state to state—was higher than Canada, Australia, France, and Sweden, but relatively lower than that of the United Kingdom, Germany, and Italy among the developed countries,
Figure 7.4: Distribution of generation from different sources in the United States in 2017.

Figure 7.5: Annual share trend in United States for electricity generation by source from 1950 to date.

Figure 7.6: Evolution of the generation mix contributing to the United States electricity generation over the time.
and the average residential bill was noted $100 per month. A statistics of 2008 shows the United States average electricity tariff was 9.82 Cents/kWh, varying from 6.7 Cents/kWh (in West Virginia) to 24.1 Cents/kWh (in Hawaii).

Compared to that, data of October 2018 reveals that the average electricity tariff is 12.87 Cents/kWh, varying from 9.11 Cents/kWh (in Louisiana) to 32.46 Cents/kWh (in Hawaii). It demonstrates a 0.5% rise in price compared to 2017.

United States grid is organized administratively in the following order: Reliability organizations; Balancing authorities that include independent system operators, regional transmission organizations, and vertically integrated utilities; Generators comprised of utilities and independent power providers; and Load
Serving Entities [256]. NERC (North American Electricity Reliability Corporation) is the not-for-profit organization to assure reliability of the north american bulk power system. They are in charge of monitoring and enforcing compliance with standards, besides being the authority of the data source for system reliability and system failure. Since the United States power system is interconnected physically, any problem occurred in one area may influence other interconnected systems, and NERC is centrally responsible to take care of it and assure reliability. Besides, NERC’s major responsibilities include working with all the stakeholders to develop well-defined standards for power system operation, monitoring; and enforcing compliance with those standards, assessing resource adequacy, and providing educational and training resources as a part of accreditation program to ensure power system operators remained qualified and proficient in operation. NERC oversees eight regional reliability entities. The sub-parts of NERC are WECC (Western Electricity Coordinating Council), MRO (Midwest Reliability Organization), NPCC (Northeast Power Coordinating Council), SPP (Southwest Power Pool), RFC (Reliability First Corporation), SERC (SERC Reliability Corporation), FRCC (Florida Reliability Coordinating Council), TRE (Texas Reliability Entity). Traditional wholesale electricity market, which are vertically integrated so that they own and are responsible for the generation, transmission and distribution systems to serve the electricity consumers, exists in the south east, south west, and north west. In other part of the United States, the power systems are managed by Independent System Operators (ISO) and Regional Transmission Organizations (RTO), facilitating open access to transmission. In particular, ISO operates the transmission system independently, and foster competition for electricity generation among the wholesale market participants [257, 258]. The extent of ISOs are visualized by Figure 7.8.
Figure 7.8: Extent of ISOs in the United States.

Each of the ISOs and RTOs have energy and ancillary services markets in which buyers and sellers can bid for or offer generation. These ancillary services include reserves, frequency regulation (grid needs to be operated at 60Hz in the United States), and demand response. Though the vital sections of the United States operate under more traditional market structures, two-thirds of the nation’s electricity load is served in RTO regions [257, 258, 259].

7.3 Load forecasting

7.3.1 Weather and Climate of the United States

Weather and climate have an impact on both sides of the electricity industry- it drives the energy consumption demand, affects most of the noncombustible generation, and has an effect on electricity transmission and distribution. Before the expository narratives on the impacts, challenges, and state-of-the-art solutions, we need a proper understanding of three factors and its interpretation: weather, climate, and extreme weather. Here, temperature change, precipitation, humidity, and wind speed are interpreted and interchangeably used as the weather. Climate refers to the average seasonal
conditions for a particular geographical area. In our study, extreme weather will include droughts, floods, hurricanes, heat waves, and cold snaps; statistically rare weather events that have cataclysmic impacts. The weather elements that directly affect the demand are as follows: temperature, wind speed, cloud, visibility, and precipitation. For example, the temperature, being allied to wind speed, regulate heating or cooling demand. Besides, cloud, visibility, and precipitation are considered to estimate the level of daylight illumination, therefore affecting the lighting demand. Research shows that each of the attributes of these meteorological elements has weighted sensitivity to demand and the sensitivity weight varies with the geographical location of the representative region. To compare the impact of weather elements, the meteorological elements are scaled down to three specific factors: effective temperature, cooling power of the wind, and rate of precipitation [260, 261, 262].

![Electricity Consumption in US Homes](image)

Figure 7.9: Electricity consumption in US homes during 2018 (kWh/year).

Figure 7.9 shows why the demand side of the energy system is related to the weather and climate which is well-reflected by the electricity consumption in US homes during 2018 (kWh/year). As has been observed from this figure, the cardinal electricity consumption in US homes is for space heating in wintertime
and air conditioning in the summertime. Furthermore, the average length of wintertime in the US even intensifies the case. Besides, lighting and space heating engender a considerable amount of electricity consumption. Though lighting over the year is correlated with the weather and can be influenced by other factors, space heating in the wintertime and air conditioning in the summertime primarily depend on weather and climate. The relationship of Demand and Temperature is parabolically nonlinear, and the rationale behind that is when the temperature is low, it requires heating demand, and with the temperature rise the heating demand decreases, and there is a sweet temperature zone, in between 65F to 70F, when we do not need any heating and cooling.

Again, after that point, we need cooling demand, and it increases with the rise of the temperature. Such parabolic nonlinearity encouraged us to study the impact of weather parameters on electricity demand separately: heating and cooling demand; in particular, using the concept of heating degree days and cooling degree days.

A degree day compares the ambient temperature to a standard temperature of 65F. The more severe the temperature, the higher the number of degree days. A higher number of degree days will require more energy for space heating or cooling. Figure 7.10 classifies the United States based on heating degree days [263, 264]. Here, the darker the red, the more the heating is required in the winter time. As has been observed from the Figure 7.10, the west north central region of the United States, which includes North Dakota, South Dakota, Minnesota, Nebraska, Kansas, Iowa, and Missouri, requires most of the heating degree day demand. Figure 7.11 depicts the cooling degree day demand distribution over the United States [263]. Here, the darker the blue, the more the cooling is required in the summertime. It is evident from the figure that west south central region, which includes Oklahoma, Arkansas, Texas, and Louisiana,
Figure 7.10: Region classification of the United States based on heating degree days.

Figure 7.11: Region classification of the United States based on cooling degree days.

requires most of the cooling degree day demand in the summertime, and conversely, east north central and west north central region exhibit the lowest cooling degree day requirement in summer.

There has been experiencing a continual net temperature increase in the United States over the years, so the electricity demand has been. The annual average temperature over the contiguous US (48 states excluding Alaska and
Hawaii) has increased by 1.2°F for the period 1986–2016 comparative to
1901–1960 and by 1.8°F based on linear regression for the period 1895–2016.
Both surface and satellite data consistently support the fact of rapid warming
since 1979. Furthermore, Paleo-temperature evidence reveals that recent decades
are the warmest of the preceding 1,500 years. As a result, the number of
high-temperature records placed in the previous two decades considerably
outstrips the number of low-temperature records. However, the Dust Bowl era of
the 1930s remains the peak period for the extreme heat. Moreover, the annual
average temperature over the contiguous US is projected to ascend about 2.5°F
for the period 2021–2050 corresponding to 1976–2005 in all RCP scenarios. In
particular, much higher rises are expected by late century (2071–2100): 2.8–7.3°F
in a better case and 5.8–11.9°F in the exacerbated scenario [265]. Figure 7.12
illustrates the temperature anomaly from 1901 to 2015.

Figure 7.12: Temperature anomaly from 1901 to 2015 in the contiguous 48 states.

Since population distribution is not uniform over the entire United States
and population density has a strong impact on energy consumption, the US EIA
(Energy Information Administration) use population-weighted degree days to
model and project energy consumption. Mathematical modelings are involved in
incorporating the impact of weather and climate on national electricity consumption. [266] contributed to model the effect of summer temperature on electricity demand and consumption. This model includes three aspects: estimate the impacts of unusual weather (such as heat wave), consider the effects of governmental policies, assess the impacts of projected climate change on energy demand and supply. This model can be described as (7.1):

\[ E = a_0 + a_1 CDD + a_2 CDD_{(-1)} + a_3 Y_1 + a_4 Y_2 - a_5 H \]  

Here, \( E, CDD, CDD_{(-1)}, H, Y_1, \) and \( Y_2 \) stand for weekly national electric output in billions of kWh, weekly national cooling degree day total, previous week’s national cooling degree day total, holiday factor, penultimate year growth factor, and last year growth factor. This is one of the earliest mathematical modeling that considers weather and climate change into account shows \( R^2 \) of 0.96 and RMSE of 0.544. The recent models investigate the additional explanatory content of the weather and climate [266, 267]. [268] incorporates residual temperature, along with specific humidity, in forecasting weather-dependent warm-season electricity demand. Apart from that, A hierarchical Bayesian regression model is presented in [268, 269] to predict summer residential electricity demand across the United States.

The change of weather and climate directly impacts the variable renewable energy productions (hydropower, wind, and solar-based generation) besides that of conventional fossil fuel. A study shows that the north-west region of the United States confronts most of this challenge since Washington, Oregon, and Idaho essentially depends on renewable energy sources, particularly on hydropower. Though the blessings of immense hydropower resources assure extremely low carbon generation in these states, they experience a significant cut in their generation during drought time. Federal Columbia River Power System
operated by Bonneville Power Administration, which extends through Canada, Montana, Idaho, Washington, and Oregon, is an excellent example to study the impact of weather and climate in a hydropower-based generation. Collectively, it is about 23GW of hydropower capacity and meets 60% of the regional demand. Unlike the hydropower systems in the east coast, this system is snowmelt-dominated where most of the precipitation occurs as snow is stored throughout the winter time. Then, in the summertime, when the snow starts melting as water, and coincidentally the demand of the electricity generation gets high, it helps in gearing up the generation. In recent years, the dynamics of this hydropower-based system, such as the amount of precipitation, the amount of melted water, and the timing of water melting, has been affected by climate change.

Figure 7.13: Anomaly between a normal atmospheric condition and forecasted catastrophic condition in the dynamics of hydropower-based system.

Figure 7.13 shows the anomaly between a normal atmospheric condition and the forecasted catastrophic condition (due to climate change) in the dynamics of this hydropower-based system. As has been seen for the normal condition, a snowmelt-dominated system naturally boosts up the resource flow to generate electricity when the demand is maximum [270, 271, 272, 273]. Here, we observe when the electricity demand is relatively low in the wintertime, the flow is low as the snow precipitation is getting stored in the mountains and not
contributing in the streams. Then, in the early summer, when the snow is melting down, we experience a sharp increase in the flow. In the middle of the summer, when the electricity demand is maximum, the melt rate is maximum, we experience the peak flow. Then, again with the decrease in the melting rate, the flow gradually decreases in the fall, coincidentally with the drop in the demand. In the catastrophic condition, we expect to experience higher temperature, which will result in less snow storage and more melting water in the flow during wintertime. Consequently, we will experience higher stream-flow in the winter, resulting in more generation in the time when the demand is not high. Whereas in the early summer, it will undergo a less sharp increase in the flow, and the peak flow in the middle summer is significantly dropped down, and eventually, we will not get the necessary flow for power generation when the demand is maximum. In short, climate change will result in less precipitation falls as snow, more falls as rain (no winter storage) because higher temperatures initiate spring snowmelt earlier [274].

The weather and climate engender variability in wind energy. Research shows regional climate is crucial in terms of resource development because, in the
United States, there are some parts of the country which are significantly more wind-rich than the others. A considerable change has been observed in year-to-year wind power generation due to the climate and weather variability, resulting in difficulty to plan around. The historical wind speed distribution for a particular region can help to plan the energy generation mix of that specific region; however, inadequate historical wind data to figure out the distribution and the prospective computational complexity are the major challenges. Figure 7.14 depicts the pattern of seasonal wind variability in different regions of the United States. Here, the dotted straight line illustrates the yearly median of the wind-energy generation capacity factor for each of the geographical regions. It is evident that the upper plains (dark blue) and the lower plains (brown) have the high average capacity factor; hence they can be considered the perfect area for the wind energy generation. Besides, this figure implies how the geographic wind-generation capacity factor pattern overlaps with the electricity demand patterns. On top of that, like the localized heating and cooling requirement, the wind flow and volume follows a diurnal pattern for a particular geographical locality [275].

Most of the assessment and planning regarding solar energy systems assume that the amount of solar radiation on the Earth’s surface is more-or-less constant over the years. However, due to change in climate, along with air pollution as a factor, solar resources will inexorably experience substantial decadal changes. Several research confirms long-term changes in dimming and brightening quantity. The prospective aberrant changes in the surface solar radiation projected by the available climate models may unfavorably affect solar power production, including both PV and CSP (Concentrated Solar Power). Apart from the renewables, conventional fossil fuels—besides experiencing the inevitable impact of weather and climate change—exhibit strong seasonality in
availability and cost, resulting in relatively less expensive in the winter time and more expensive in the summertime [276].

Extreme weather condition has a severely adverse effect on fossil fuel production. Figure 7.15 illustrates the impact of extreme weather condition, in particular, Hurricanes on oil and natural gas production. It shows how the production in those regions experienced a sharp decline just after the incidents. It is indeed notable how Hurricane Frances made a significant loss in oil refinery, and Hurricane Katrina came with an unprecedented loss in the natural gas refinery. The loss due to the strike of Hurricane Dennis in the natural gas refinery is also striking. Table 7.1 summarizes the crucial events that caused power outages in the United States from 1984 to 2006 [276]. We observe significant research in effort to reduce storm-related outages in the literature. These mostly suggest tree-trimming schedules, undergrounding distribution and transmission, implementing smart grid improvements, distributed generation, reliability-centered maintenance regulations, and mutual assistance agreements to mitigate the impact of extreme weather condition.

Table 7.2 summarizes the impact of weather elements in electricity
Table 7.1: Statistically Significant Blackouts’ Cause Categories in the United States.

<table>
<thead>
<tr>
<th>Cause Category</th>
<th>Mean size in MW</th>
<th>Mean size in Customer</th>
<th>% of events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earthquake</td>
<td>1,408</td>
<td>375,900</td>
<td>0.8</td>
</tr>
<tr>
<td>Hurricane/Tropical Storm</td>
<td>1,309</td>
<td>782,695</td>
<td>4.2</td>
</tr>
<tr>
<td>Ice Storm</td>
<td>1,152</td>
<td>343,448</td>
<td>5</td>
</tr>
<tr>
<td>Wind/Rain</td>
<td>793</td>
<td>185,199</td>
<td>14.8</td>
</tr>
<tr>
<td>Other External Causes</td>
<td>710</td>
<td>246,071</td>
<td>4.8</td>
</tr>
<tr>
<td>Other Cold Weather</td>
<td>542</td>
<td>150,255</td>
<td>5.5</td>
</tr>
<tr>
<td>Operator Error</td>
<td>489</td>
<td>105,322</td>
<td>10.1</td>
</tr>
<tr>
<td>Fire</td>
<td>431</td>
<td>111,244</td>
<td>5.2</td>
</tr>
<tr>
<td>Equipment Failure</td>
<td>379</td>
<td>57,140</td>
<td>29.7</td>
</tr>
<tr>
<td>Tornado</td>
<td>367</td>
<td>115,439</td>
<td>2.8</td>
</tr>
<tr>
<td>Supply Shortage</td>
<td>341</td>
<td>138,957</td>
<td>5.3</td>
</tr>
</tbody>
</table>

Table 7.2: Impact of Weather Elements in Electricity Demand and Generation.

<table>
<thead>
<tr>
<th>Weather Elements</th>
<th>Demand</th>
<th>Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>Wind Speed</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td>Cloud</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td>Visibility</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Precipitation</td>
<td>**</td>
<td>**</td>
</tr>
</tbody>
</table>

demand and generation. Here, the number of asterisks indicates the degree of impact in the corresponding domain of concern: three asterisks symbolize profound impact, two asterisks express moderate impact, and a single asterisk implies it somehow has an impact.

One interesting challenge in power system planning, particularly in a step towards replacing baseload (coal generation) with the variable renewable energy like wind and solar, is how to combine different variable renewable energy sources in such ways so that it is possible to complement each other and reduce the inherent uncertainty comes from the renewable energy sources. Specifically, how can we place renewable energy projects to take advantage of less covariance when they are producing energy? The implementation of statistical law of large number can be helpful. Figure 7.16 depicts the scenario of wind energy generation output normalized to mean for 200 and 15 wind turbines placed in dispersed positions [276].
As can be seen, the bottom one shows more zigzag (which means more variability and uncertainty) along with the lesser mean value (though the mean value is self-evident) compared to the top one. This analysis provides an insight that the curve can even be smoother with high output mean value if we place 500 wind turbine in different places. The insight from the law of large number also applies for the solar power generation aggregation, implying the more it integrates with the solar plants located in different places, the more predictable the generation curve becomes. One immediate question, in this regard, can be how far apart solar plants need to be placed to gain the advantage in predictability from the law of large numbers because only building a number of wind turbines or solar panels (right next to each other) cannot guarantee optimal generation. Research shows we experience more covariance with the lesser distance between each plant-site. And after a certain range, we experience more-or-less constant covariance. This insight can be helpful in deciding the minimal distance to get the optimal output.
7.3.2 Load Forecasting Techniques

Load forecasting is a technique used by the electricity providing companies to predict the required energy to attain a dynamic demand-supply equilibrium. The predictive accuracy of load forecasting is of profound importance for the operational, as well as managerial loading, of a utility organization. Load forecasting, precisely peak load forecasting, is an integral and indispensable process in strategic planning and efficient operation of electric utilities. Primarily, reliability and low cost are the two significant motivation behind load forecasting since electric utility is expected to operate without having a failure in continually balancing supply and demand, and within as low as possible cost. In recent years, to mitigate the environmental challenges and promote renewable resources in the generation infrastructure, lowest possible emission is considered as one of the crucial factors in predicting both the load magnitude and geographical location of the load in a certain planning horizon. Assuring reliability is a multi-scale challenge that involves balancing between supply and demand on a second-to-second, minute-to-minute, hour-to-hour, daily, seasonal, and all the way up to years and decade. Frequency regulation is one of the most obvious reliability issues that require active dynamic management. It is known that different power systems may have different frequencies, and in the United States, all the power systems operate at 60 Hz, unlike in Europe and Asia at 50 Hz. If we do not stably balance between demand and supply, the frequency will increase or decrease: If the demand is greater than the supply, the frequency increases, and if the supply is greater than the demand, the frequency decreases. If it deviates considerably away from 60 Hz, we may have a grid-scale failure. Hence, to assure reliability, it requires to continually adjust the availability of supply to match it to demand within a
Table 7.3: Applications of Different Load Forecasting in Energy Workflow.

<table>
<thead>
<tr>
<th>Operations</th>
<th>VSTLF</th>
<th>STLF</th>
<th>MTLF</th>
<th>LTLF</th>
<th>VLTLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Producing, purchasing and selling electric power</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Transmitting and distributing electric power</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Fuel Allocation</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inrush current stabilizer</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Security assessment and analysis</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintenance scheduling</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Searching for renewable resources</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Environmental policies planning</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>System planning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Economic dispatching</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Load dispatching coordination</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Staff recruitment</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optimal generator unit commitment</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Price deciding to meet demand with fixed capacity</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Sensitivity analysis of electrical equipment</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Load flow estimations</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Scheduling construction of new generating capacity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 7.3 points out the possible applications of different forecasting based on time horizon.

certain range and keep frequency close to 60 Hz [277, 278].

From the aspect of the duration of the planning horizon, load forecasting can broadly be classified into five categories: VSTLF (very short-term load forecasting that ranges from few minutes to an hour ahead), STLF (short-term load forecasting that concentrates on hourly forecasts for one day to one week ahead), MTLF (medium-term load forecasting that ranges from few months to one year), LTLF (long-term load forecasting that includes from one year to five years), and VLTLF (very long-term load forecasting that includes ten years ahead) [279]. In short, VSTLF and STLF are mostly required for balancing operation of the grid system. Besides, they help in trading strategies for the day-ahead electricity market. MLTF is essential for planning major tests, commissioning different events, determining outage time for plants and key parts of equipments, besides the trading strategy. On the other hand, LTLF and LTLF are crucial for resource planning for the power system, and the subsequent price evaluation of the energy contracts. Table 7.3 points out the possible applications of different forecasting based on time horizon.

Long term load forecasting conveys design implication in system planning. One of the critical elements of system planning in most power systems
Figure 7.17: Trade-off in flexibility of resources, though we need flexibility in resources to address variability in demand, flexible resources are costly.

is the diversity of the resources. It requires different levels of flexibility with different types of generations so that it can meet the electricity demands that continually changes on an hour to hour, week-to-week, and seasonal basis that entails some resources which are flexible. It won’t be wise to have all the resources flexible since they are considerably costly compared to the other options. That’s why we need an optimal balance between the resources that are inexpensive but not flexible (such as coal and nuclear), and flexible but more expensive (such as natural gas and oil). Another key element in system planning is redundancy. It is a very pronounced trade-off in the infrastructural planning and engineering that increase in redundancy comes with an increase in reliability, but also increases the cost. Therefore, it is an optimization challenge to investigate up to how much redundancy it is worth investing to secure a balance between the cost and reliability. Figure 7.17 explains the scenario of flexible and not flexible resources where the horizontal-axis represents cumulative quantity supplied in MWh, and the vertical-axis means the marginal cost in $/MWh. As has been seen, oil, in general, is costly and suggested not to use until it is of absolute necessity [280, 281].
Thermal power plants can typically be categorized into three levels: baseload, shoulder load, and peaking. Shoulder power plants lie in an intermediate category that can ramp up a bit but not enough to consider it a peaking power plant. Thermal power plants can be classified in another way: by its fuel type such as coal, natural gas, nuclear, and oil. Fuel type is pertinent to the capacity of the plant. Capacity factor, the fraction of installed capacity and of getting used throughout the year, is another parameter to distinguish thermal power plants. Capacity factor ranges between one and zero, and the higher value indicates more usage throughout the year. The value close to one implies that they are online almost every hour of every day throughout the year except some days when the plants are shut off for maintenance purpose. On the flip side, the peaking power plants, such as oil and natural gas power plants, shows capacity factors close to zero, indicating seldom usage in generation-flow [282].

Addressing the reliability and redundancy trade-off mentioned before, system planners optimize it by building enough plants to cover future peak demand plus a 15% reserve margin. Reserve margin is the ratio between the difference of total available capacity and peak annual load to the peak annual load as shown in (7.2).

\[
\text{Reserve Margin} = \frac{\text{Available Capacity} - \text{Peak Annual Load}}{\text{Peak Annual Load}} \tag{7.2}
\]

The load forecasting techniques can broadly be clustered into nine categories as far as the mathematical approaches are concerned: Multiple Regression; Exponential Smoothing; Iteratively reweighted least-squares; adaptive load forecasting; stochastic time series such as Auto-Regressive (AR) model, Auto-Regressive Moving Average (ARMA) model, and Auto-Regressive Integrated Moving Average (ARIMA) model; Auto-Regressive Moving Average with eXogenous terms (ARMAX) models based on genetic algorithms; fuzzy
logic; neural networks; and knowledge-based expert systems.

Multiple regression analysis, leveraging the weighted least-squares estimation for each of the factor variables based on the statistical relationship between total load and factors’ influence, is the most common technique for load forecasting. [283, 284, 285] suggested the fundamental model for the multiple regression analysis as shown in (7.3).

\[ Y_t = v_t a_t + e_t \]  \hspace{1cm} (7.3)

Here, \( t \) is sampling time, \( Y_t \) is measured total load, \( v_t \) is vector of adapted variables, \( a_t \) is transposed vector of regression coefficients, and \( e_t \) is model error at \( t \). In multiple regression, \( v_t \) can be expanded based on the different insights regarding historical metered load, expected distributed generation, calendar effects (day of the workdays, weekends, month of the year, etc), weather data (degree days, wind speed, humidity, light intensity, etc), and economic and demographic drivers. On top of that, though linear dependency demonstrates best results in most of the cases, multiple regression offers select the polynomial degree of influence ranged from 1 to 5.

Then, exponential smoothing, one of the classical techniques used in load forecasting, with a fitting function is used for load forecasting as presented in (7.4) [284, 285].

\[ y(t) = \beta(t)^T f(t) + e_t \]  \hspace{1cm} (7.4)

Here, \( f(t) \) is fitting function vector of a process, \( \beta(t) \) is coefficient vector, and \( e_t \) is white noise. Exponential smoothing can be augmented with power spectrum analysis, as well as adaptive autoregressive modeling, to address the challenges induced by a unique pattern of energy and demand in fast-growing
regions.

Iterative reweighted least-squares—through an operator that controls each variable at a time—is used to identify the order, including parameters as well, of the model. It initiates with an optimal starting point determined by the operator, and then, uses the autocorrelation, as well as partial autocorrelation, of the resulting differenced preceding load data to identify a suboptimal model for the load dynamics. In the case of iterative reweighted least-squares techniques, the weighting function, along with the tuning constants and the weighted sum of the squared residuals, form a three-way decision variable to determine an optimal model and the subsequent parameter estimates [284, 285].

Adaptive load forecasting is one of the commonly used techniques in recent days. In this technique, to keep track of the continually changing load conditions, the model parameters are automatically corrected. Here, regression analysis is implemented based on the Kalman filter theory that incorporates the current prediction error and present weather data acquisition programs to estimate the next state vector. To determine the state vector, it not only analyses the most recent measured load and weather data, but also takes the historical data into account, and the mode of operation is facilitated switching in multiple and adaptive regression analysis [284, 285].

Though time series modeling is not a suitable forecasting approach for long term load forecasting because of the frequent unique change in demand pattern in the developed and fast-developing regions, it is one of the most popular methods in short term load forecasting. In simple, time series modeling is initially generated based on the previous data, and then, the future load is predicted based on the model.

The autoregressive model can be adapted to model load profile as follows if the load is considered as the linear combination of previous loads as presented
Here, $m$ is the order of the model, $w_k$ is random load disturbance, $a_{ik}$ are coefficients tuned from least mean squares algorithm. $\hat{L}_k$ represents predicted load at time $k$. 

In addition, the ARMA (Auto-Regressive Moving-Average) model considers the current value of the time series linearly regarding values from previous periods and previous white noise values. A $p$ and $q$ ordered ARMA model can be represented as (7.6). 

$$y(t) = \phi_1 y(t-1) + \ldots + \phi_p y(t-p) + a(t) - \theta_1 a(t-1) - \ldots - \theta_q a(t-q)$$

In addition, the ARMA model considers the current value of the time series linearly regarding values from previous periods and previous white noise values. A maximum-likelihood approach or a recursive scheme is generally used for parameter identification in ARMA model.

If the process is not stationary, it requires to transform the series to a stationary form first by a differencing operator. An ARIMA (autoregressive integrated moving average) model of order $p, q, d$ can be presented in (7.7) where the series of $p$ and $q$ ordered autoregressive and moving average component is required to be differenced $d$ times.

$$\phi(B) \nabla^d y(t) = \theta(B) a(t)$$

ARIMA model, using the trend component, is deployed to forecast the
growth of the system load.

Apart from the time-series-based short-term forecasting, ARMAX, leveraging genetic algorithm, is a popular technique for the long term load demand forecast. Through simulating the natural evolutionary process, it allows the ability to converge towards the global extremum of a complex error surface [284, 285].

Apart from that, leveraging the idea that the fuzzy logic system with a centroid defuzzification can successfully identify and sufficiently approximate an unknown dynamic system on the compact set to arbitrary accuracy, fuzzy logic can be implemented in the case of load forecasting. The fuzzy logic-based forecasting method follows two stages: Training, and then, On-line Forecasting. In its training stage, a $2m$-input and $2n$-output fuzzy-logic based forecaster are trained using the metered historical load data to generate patterns database and fuzzy rule base patterns database and a fuzzy rule base from first-order and second-order differences of the data. After the training stage, it will be connected with a controller to forecast the load change online. An output pattern is generated through a centroid defuzzifier if it attains a most probably matching pattern with the highest possibility [285].

Neural networks, such as multilayer perceptron network and self-organizing network, have a strong potential to overcome the sole reliance on a functional form of the predictive model. It makes the neural network-based forecasting a very active area of research. It facilitates improving the accuracy of load forecasting by neural networks integrated with several other techniques such as stochastic time series methods, weighted least squares procedure, a combination of fuzzy logic and expert systems, etc. [286] grouped the commonly used load forecasting techniques based on the duration of the planning horizon.

Different forecasting horizons, such as STLF, MTLF, and LTLF, have
different challenges in forecasting. The peril of long term load forecasting is profound since uncertainty is rampant regarding climate, technology, population growth, and economic conditions. Overestimating demand might seem like the prudent modeling choice from the reliability aspect, but it can be costly, and hence unwise. In addition, it has been observed the usage behavior differs between the consumers using different types of meters, in particular between the consumers using smart and traditional meters along with different tariffs. The utility must take this into account and develop separate forecasting model for each of the metering systems and then plug-in them up for the final forecast value. Otherwise, they may come up with an inaccurate forecasting. In the case of STLF and MTLF, it is sometimes overly complicated to precisely fit various complex factors affecting demands for electricity into the forecasting models. In addition, it may not be easy to obtain an accurate demand forecast based on parameters such as change in temperature, humidity, and other factors that influence consumption. The utility may suffer losses if they do not understand and decide on an acceptable margin of error in load forecasting [287, 288].

7.4 Penetration of Renewable Energy

7.4.1 Variability of Renewables

Electricity demand frequently fluctuates throughout the day, week, and year. Albeit having noise and uncertainty, how we are going to use electricity uniquely shows a tremendous amount of predictability. This predictability lets the generation planning and integration in a prudent manner, such as meeting the baseload with not-flexible and not-able-to-ramp-up resources, intermediate and peak load with must-take (like renewables), flexible and able-to-ramp-up resources. With the recent surge of the renewable energy-based generation in generation mix of the United States, variability in the renewable resources,
particularly solar and wind, has been developing into a critical challenge in generation planning and integration. In this section, the discussion will be limited to the variability of solar and wind for two reasons [289]. First, apart from the fact that wind and solar are emission-free, compared to rest of other types of resources wind and solar are not dispatchable and controllable. So, we cannot consider them as baseload: turn them on and leave them on, and they cannot provide a steady amount of electricity. We cannot consider them as peaking resources: leaving them off most of the time and only turn them on when the electricity demand is highest. We consider them as must-take resources: when they are available, we will use them, when they are not, we won’t use them. So, from the grid operators’ point of view, wind and solar are considered for demand reduction since they have negligible operational costs, their capital cost is all about building those projects. Second, as has been noticed from the yearly US per capita consumption in kWh by renewable sources from 1999-2016 with trendlines and forecast depicted in Figure 7.18, only solar and wind exhibit incrementing trends with an exciting prospect to be the renewables of choice in generation planning [290].

![Figure 7.18: Yearly US per capita consumption in kWh by renewable sources from 1999-2016 with trendlines and five year-forecast.](image)
When we model the solar PV production, we take the following variables into the account: size of the panel array, solar insolation (determined by hour of day, day of year, latitude, aspect, and tilt), efficiency (conversion of solar energy to electricity), and performance losses (temperature and inverter). Among these variables, solar insolation is not dispatchable, and hence a key driver in solar PV production. It has an immediate impact on the amount of generation from a particular project and distinctively varies throughout the United States. For example, the further south and southeast we go, the higher the availability of solar insolation is reported. Therefore, it is intuitive if we assume the solar installation cost in Wisconsin or New York is as same as that of Arizona and Florida, we would prefer to install the solar panels in the region from where we can get the most energy out. However, it is not the only parameter to analyze its financial viability and economic competitiveness [291, 292].

On a different perspective from the solar insolation, solar irradiation, in particular, is very predictable, at least theoretically. We—more or less—can perfectly model the solar irradiation as it changes throughout the day. However, cloud dynamics regards as the most stochastic element of solar power production, and in several instances, adds a tremendous amount of uncertainty in the case of incorporating large amounts of solar into power systems. Figure

Figure 7.19: The effect of sky conditions on solar panel power output.
7.19 portrays the impact of sky condition, in particular, cloud, on the solar generation. It is evident how the heavy dark clouds add noise in the electricity generation streams. Consideration of the cloud factor in generation makes the forecasting a way more stochastic and uncertain process, and in practice, exhibits a significant difference between the day-ahead-forecast, hour-ahead-forecast, and the actual generation. Even more, it is critical to predicting solar power even just an hour before its generation because of the cloud factor albeit efficient prediction of solar irradiation [293].

Now, delving into the wind-based power, power generation depends on three cardinal variables: the amount of air (volume), the speed of air (velocity), and the mass of air (density) flowing through the area of interest (flux). The generated power from the wind turbine follows the equation as presented in (7.8):

\[
P = \frac{1}{2} \rho \cdot A \cdot v^3
\]

From this equation, it is apparent that power production from the wind turbine is very sensitive to the wind speed or the velocity \(v\), and algorithmically, if the wind velocity increases by a small amount, power generation goes up by the function of this cubic relationship. However, according to the Betz limit, the power coefficient is the quotient of the power extracted by the turbine to the total energy contained in the wind resource [294]. This coefficient helps us to estimate the generated wind power in the real case scenario. Betz limit is of the maximal possible \(C_p = 16/27\) which indicates 59% of efficiency for the conventional wind turbine in extracting power from the wind. Since it is identified from the equation that wind velocity is the most impactful parameter in wind power generation dynamics, Figure 7.20 delineates the wind velocity distribution over the United States.

Figure 7.20 implies where installing wind turbines is more advantageous.
Since, like other renewable energy resources, wind energy-based power generation only requires the capital cost for installation—requiring no cost for fuel—and we can presume the installation cost of wind turbine is—more or less—same across the entire United States, we can infer from this figure that US midwest, west-offshore, and east-offshore are the most convenient places to install the wind power generators [295, 296]. However, there are some additional confounding variables. First, financial viability of the project can be crucial since there is a notable population dearth in the US midwest region and it necessitates significant added cost for high voltage transmission to deliver the wind power to load centers or to the communities of high electricity consumption. Second, as can be seen from the year-to-year changes in wind power production for a single wind project depicted in Figure 7.21, the more uncertain variability and less predictable generation compared to solar energy are imperative in this dynamics. Though the generation variability in the case of a single plant is considerable, this evidence is not sufficient to have a conclusive idea regarding the impact of unpredictable variability and integrating the portfolio of several wind plants may...
come up with a different insight. On top of that, it may seem more uncertain in a smaller time scale, and considering a larger time scale may address a different view on this point.

Third, the variation in wind power generation on an hourly basis makes it confounding to incorporating into electric power systems. In particular, with the rising share of wind energy in the United States’ electricity generation mix, having the possibility that a big chunk of it goes away unexpectedly during the day is a significant concern, and it is intricate for utilities since it impacts the maintenance of the power system operations in more than one way.

Next, to understand the grid integration challenges in variable renewables, Figure 7.22 – besides introducing the concept of duck diagram – illustrates the hourly distribution of the net demand with increasing PV penetration considering overall demand remains unchanged [297, 298]. As can be observed from Figure 7.22 and 7.23, with the increase of PV penetration, the non-PV supported portion of the net demand curve gets dropped down (consider the drastic drop down in case of 58% penetration) from 9 am to 5 pm when the sun is available. It- putting the net load in the context of increasing PV penetration- implies a trend that anticipates two major issues: over-generation risk and ramp requirement.
As can be interpreted from Figure 7.22, it requires a moderate amount of generation online to meet demand in the early morning and the late afternoon (when the sun goes down), and in the middle part of the day, it does not need much because of having greater PV penetration. In many cases, it has been observed it may be less costly to leave generation on around 9 am compared to achieve high ramp up, and kind of waste it throughout the day in order to have that generation online later on the day. This problem is called overgeneration.
Since we have been experiencing greater penetration of PV (in general, renewables) in the conventional systems, the risk of overgeneration becomes greater, makes physical issues of safety and reliability. Another trade-off of having greater renewable-penetration is if significant changes in wind and solar availability take place very quickly—without warning—that can pose a challenge to system reliability. In the case of operations, it can be minute-to-minute, hour-to-hour, and day-to-day. For minute-to-minute, frequency regulation is needed, since it requires to maintain 60 Hz of frequency for AC system in the United States, and undersupply of generation can cause that frequency to deviate from 60 Hz, and if it goes too far, then it may experience a significant instability on the grid. It has to be actively managed through automatic generation control at generators. To address the hour to hour variability, load following and reserves are crucial. It requires certain power plants to increase or decrease their production to follow the net electricity demand patterns. Reserves are online sometimes, and offline sometimes, they are not primarily producing electricity, but they can quickly ramp up and produce electricity at the right frequency in order to account for any unexpected change in the availability of wind and solar. For day-to-day, unit commitment is critical to make decision to turn a plant on and off. It is crucial as it is exorbitantly costly to turn a plant on and off.

In the case of planning, it is about a year-to-year basis: capacity planning based on the pick load forecasting, considering the likelihood that given uncertainty in electricity demand because of the weather and given the uncertainty in renewable energy production on year-to-year basis, and the reserve margin being below a certain point of inability to meet electricity demand. Another interesting point to discuss is that there has been observed a steady downward trend in wind speed globally over the last fifty years. The trade-off in considering the historical data over the recent trend in generation forecasting can
pose a critical challenge in planning because of the third order relation of wind speed in wind energy generation, and eventually, end in a serious failure [299].

7.4.2 Effect of Commodity Prices

Commodity prices are one of the key drivers in the dynamics of United States Internet of Energy, and hence, is imperative to be discussed explicitly in an individual section. Previously, energy commodities were essentially conceptualized as including natural gas, petroleum products, and coal. With the recent surge of renewable-based generation, the raw materials used in the fabrication (along with cutting, bending, and assembling) of renewable energy and storage technologies are considered as energy commodities. In this section, we will briefly discuss the factors that influence commodity prices and how these propagate to electricity prices. The direct impact of the commodity market on electricity prices is observed in the fossil fuel power plant that ultimately gets incorporated into the wholesale prices, and eventually, retail prices for customers. Unlike this, renewable energy-based generation is considered as immune to year-to-year changes in fuel cost. The reason is though it is uncertain how much energy we will get from the renewable energy plant, such as solar and wind, we know exactly how much we will cost for it. However, it is required to factor in the availability of the renewable resource across the year, since it ends up impacting in the levelized cost of electricity.

Nevertheless, commodity prices do matter—albeit in an indirect fashion—for renewable energy, since the majority cost (compared to the fixed operation and maintenance cost) of the renewable energy is drawn from compensating the annualized capital cost. If commodity prices fluctuate the capital cost of renewable energy projects, this capital cost aggregate into the cost of renewable energy over the entire project lifetime. So, if the solar plant or wind turbine is
developed in a year when steel and copper prices are high, the long term electricity selling price to adjust this cost will be significantly high; and there is no chance for the commodity prices to go back and lower the price of electricity from solar or wind firm. For example, Figure 7.24 depicts the instances of how the copper and steel price reflects the levelized Power Purchase Agreement (PPA) from 1990 to 2010 across the different regions in the United States. As can be noticed from this figure, the price of copper and steel experienced a significant increase in 2006-2008, so the different regions in the United States did in their levelized PPA. Typically, these agreements are of twenty-five to thirty years which implies the plant developed in 2006-2008 reflects into the higher price of electricity for the next thirty years, not just in the year it was developed.

The United States Critical Materials Institute (CMI), an entity associated with the DOE, concentrates on technologies that make better use of materials indispensable for the United States’ competitiveness in clean energy; and identify and eliminate the demand for materials that are crucial to supply disruptions. They have four principal objectives. First, diversifying supplies: if on geographical source goes offline or out-of-function, a different source can take its place in operation. Second, developing substitute materials that can functionally
serve the same purpose compared to the materials currently used. Third, using the available materials more efficiently by reducing waste and adopting recycle in manufacturing. Finally, last but not least, forecasting which materials might become critical in the future. Table 7.4 reports the CMI’s investigation on materials used in clean energy technologies and components. It incorporates the materials including rare earth elements, and their applications in photovoltaic films, wind turbines, vehicles, and lighting. Red rows in Table 7.4 indicate the rare earth elements.

Besides, the CMI classified the materials used in clean energy (mostly in photovoltaic cell and energy storage systems) into three categories based on two evaluation metrics: supply risk and how important it is in clean energy. Both are considered on a scale of 4, indicating low as 1 and high as 4. The materials which are of high (4) or high-medium (3) in both metrics are identified as critical materials, the materials which are of high-medium (3) or medium (2) in both metrics are considered as near-critical materials, and the materials which are of medium (2) or low (1) in both metrics are studied as not critical materials. The third parameter is essentially the time-frame of the supply availability that reflects on the categorization, and eventually, necessitates forecasting of resource
availability. Figure 7.25 illustrates the criticality matrix of materials used in clean energy for the short-term and medium-term.

![Criticality Matrix Diagram](image)

Figure 7.25: Criticality matrix of materials used in clean energy for short-term and medium-term.

Unlike the materials used in clean energy, commodity prices have a direct short-term influence in the case of conventional fossil fuel-based generation. Similarly, in the case of conventional fossil fuel-based generation, commodity prices vary with a number of reasons such as energy crisis, natural calamities, inexplicable tracking, global financial crisis, polar vortex, and excess supply from fracking.

### 7.5 Understanding Energy Consumption Dynamics

#### 7.5.1 Energy Disaggregation

To meet the environmental challenges and continually depleting energy resource dilemma, energy demand reduction, along with improving energy efficiency, is considered as the safest and most sustainable approach. It has been reported that the residential sector occupies approximately 22% of total energy in the United States which reflects in 37.8% of total electricity consumption in
the US (electricity consumption by different sectors and household-electricity consumption distribution of the United States depicted in Figure 7.26).

Consequently, household energy usage shares about 38% of the total yearly carbon emissions in the US. Research shows approximately 27% of the current households’ energy, so as the electricity, can be saved through efficient demand-side energy management. Household-demand-side management majorly concentrates on six objectives, namely, peak clipping, load shifting, valley filling, strategic load growth, strategic conservation, and flexible load shape [300]. These require to classify the factors that affect household energy usage into different categories, such as demographics and socio-economics, location, temperature, energy prices, and building characteristics, and eventually, understand the household energy consumption behavior. Electricity consumption patterns of different users in different time granularity, which is affected by both objective and subjective factors, can be discovered through effective analysis of electricity consumption data accumulated by different data acquisition terminals, such as smart meters. Therefore, energy disaggregation is essentially being an integral part of the Advanced Metering Infrastructure (AMI) in this effort [301, 302].

![Figure 7.26: Electricity consumption by different sectors and household electricity consumption distribution in the United States.](image-url)
The benefits of energy disaggregation are manifold. It ranges from raising awareness regarding energy usage to empower consumers across different dimensions in making better decisions, offering sophisticated options for automated commissioning, diagnosis, and fault detection of residential buildings to providing simplified and improved load studies leading to the identification of specific end-use equipment and facilities. Thus, it encourages considerably more efficient, relatively cost-effective, and comprehensive quality assurance programs in order to achieve substantial savings from energy efficiency measures and demand response [300, 303].

![Iterative process for developing models to understand energy usage behavior](image)

In simple, energy disaggregation can be defined as an approach that allows taking a whole building (aggregated) energy signal into consideration, and then classifies it into appliance-specific data, such as a plug or end-usage data, by a set of IOT-based computational techniques. It is an effort motivated to delve into understanding energy usage behavior and modeling. In general, energy modeling involves iterative approaches for finding variables and parameters using more nuanced information and features as depicted in Figure 7.27 which
eventually minimize the model error. It necessarily starts out with an extensive set of training data. Then, the training data set is employed to come up with models for energy consumption for individual activity based on a number of features across different dimensions. After that, it gradually eliminates any kind of statistically insignificant variable. After a certain iteration, a model is finalized which is as accurate and, at the same time, as parsimonious as possible [304].

Different energy models of different dimensions unpack different energy usage behaviors. Among them, behavior that incorporates different time granularity and sectors are regarded as crucial for knowledge extraction for resource planners. For example, the consumption patterns of different sectors, such as industrial, residential, and commercial, are illustrated in Figure 7.28 for both monthly and sub-daily basis. It is evident from Figure 7.28 that the consumption patterns of the various sectors are strikingly different, and the residential electricity consumption is the critical driver, as well as the most reactive sector with changes of weather and climate, in total demand [305].

From this stage, it requires special techniques to acquire insights on the household level, helping individual consumers make a smart decision about their electricity consumptions based on multiple parameters, such as price and
Figure 7.29: Almost real-time household level total power provided by the advanced meter and the desired disaggregated power.

availability of renewable energies, and therefore AMI is deployed into operation. It collects information about electricity consumption at the household level on a minute-to-minute basis, and then, transmits this information back to the central console system, facilitating two-way communication and almost real-time sampling. However, the information that comes from the advanced meter is not apparently comprehensive and it requires advanced analytics to leverage the advantages of this information. Figure 7.29 explains what we receive from the advanced meter (total power) and what we desire to know for the smart decision (disaggregated power) [306].

The initial approach to obtaining disaggregated power was sub-metering, installing the separate individual smart plug in each major appliance in residents. It worked and met the fundamental objectives that we want, however, the cost for integrating a number of smart plugs in each house and implementing it in residential level in the entire United States challenges the overall purpose of efficiency and cost-effectiveness. Table 7.5 compares the hardware-based and software-based disaggregation techniques from the consumer-level costs, installation effort, and adoption aspects. Table 7.5 lets us conclude that the smart meter can be the most efficient and cost-effective option if advanced analytics can be incorporated to obtain appliance-level information [306, 307].
Table 7.5: Comparison between Hardware-based and Software-based Disaggregation Techniques.

<table>
<thead>
<tr>
<th>Sensing Technology</th>
<th>Hardware Disaggregation</th>
<th>Software Disaggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hardware Monitors</td>
<td>Smart Appliances</td>
</tr>
<tr>
<td>Plug Level</td>
<td>30 – 50/plug</td>
<td>$100+ additional for other non-smart appliances</td>
</tr>
<tr>
<td></td>
<td>300 – 600/home</td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td></td>
<td>House Level Current Sensor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$200+/home</td>
</tr>
<tr>
<td>Installation Effort</td>
<td>Most plugs– Moderate</td>
<td>Easy</td>
</tr>
<tr>
<td></td>
<td>240V plugs- Hard</td>
<td></td>
</tr>
<tr>
<td>Adoption</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Smart Meter</td>
</tr>
<tr>
<td></td>
<td></td>
<td>None</td>
</tr>
</tbody>
</table>

NILM (Non-Intrusive Load Monitoring) or NIALM (Non-Intrusive Appliance Load Monitoring) is an analytic approach employed to disaggregate the building loads primarily based on a single metering point. This advanced load monitoring and disaggregation technique have the potential to come up with an alternative solution to high-priced sub-metering and facilitate innovative approaches for energy conservation, energy efficiency, and demand response.

From the functional point of view, NILM can be explained as a three sequential operation, namely, signal acquisition, feature extraction, and finally appliance classification. The state-of-the-art NILM and NIALM techniques for energy disaggregation are briefly discussed below. [303] proposed a cluster splitting approach to disaggregate the overlapping home appliances’ consumptions. It addresses the challenges in disaggregating energy consumption by each of the appliances when several home-appliances have power consumption-levels that overlap (loosely or tightly) with each other. This approach initiates with analyzing the cohesion between devices’ clusters to determine whether a cluster is required to be split into two or multiple clusters. This proposed technique—using REDD public data sets—was tested on overlapping devices’ clusters of six residences, and it was evident from results that the degree of overlapping in devices’ clusters and the sizes of individual clusters are crucial in its performance.

After that, for energy disaggregation, committee decision mechanisms (CDMs) have been introduced by [304] to disaggregate load at the metering
level. Their investigation shows load signatures inherently embedded in the
patterns of typical electricity consumptions are able to provide critical
information about the characteristics of the appliances as well as their usage
patterns. Multiple evidence bolstered that all CDMs- through Monte Carlo
simulations- outperform any single-feature and single-algorithm-based energy
disaggregation methods, and are considerably less sensitive to any load dynamics
and noise. They reported some case studies using this technology in appliance
usage tracking and energy consumption estimation. In [308], Misbah et al.
proposed sparse optimization for end-use disaggregation, a novel nonintrusive
appliance load monitoring (NIALM) algorithm, that can characterize the
appliance power consumption profiles accurately over time. The primary
assumption of this algorithm is that power consumption profiles of the unknown
appliances are piecewise constant over time, and it leverages the knowledge on
the time-of-day probability in which a particular device might be used. Here, it
formulates the energy disaggregation problem as a least-square-error
minimization problem, including an additional penalty term to enforce the
disaggregate signals to be piecewise constant over time. Testing this algorithm
on the household electricity data is reported in [308] with satisfactory accuracy.

Next, in [309], the authors proposed a dictionary learning-based approach
in addressing energy disaggregation problem. This technique is usually a
synthesis formulation, involving in learning a dictionary for each device and then
applying the learned dictionaries as evidence for the blind-source separation
during energy disaggregation. It facilitates disaggregation as drastically reduces
the sensing cost. In [310], Singh et al. presented a distributed and scalable
method for semi-intrusive large-scale appliance load monitoring. They--with
sufficient conditions considered for unambiguous state recovery-- incorporate an
SSER model (sparse switching event recovering) for retrieving appliances’ states
from the aggregated load data stream. This approach demonstrates satisfactory
results in improving the accuracy of load disaggregation for large-scale
appliances with a small number of meters. Then, in [311], Xia et al. proposed a
depth dilated convolution residual network- based non-intrusive sequence to
sequence energy disaggregation approach in an effort to reduce the network
optimization intricacy and explain the vanishing gradient problem. They,
initially, normalized the primary data, and then, applied the sliding window to
formulate the input for the residual network. Here, they met the challenges of
learning long time series data by increasing receptive fields and capturing further
data through the dilated convolution. Several case studies bolstered the
improved efficacy of this proposed deep dilated convolution residual network-
based sequence to sequence disaggregation method in energy disaggregation.

On a different note, [312] presented a GSP-based approach (graph signal
processing) to disaggregate the entire energy consumption down to individual
appliances’ level. The authors addressed the complexity of general graph-based
methods associated with large training overhead employing event-based graph
approach. This paper showed two approaches leveraging the piecewise
smoothness of the power load signal. The first one searches for a smooth graph
signal under known label constraints following the principle of total graph
variation minimization under some known label constraints. The second one
initiates with the total graph variation minimizer and delves into further
refinement through simulated annealing. The paper reported a competitive
performance using the proposed approach compared to the decision tree and
hidden Markov model-based approaches. After that, considering the fact that
the aggregated or smart meter signal can be expressed as a linear combination of
the basis vectors in a framework for matrix factorization, Alireza et al. presented
a technique to disaggregate energy data using non-negative matrix factorization
with sum-to-k constraint [313]. This technique—through imposing non-negative constraint as well as sum-to-k constraint—can extract perceptually meaningful sources efficiently from the complex mixtures. They compared its performance with the state-of-the-art decomposition-based disaggregation algorithms and reported superior results. In general, all the state-of-the-art nonintrusive energy disaggregation techniques can be broadly classified into two categories: optimization-based approaches and machine learning-based approaches.

### 7.5.2 Generation Expansion and Decision Analysis

Capacity expansion, in broad generation expansion, is an indispensable part of the infrastructural planning of the power industry, and subsequently, the internet of energy; and hence to be highlighted in this sub-section. In simple, capacity expansion is the process adopted by the utilities to increase their capacity of the generating-resources gradually to meet either of the following objectives: primarily, meeting electricity demand growth, then, making replacement of the existing generation that comes offline or retires because of aging infrastructure, and confronting relatively more stringent circumstances or regulations. In other words, mostly from the aspect of the electric power industry, it is the process of adding additional facilities of a similar type over time in order to meet the rising demand. Capacity expansion is a multi-faceted decision that concerns the timing, scale, and location of the major projects in the face of uncertain—often with the considerable unpredictability—demand forecasts, costs, and completion times [314, 315, 316]. The simple pictorial depiction of capacity expansion is shown in Figure 7.30.

In literature, it has been documented that the highly unpredictable uncertainty often resulted in surprise or shock to the system planners either as a critical shortage or provision of gross amounts of unwanted capacity. Both of
Figure 7.30: Capacity expansion over time.

them are highly undesirable. Figure 7.31 illustrates the impact of the critical shortage (building not enough capacity) challenges in the case of the Pacific Northwest of the United States. It shows the causal relation of unpredictable growth of electricity demand, having not enough capacity, and market deregulation to market manipulation; and shows how market manipulation and hot summer and drought can lead to increase in natural gas prices [317].

Figure 7.31: Impact of the critical shortage challenges in the case of the Pacific Northwest of the United States.

A popular example of the impact of building too much capacity is what
Figure 7.32: Impact of overestimating electricity demand resulted in a higher price. Happened in 1970 in the United States. Since the end of World War-II till 1970, there observed an unprecedented industrial growth, and hence an exponential increase in electricity demand. This phenomenon inspired the utilities overbuilt the nuclear power capacity to meet the exponentially growing demand. After 1970, a change was observed in the pattern of demand growth; it started showing an almost linear behavior in contrast to the previous exponential behavior. When demand started increasing in a considerably slower fashion, the annualized capital costs associated with these plants had to be spread over fewer individual units of electricity, and eventually, retail prices increased. Figure 7.32 is the illustration of the impact of overestimating electricity demand resulted in a higher price. This overinvestment-underinvestment dilemma posits that If it is
an overinvest, then it just sets to have higher electricity prices, if it is an underinvest, it sets up for the actual physical failure of the grid [317].

Apart from the demand, there is another source of uncertainty in analyzing capacity expansion: Market, or in particular, fuel prices. Fluctuations in fuel prices can have a tremendous impact on which technologies are more preferable. Sometimes, the change in fuel prices may experience a behavior which is not predictable, or even statistically not well-characterized, before making decisions regarding capacity expansion. The other considerable sources of uncertainties are technologies, regulations, construction time, and retirement. The technological and industrial innovations are correlated to the price projections about future capacity costs. In particular, renewable energy technologies can be a crucial driver in predicting future capacity cost; and its predictability in generation mix over decades can change the capacity cost dynamics dramatically. After that, the uncertainties come with regulations are beyond the scope of describing it in a statistical manner. For instance, it is near to impossible describing the likelihood of the United States enacting a carbon tax or some other legislation that eventually makes the fossil fuel power plants more expensive. Construction time is also a crucial factor since the decision regarding plant construction needs to be made many years before they are actually built. Retirement, though studied as a factor of uncertainties, can–more or less–be planned. These retirements of existing capacity ultimately add to the need for new capacity. Besides, in the case of capacity expansion, it is incumbent to consider the scenarios, such as if the growth in electricity demand is considerably lower than expected, if the future electricity demand is related to the cost of solar and batteries, if the technological innovation happens faster than that was expected, and most importantly, if the US transportation is shifted to be electrified [318, 319, 320]. Addressing these all sorts of possibilities
is a pressing concern of infrastructural planning for the power industry, and with the time being, the electric power industry is continually being linked to irreducible and unquantifiable uncertainty. Ensemble prediction [320, 321] followed by the decision analysis is the most frequently used academic approach to address these challenges [321, 322]. Decision Analysis is a formal structure for decision making under uncertainty that includes numerous methods for adequately identifying, clearly representing, and precisely assessing the essential aspects of a decision, and for suggesting a course of action by applying the maximum expected value axiom [322, 323]. A decision tree is a commonly used tool in the decision analysis that involves decision nodes, chance nodes, and end nodes to interpret the flow of time, decisions, uncertainty, and consequences to come up with the evaluation measures realizing how well the objectives are achieved in the final outcome [323].

7.6 Internet-of-Energy

In the last section, we intend to capture the most crucial areas centered around the concept of Internet-of-Energy. In a nutshell, we discuss the Grid Overview of the United States; Weather and Climate and its impact on the entire energy generation and consumption dynamics; Peak Load Forecasting and its techniques and burgeoning challenges; Variable Renewable Energy, its reliability challenges and how we can take advantage of this variability; Commodity Prices and its criticality; Energy Disaggregation and its impact on consumption awareness; and Generation Expansion and Decision Analysis and trade-offs.

Internet-of-Energy, as well as IoT, preserves the essence of sustainability–coordinated development of life and its habitat, society, culture, work, and material production environment, well-reflected by the social-economic-natural complex ecosystem theory. Though the conceptualization of the Internet of
Figure 7.33: Essential layers of IoT deployment with the smart grid in the realization of the internet-of-energy.

Energy is centered around the motivation of assuring electric mobility and full deployment of the must-take-resources, such as renewable sources. Internet of Energy can answer numerous energy and reliability challenges, and provide solutions and theoretical underpinnings leveraging the recent advancements in microsystems, nanoelectronics, embedded systems, control, communications, algorithms and analytics, software, and the internet technology. In the Internet-of-Energy, the area for IoT realization can be manifold. From the aspect of energy delivery and peak demand, IoT realization is incumbent for online generation monitoring, smart meter reading, and advanced control system for transmission and distribution. From commercial, industrial, and residential point-of-view, demand response modeling, electric vehicle charging, and home energy management are crucial for IoT effectuation. Besides that, utilities or consumers are one of the key sectors to be realized using IoT. Microgeneration and asset management are crucial in this regard. Figure 7.33 captures the essential layers of IoT deployment with the smart grid in the realization of the internet-of-energy. There may have multiple avenues in IoT deployment yet to
be explored to enact smooth and effective communication between the smart meters attached at the consumers’ place and the sensors [239, 228, 324, 325].

There are four key functionalities of IoE: Motivating consumers, self-healing, improve power quality and resist attack. IoE offers interactive options in transferring consumption and reliability information between the user and utility, and thereby, motivates users to plan their cost and select suitable tariff, creates awareness regarding demand response features and their impact on reliability and cost, and eventually, lets the consumers control their power usage more effectually. With the capacity to analyze on the fly, IoE can identify and react to the major faults swiftly and in a more intelligent way. In particular, smart metering approaches with wireless connectivity facilitate identifying black-outs immediately and in a nonintrusive manner. Next, IoE promotes improving power quality. The major consumer demands in all the commercial, industrial, and residential sectors are of constant voltage, and abrupt fluctuations in the voltage may be detrimental to electric appliances. IoE has tremendous potential to maintain constant voltage, thereby reducing commercial productivity loss. Apart from that, IoE adopted numerous privacy preservation methods for smart grids to protect itself from cyber and physical attacks [238, 326, 327].

The technology synthesis allows perceptive technology, advanced analytics and machine learning, advanced network technology, artificial intelligence and automatics to be employed together into machine-to-machine and human-to-machine interactive systems to realize the functional interconnection of humans and objects. It motivates the internet-of-energy to leverage the elements and functionalities of IoT, such as flexible structure, autonomous process, multi-role participants, scalability, event sharing, interconnectivity, and semantic sharing. Besides that, third parties are welcome to develop complex and compound applications with the provision of APIs. Figure 7.34 illustrates
the concepts of the internet-of-energy integration— a framework realized by the approach of IoT paradigm with the smartgrid [238, 328, 325].

7.6.1 IoE Architecture and IoT Integration

Internet-of-Energy architecture is dynamic and progressive, as such with respect to time factor, the system elements can be reconfigured. However, the myriad number of devices, functionalities, and technologies in IoT, and consequently, in the internet-of-energy, makes interoperability a crucial issue. Thereby, data deluge (by smart metering), extensibility, and scalability should be taken into consideration, resulting in enormous computational tasks. Parallel computing may obtain a significant speedup and get the analyses and results
faster. However, extrapolating the performance from the small size of the problem on small systems to the larger size of the problem on larger configurations is a primary concern. For a given problem size, computational overhead increases with the increase of the number of processing elements. Hence, the overall efficiency of the parallel program goes down in a meaningful manner. Besides that, according to Amdahl’s law, speedup tends to saturate with the increase in the number of processing elements. On the other hand, since the total overhead function is necessarily a function of both of the number of processing elements and the size of the problem, in many cases, we observe the overhead grows sublinearly with the increase of the problem size. If we keep the number of processing elements constant for such cases, the efficiency will increase with the increase of the problem size. Leveraging this insight, we can simultaneously increase the number of processing elements and the problem size at a particular rate to keep the efficiency of the system constant. Such a system is called the scalable parallel system, and assuring scalability of a system is a critical challenge in large scale IoE deployment. Another major concern, which may lead to severe repercussions, in this technology is privacy and security. The security and privacy threats are even more serious in the case of smart meters in residential buildings. The privacy concern with residential users are easily susceptible to the hackers, and sometimes, to other consumers intending their per day energy consumption reduction. These challenges and concerns come up with future research opportunities regarding suitable remedial measures, such as encryption methods, authentication schemes, public key infrastructure, and standardized application program interfaces [238, 324, 329].

The principal features of internet-of-energy is acquainted as follows in the lens of advantages and disadvantages. To begin with, automation realizes the control of numerous smart devices, leading to the uniformity of tasks. This
Figure 7.35: benefits of adopting internet-of-energy from the functional aspect

secures a transparent process over the entire machine to machine communication. Then, the efficiency of the system can be perceived in two aspects: the ratio of useful output energy and total input energy, and the opportunities it creates to retarget human efforts in other fields. The internet-of-energy facilitates more machine to machine interaction; the more the interactions between machines, the more the opportunities are created to target on other jobs that require human efforts. Besides, advanced analytics help optimizing the efficiency (the first aspect) of the energy production and management ecosystem. It also brings cost-effectiveness as another advantageous aspect of IoE. Again, communication is crucial to improve the quality and time factor; internet-of-energy facilitates a platform for daily basis communication with the devices. Implementation of IoE may facilitate instant data access (with proper authentication and user verification), which further helps the research community to conduct exploratory research in this domain and delivers useful data-driven insights. Figure 7.35 depicts the benefits of the internet-of-energy from the functional aspect [330, 238, 329, 331].

One of the major disadvantages of internet-of-energy deployment is
paramount privacy and security concerns. The more the appliances and services are dynamically connected, the more the information stored are readily available, the more the risks of the data-security breach as the information may get susceptible to hackers and unauthorized concerns. It brings a surge in multidisciplinary research opportunities regarding more robust data authentication tools, privacy policies and standards, and firmware standards. Again, due to the lack of sufficient international compatibility standards available for internet-of-energy, it is tricky and confusing both for the manufactures and stakeholders to interact with the services; thereby, compatibility is a significant concern in the massive deployment. In this regard, new standards with common protocols are being developed for residential, commercial, and industrial sectors. Next, as far as the complexity is concerned, an extremely large network is connected across in the IoE; a small failure in the software and hardware components may lead to a damage in the entire system. On the flip side, the immediate failures at the junction of nodes can be addressed through a common control center; remedial action is next to instant [238, 332, 333].

7.6.2 Broader Impact

Adopting the internet-of-energy comes up with a tremendous social impact as it steps forward into the future energy ecosystem with smart technologies and new regulatory structures and services. First, it changes the classical perception regarding generation, transmission, and consumption to both the consumers and utilities. From the consumers' aspect, this contemporary avenue is critical for ecological awareness and convoluted for energy management, underscoring their daily comfort behaviors as a dynamic factor in the complex system. Furthermore, the intricacy involved in adopting and controlling different smart devices with numerous distinct sensors governed by
different operating systems leads to interoperability concerns for the consumers, in particular, the senior citizens and the people from a non-technical background. As technology progresses, the internet-of-energy– to enhance the acceptability of these new technologies– requires training as well as mentoring opportunities for a diverse group of consumers and operators. This new technology opens up scopes for researchers around the world to study and understand the concept properly. The results and insights generated on this new technology need to be widely disseminated through publications, professional presentations, and online access to raise awareness and motivate advancement. These collective efforts will– soon– change the understanding of consumer devices, from a black box to a source of multivariate information based on the pricing scheme [238, 334, 335].

On the bright side, the recent technology transformation makes the world propelling at a rapid and exponential change, creating a tremendous impetus, as well as brunt, at different avenues and courses of action, such as professional and personal aspects. Again, embodying the consumer in the intersection of multiple domains centered around embarking fourth industrial revolution, technology innovations, and social impact, the internet-of-energy has the potential to make the user more empathetic about consumption, and hence reduce the wastage. The interactive energy system– enabled by internet-of-energy leveraging new intelligence in information technology infrastructures– makes the user not only aware of the consumption but also active in controlling. The advent and evolution of the internet-of-energy have an impact in other sectors of the economy, in particular, the development of many fast-growing smart cities. predicts a full-fledged IoT eon by 2030 [336, 337].
7.6.3 Challenges and Future Research Opportunities

Before moving on to the full capability of IoE, it is incumbent to have a proper understanding of the challenges that the combination of IoT in the smart grid may bring into the dynamics. The most cardinal challenge is the possible data leakage; consumers’ sensitive information can be revealed from the data obtained from the appliances scheduling. For example, heater usage data in the wintertime or air-conditioner usage data in the summertime implies the availability (or absence) of the residents. This data, if leaked, can lead to burglary or undesirable events and practices. Again, as all the consumers’ information are readily available in the central server of the utility provider, consumers’ privacy in the network can be compromised by cyberattacks. Cyberattack is another major concern. Cyberattackers can—by rifiting the IoT-enabled-smart-grid-infrastructure—manipulate the data transferred between users and utilities and present incorrect decisions to the sensors connected to all the smart meters. Subsequently, the appliances operate in an incorrect way and get damaged, thereby causing a serious financial meltdown. Especially, these challenges involved in commercial and industrial sectors can lead to an economic catastrophe around the world. For instance, any industrial enterprise integrated with the internet-of-energy, if subjected to cyberattacks, may need to compromise their functioning, and it can discredit the entire production.

Unreliable or unpredictable internet connectivity is another concern followed by swifter connectivity requirements for on-the-fly energy management analysis [338, 339, 238].

The future directions and research opportunities regarding the IoT enabled smartgrid are multifacted. In the physical layer of the internet-of-energy, energy acquisition and considering IoT based devices for different conditions,
situations, and environment opens up research opportunities for scholars and new entrants in the future. In the network layer, more research is required in data fusion technologies, deployment techniques for new power supply products, and communication technologies. As the number of data sources grows with the deployment of IoE, a single source may not be effectual in providing useful insights and information. On the flip side, it is cumbersome and expensive – from the data collection and management point of view – to store data from all the available sources. Advance data fusion techniques can help integrating multiple data sources and deliver more accurate, consistent, and useful insights. In the transport layer of this new technology, data transfer at data centers avoiding network congestion and data traffic can be the future research challenge and directions. Network congestion is the reduced quality of service in a network due to carrying more data in its link or node than that it can typically handle, affecting queueing delay, blocking of new connections, and packet loss. In the application layer of the internet-of-energy, research challenges centered around the integration of IoT enabled devices, edge servers, and data handling issues are required to address more efficiently and consistently in the future. The integration of IoT enabled devices requires logical connectors, commonly known as APIs, allowing applications to communicate with other IoT devices. They expose data that enables devices to transmit data to applications, functioning as a data interface. The other avenues of the internet-of-energy that may draw the attention of the research community in the future for further research and development are standardization, authorization and privacy with authentication, and avoiding cyberattacks with robust security management [340].
7.6.4 IoE and Computational Sustainability

The overarching goal of the computational sustainability network (CompSustNet) is to promote a platform that unites and helps more and more scholars, across the domain, use data mining techniques to solve the most complex and pressing problems of this time. The most compelling aspect of this virtual network—besides making a platform for computational science researchers to put their muscle towards making the world a more sustainable and livable place—is that a new method or solution created to solve one particular problem can be repurposed for another distinct problem. Table 7.6 and 7.7 present the broader computational techniques addressed in the CompSustNet publications from 2016 to date and their prospective application in the IoE conceptualizations [341]. We followed a multi-blind Delphi method to extract the broader (mother) computational techniques from more than 175 papers indexed in the CompSustNet publication section. Besides, this study relied on the comments and suggestions by the domain experts at Oregon Renewable Energy Center (OREC) at Oregon, United States, to summarize the prospective application in the IoE conceptualizations.
Table 7.6: Transferable Computational Techniques and Their Prospective Applications in IoE Conceptualization (1)

<table>
<thead>
<tr>
<th>Computational Techniques</th>
<th>Prospective Applications in IoE Conceptualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathematical Modeling</td>
<td>✓Physical System Modeling ✓IoE Characterization ✓Interpreting System Dynamics</td>
</tr>
<tr>
<td>Statistical Modeling</td>
<td>✓Energy Predictive Modeling</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>✓Generation and Demand Variability Forecasting ✓Machine Translation for Energy Systems ✓Classification, Categorization, and Clusterization of Energy Prosumers and Consumers</td>
</tr>
<tr>
<td>Reinforcement Learning</td>
<td>✓Understanding Consequences of Different Strategies ✓Resources Management in Energy Systems</td>
</tr>
<tr>
<td>Linear and Nonlinear Algebra</td>
<td>✓Matrix Representation of Physical System for Convenience of Analysis ✓Identifying Eigen Value of a System ✓Determining Overdetermined and Underdetermined System of Energy Dynamics ✓Dimensionality Reduction to Analyze Higher Dimensional Matrix</td>
</tr>
<tr>
<td>Data Mining and Visualization</td>
<td>✓Data Management System ✓Decision Analysis</td>
</tr>
<tr>
<td>Bayesian Modeling and Causal Inference</td>
<td>✓Generation Expansion ✓Generation and Demand Variability Forecasting ✓Energy Storage and Analytics ✓Tariff Designing ✓Causal Discovery ✓Preventive Maintenance</td>
</tr>
<tr>
<td>Stochastic Optimization</td>
<td>✓Generation Planning ✓Advanced Controlling</td>
</tr>
</tbody>
</table>
Table 7.7: Transferable Computational Techniques and Their Prospective Applications in IoE Conceptualization (2)

<table>
<thead>
<tr>
<th>Computational Techniques</th>
<th>Prospective Applications in IoE Conceptualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generative Modeling</td>
<td>✓ Photo/Video Prediction for Fault Detection&lt;br&gt;✓ Generate Examples for Datasets for Different Generation, Demand, and Maintenance Contexts&lt;br&gt;✓ Aging Determination of Energy-related Instruments</td>
</tr>
<tr>
<td>Time Series Analysis</td>
<td>✓ Seasonality and Trend Analysis for Decision-makers&lt;br&gt;✓ Generation and Demand Variability Forecasting&lt;br&gt;✓ Visual Analytic Interpretation of Physical Events</td>
</tr>
<tr>
<td>Game Theoretical Modeling</td>
<td>✓ Developing Multiplayer Oligopoly Games to Promote Renewable Resources&lt;br&gt;✓ Energy Democratization&lt;br&gt;✓ Algorithmic Trading&lt;br&gt;✓ Smart Home/Smart City&lt;br&gt;✓ Coordination of Decentralized Plants&lt;br&gt;✓ Coordination of IoE Maintenance</td>
</tr>
<tr>
<td>Artificial Intelligence</td>
<td>✓ Predicting Cataclysmic Events&lt;br&gt;✓ Energy Market Price Forecasting&lt;br&gt;✓ Energy Storage and Analytics&lt;br&gt;✓ Fault Maintenance</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>✓ Renewable Energies Intermittent Data Processing&lt;br&gt;✓ Cluster Computing of Distributed Energy to Find&lt;br&gt;✓ Hidden Patterns or Grouping in Data&lt;br&gt;✓ Anomaly Detection for Preventive Maintenance</td>
</tr>
<tr>
<td>Unsupervised Learning</td>
<td>✓ Summarize Policy Contents for Users&lt;br&gt;✓ Sentiment (Satisfaction) Analysis of Consumers&lt;br&gt;✓ Scam Detection&lt;br&gt;✓ Encryption/Decryption/Deidentification</td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td>✓ Analyzing Big Data and Controlling Many Generation&lt;br&gt;✓ Units at Different Time Scales&lt;br&gt;✓ Faster Decision Making&lt;br&gt;✓ Bringing Concurrency in Energy Analytics</td>
</tr>
<tr>
<td>Parallel and Distributive Systems</td>
<td>✓ Identifying Failures and Problems in Energy Networks and Fixing Them Virtually&lt;br&gt;✓ Addressing Cyber Security Concerns&lt;br&gt;✓ Addressing Outlier and Anomalous Data Problems in IoE</td>
</tr>
<tr>
<td>Change Point Detection</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 8
WildFire and Public Safety Power Shutoffs: Implementing Contextually-tailored Online Change Point Detection

The escalating frequency and scale of wildfires in the recent past and associated hazards and loss worldwide are increasing concerns [342, 343]. To understand the intensity of the burden of these hazards, Figure 8.1 portrays the annual acres burned in wildland fires from 1980-2020 in the US [344]. As evinced, the wildfire burned more than 8M acres of land annually four times in the last six years. The consequences of these wildfires in wealth and life and their physical, epidemiological, and economic aspects are even more devastating. Figure 8.2 captures these ravaging statistics of overall wildfire losses in the US, 2010-2019 [344]. As noticed, the wildfire losses from 2015 to 2029 are particularly alarming.

Among the wildfire incidents in the US, the scenario in the west of the United States, in particular California, is more dangerous. Figure 8.3 illustrates

Figure 8.1: Annual acres burned in wildland fires from 1980-2020 in the US.
the biggest wildfires (with respect to land acres) in the history of California and associated losses [345]. [345] reports that Pacific Gas Electric Co. (PG&E) power lines are responsible—directly and indirectly—for more than 1500 wildfires in California from 2014 to 2019. It also indicates PG&E equipment issues may
be accountable for Camp Fire, the deadliest fire incidents in California history, reporting the death of 85 people, loss of $16.65 Billion (worldwide costliest), and more than 153K acres of land burned. PG&E, California’s largest utility, took the onus to lead endeavors to save California from frequent, deadly, and destructive fire events [346].

PG&E investigation identifies that vegetation may ignite amidst high temperature, strong wind, and low humidity when it rubs against high voltage wires. Unlike the low voltage power lines connected to homes, usually, the high voltage lines aren’t covered in insulation. Besides, the electrical activity inside the electric transformer, which facilitates electrical current moving from high voltage to low voltage, may initiate sparks and is considered as a frequent wildfire causes. To address these catalysts and causes and prevent these hazardous events, PG&E adopted two-fold plans: long-term and sustainable plan and immediate action plan. The proposed long-term and sustainable goals include: replacing aging hardware and upgrading infrastructure to make the grid more resilient, installing transformers that contain fire-resistant fluid, moving wires underground in particularly fire-prone areas, replacing wooden poles with one from sturdier and less flammable materials, such as steel or concrete, compartmentalizing the grid into smaller segments, installing more high-definition cameras and stations in the next 5-10 years, and most importantly, insulating high-voltage power lines to prevent sparks of molten metal from erupting. A recent study shows that, in the last fifteen years, approximately 75% more forest got dried out, and on average, nine extra days added to high fire potential a year in the western US. Considering these alarming threats, PG&E plans short-term urgent actions to prevent wildfire: Public Safety Power Shutoff (PSPS). PSPS is a precautionary safety measure that proactively turns off power lines in communities in the higher fire threat area. This situation
may continue for few days, and when the danger goes away, the utility will resume its service. In other words, people sacrifice and accept the unpleasant situations of PSPS for their broader safety [346, 345].

PSPS criteria include low humidity levels, a forecast of high wind, dry materials on the ground, red flag warning of fire declared by the National Weather Service, real-time ground observations. The PG&E meteorologists jointly work with Wildfire Safety Operations Center to monitor weather conditions and potential fire threats through the lens of these parameters [347]. In this effort, they primarily use two models: OPW (Outage Producing Wind) and FPI (Fire Potential Index), built on extensive historical data and experiences and state-of-the-art data analytics techniques. OPW findings, factoring wind speed for every unplanned outage, report that consistent high wind may initiate tree branches or debris contact energized electric lines or cause damage to transmitting equipment and engender major wildfire. Apart from that, FPI factors weather, fuel moisture, climatology, and wildfire data [346]. It indicates the growth potential of wildfire escalates as vegetation dries and wind speed increases. PG&E reports interesting findings while integrating considerations from OPW and FPI models. It identifies four scenarios: Blue Sky Day (conditions are low/mild in both OPW and FPI spectrums), Hot/Dry Day (low in OPW spectrum, but high in FPI spectrum, indicating the possibility of large fire ignition), Winter Storm (high in OPW spectrum, but low in FPI spectrum, also showing a PSPS outage probability), Windy plus Dry Fuels (conditions are high/severe in both OPW and FPI spectrums, indicating a high probability of large fire ignition and PSPS outage). PG&E uses this analysis and technology not only to drive the PSPS decisions but also plan to prioritize system hardening and infrastructure improvement work. To reduce the impact of PSPS events, PG&E delves into further research to reduce the number of customers impacted,
duration, and frequency, improve coordination with communities and customers, and provide better access to customers with independent living needs [347, 346].

PG&E reports indicate that both the OPW and FPI-based models and public safety power shutoff criteria follow subjective interpretation and suspicion. This is primarily due to the cascaded nature of the problem. First, the red flag warning declared by the national weather service indicates a strong probability of ignition initiation. PG&E investigation identifies that vegetation may ignite amidst high temperature, strong wind, and low humidity when it rubs against high voltage wires. Unlike the low voltage power lines connected to homes, usually, the high voltage lines aren’t covered in insulation. Besides, the electrical activity inside the electric transformer, which facilitates electrical current moving from high voltage to low voltage, may initiate sparks and is considered as a frequent wildfire causes. However, the ignition may occur due to human-made reasons as well. Second, PG&E features criteria or conditions that may propagate this ignition and further amplify it to wildfire or even conflagration. These criteria include consistent low humidity, high wind, and high temperature. When a situation meets these two-step measures, the case seeks further investigation if there is a low moisture content of live vegetation via real-time ground observation. One PG&E report points at that less than 12% of the area covered by PG&E is under the zone covered by the national weather service flag warning system. It means a significant area covered by PG&E energy administration primarily depends on monitoring crucial factors for these cascaded events. As these factors can be represented as time series, change point detection can be an excellent approach to unpack the structural changes in these factors. Besides, it will imply if any particular condition, such as low humidity, remains consistent. It is to note that the interpretation of change points is subjective, and it only indicates structural changes (either positive way or
Figure 8.4: Proposed dynamics for Wildfire-risk and PSPS event monitoring.

negative) in the parameters over time. Understanding change points in temperature, humidity, wind, and precipitation regime can interpret wildfire events and PSPS intervention.

In Figure 8.4, we propose a wildfire and PSPS event monitoring flowchart where change point detection of the vital parameters is crucial. The flowchart initiates with system readiness as the system readiness of any power station (such as PGE) is checked at the start of the cycle. The war room or Emergency Operations Center (EOC) is then examined for reaction. When the war-room is
ready, real-time data monitoring begins. The existing real-time monitoring dynamics include Ground-based, UAV-based, and weather-forecasting monitoring to understand the possibility of wildfire and PSPS events. This monitoring is essential. We further stress the necessity to navigate the structural changes in the vital parameters. If a significant structural change happens in any critical parameter, the proposed contextually tailored Bayesian online change point detection algorithm will recognize it. Understanding and interpreting the changepoints and structural changes in parameters are both subjective and objective processes. Based on these structural changes and their interpretations, the monitoring team will update the flag status. This monitoring is a continuous process, although the team will be particularly notified in the case of any change point detected for further investigation. For instance, if the combination of change points in a particular time indicates medium alert—close monitoring of the event is necessary, but there is no very high risk of wildfire, and no PSPS decision requires—yellow flags will be raised. Any region under the yellow flag will be subjected to extensive monitoring, and the ground level investigation may help adjust the alert level. If the combination of change points in a particular time indicates high alert—followed by the immediate ground-based and UAV-based monitoring—the red flag is raised. If all three monitoring (a. OPW, FPI, and PG&E existing criteria-based, b. change point detection-based, and c. in-location inspection-based monitoring) confirms a strong possibility of an imminent ignition or wildfire condition, the power line is terminated for the consumers. The consumers will get an alert one-two day before the probable event based on OPW, FPI, and PG&E existing criteria-based and change point detection-based monitoring. The consumer alert level will be updated based on in-location inspection-based monitoring within few hours. The frequent alert update will help the consumer engagement in this navigation process and
prepare them for a possible shutoff. Customers will get the final alert a few minutes before the actual shutoff takes place to assure readiness. When there is no danger available—primarily assured by the in-location inspection—the power will be restored. The event is documented once power is restored, as it may help in understanding future events.

Figure 8.5 illustrates a preliminary study on investigating parameters crucial to wildfire and PSPS events to understand the wildfire warning scenarios in 2019. We used the WRCC data source for this preliminary study. Figure 8.5 depicts—in the 2019 time span—change point detection algorithm is helpful in interpreting wildfire risk warning presented in Figure 8.5(a).
Here, we consider only three factors initially based on the availability of the data. These factors include precipitation, maximum temperature, minimum temperature (an indicative parameter for low humidity). Here, in our data source, data were available only for averaged value for a day which is not an ideal situation for this contextually tailored Bayesian online change point detection. So, for this preliminary study, we only use the regular Bayesian online change point detection to understand the utility of this algorithm in realizing wildfire events. However, we plan to integrate the proposed algorithm to the real-world PSPS monitoring dynamics’ sensor data through further collaborations. In that case, we propose to consider the 30-minutes average value as a single observation to have functional interpretation towards understanding PSPS. Here, precipitation, among these variables, has a negative correlation with wildfire. Other variables are correlated positively.
Chapter 9
Conclusion

9.1 Contribution of this Dissertation

The contribution of this dissertation can be compartmentalized into two domains: Medical Informatics and mHealth, and Computational Sustainability. In the Medical Informatics and mHealth concentration, the contribution is three-fold. First, this dissertation investigated the prevalence and underlying relation of the sepsis diagnosis criteria (qSOFA and SIRS). This research unpacked the most prevalent qSOFA and SIRS criterion, the most prevalent sepsis-3 and sepsis-2 scenario, and investigated if there exists any multicollinearity among the qSOFA parameters and SIRS parameters. Quantifying the prevalence of qSOFA and SIRS and understanding underlying relationships among parameters have important implications for sepsis treatment initiatives in ICU and informing hospital resource allocation.

Second, this dissertation discussed different avenues towards developing a more sustainable medical informatics assistive solution that will help make evidence-based judgments instead of flummoxing the caregivers in decision making. The research delved into the available machine learning-based solutions and existing research gaps, explained why change point detection-based assistive and practitioner-centric solution has potential to be more sustainable in sepsis monitoring and outlined how we can design the change point detection algorithm addressing contextual challenges for sepsis monitoring. A sustainable medical informatics assistive solution can make the treatment regime more effective and has the potential to decrease sepsis-related poor outcomes.

Third, the dissertation discussed the data-driven tool developed as a
medical informatics solution that helps ICU practitioners and researchers to monitor and intervene on the existing sepsis patients more efficiently and interactively. The discussion emphasized how this tool can assist the practitioners in data-driven decision making, how the tool addressed the recent shift in sepsis definition, and how several additional features of this data-driven software tool help ensure efficient monitoring and intervention. The tool, SepINav, facilitates both ICU practitioners and medical researchers and offers both real-time monitoring and retrospective study, and assists the practitioners’ decision-making from multiple aspects.

Then, the dissertation unraveled the computational sustainability perspective of the medical informatics research we discussed. Computational Sustainability is a movement facilitated by CompSustNet—a virtual network led by Cornell University and supported by NSF—so that a novel scientific method, algorithm, or solution innovated to solve one particular problem of one domain can be repurposed for another distinct problem of another domain with a similar computational nature. That may cover from applied mathematics, statistics, computer and information science to healthcare, electrical engineering, economics, environmental science, operational research, and policymaking. This dissertation research discussed the Internet of Energy, its different conceptual elements, architecture, integration, challenges, impacts, future research opportunities, and computational sustainability perspectives. Then, the research discussed How the Contextually-tailored Bayesian Online Change Point Detection Algorithm can be repurposed to address the PSPS issues impacting grid resiliency of IoE and Wildfire threat in the western United States. This endeavor towards Computational Sustainability has the potential to address Public Safety Power Shut-off to prevent cascaded failure in the event of a wildfire and save nature and human life.
9.2 Impact

The impact of this dissertation research can be discussed in two lenses: Immediate Impact and Long-term Impact.

9.2.1 Immediate Impact

The first research question (RQ1) introduced in this dissertation and the relevant results to address this question can help in modeling the hospital resource allocation for that particular hospital and motivate other providers to adopt data-driven allocation for sepsis management. Besides, it can help in designing treatment regimes and nurse-patient ratios in the hospitalization workflow. Apart from that, as the work bolstered that there is no multicollinearity among sepsis parameters (both for Sepsis-2 and Sepsis-3 based on MIMIC-III), it may help interpret different assumptions for statistical inferences.

The second research question (RQ2) discussed different avenues regarding sepsis monitoring and treatment initiatives. As it primarily discussed the potential pitfalls and challenges associated with predictive models developed using EMRs, it may help the research community to intensively deal with the EMR data and further consider the research gaps and pitfalls. The work may motivate toward assistive and practitioner-centric solutions, as sepsis and its associated outcomes are complicated with various host factors, pathogen factors, and treatment regimes.

The third research question (RQ3) illustrated the utility and functionalities of SepINav (Sepsis ICU Navigator): A data-driven software tool for sepsis monitoring and intervention using Contextually-tailored Bayesian online change point detection. The tool can be crucial in understanding patient sepsis trajectory and associated change points in the vital sign’s data regime and
thus has the potential to be more efficient in sepsis monitoring and treatment compared to the standalone predictive solutions.

The fourth research question (RQ4) and associated discussions help us broaden our horizon in interpretive thinking and promote multidisciplinary collaboration toward scientific problem-solving. The work itself can be crucial in monitoring PSPS and wildfire warnings and has the potential to save human life and nature.

9.2.2 Long-term Impact

This work has a considerable long-term impact and has the potential to change the current landscape of hospital monitoring toward efficient interpretation of treatment initiatives. The research team considers validation of the proposed change point detection-based monitoring from three aspects. The team proposes to use recently released MIMIC-IV data, Pragmatic retrospective clinical trial, pragmatic clinical trial to validate the results and utility and further investigate and validate causal inference perspectives of the changes in physiological data regimes centered around different host factors, pathogen factors, and treatments regimens. The team proposes further investigation on designing the backup in the case if the change point detection fails to any extent.

The research discussed in RQ4 has the potential to make the PSPS events smaller, shorter, and smarter. It may also make the event interpretable to the customers who are making the actual sacrifice. Adding change point detection-based assistive monitoring in the PGE current dynamics can help in decision-making more constructively and ahead of time. As the monitoring aims to provide weighted warnings based on the change points, it can help to plan initiatives accordingly and assure preparedness to confront the possibilities of hazards.
9.3 Future Works

Chapter 1 introduces the research questions that we explained and endeavored to address with scientific methods and approaches in this dissertation. From chapter 2 to chapter 6, we discuss RQ1, RQ2, and RQ3: exploring different avenues of Medical Informatics solutions for sepsis monitoring and intervention. In chapter 7, we broadly discuss Computational Sustainability and conceptual elements of IoE. Chapter 8 implements the contextually tailored online change point detection algorithm, which was primarily designed for sepsis monitoring, to understand and mitigate wildfire events, particularly in the western US. This chapter will discuss future plans and proposed intents to advance these two research.

9.3.1 Data-driven Sepsis Monitoring and Intervention using Contextually-tailored Online Change Point Detection

We recently submitted an NIH R21 grant application seeking further advancement of our proposed data-driven software tool for sepsis monitoring and intervention using contextually tailored Bayesian online change point detection. We briefly discuss several proposed research as follows.

Utility and Usability Assessment

Assessing the utility of SepINav considers two perspectives. First, we evaluate the system’s usability using the System Usability Scale (SUS). SUS is a well-recognized and reliable usability assessment tool, including a ten-item questionnaire with five response options for respondents. Dr. Khan convenes a group of ten practitioners serving the College of Medicine at the University of Central Florida as they will participate in this evaluation. This scale is easy to administer, provides reliable results on small sample sizes, can efficiently
differentiate between usable and unusable systems. The CoPI and convener of
the practitioners’ group, Dr. Khan, will brief the considerations before using
SUS. The briefing includes scoring systems, possible complexities and pitfalls in
interpretation, normalizing scores to deliver a percentile ranking, and overall
interpretations on the ease of use. The second phase of the assessment
incorporates the diagnostic of the application in a pragmatic environment. Dr.
Khan adopts Delphi methods to avoid hierarchy bias and group-thinking bias.
The evaluation will cover both subjective and objective interpretations of the
detected change points, rationales from the host, pathogen, and intervention
perspectives, and its capacity to predict sepsis-related poor outcomes.

**Evaluation Metrics**

For quantitative evaluation, we adopt MIMIC-IV (the latest release of the
MIMIC database). The PI and mHealth expert, Dr. Ahamed, will convene this
evaluation team of seven individuals, comprising computer science doctoral
researchers and developers at Marquette University, a statistics professor at
Marquette University for statistical validation, Dr. Khan, and practitioners at
the University of Central Florida. The medical practitioner group will conduct
the data annotation process. We adopt two evaluation metrics as we assess the
efficacy from two lenses: Classification (formulating the problem as there are
change point class and non-change point class) and Clustering (developing the
problem as each partition represents a cluster). For classification, we will
primarily use the F1 score as it reflects both precision and recall. For Clustering,
we will use covering metrics. Besides, the Adjusted Rand Index (ARI), Variation
of Information (VI), and Hausdorff Distance will also be calculated to realize the
clustering perspective.
Understanding Sepsis and its Complicated Dynamics

SepINav embodies a step towards bridging the perceptual gap between explainability and computational intricacy centered around machine learning-based medical informatics solutions. It is represented by the facility to navigate both patients’ trajectory and the interventions made by the practitioners and to detect the changepoints in vital signs that may harbor prior to septic shock onset. We will study how practitioners can take advantage of the insights they develop from using this application into intervention dynamics and eventually improve the treatment management’s efficacy. SepINav facilitates interpreting different host factors as it will make demographic information (related to host factors) available to practitioners in the same interface where we illustrate the change points in the vital signs, practitioners’ interventions, patients’ conditions in the sepsis spectrum. The change point detected in the patients’ vital signs regime will help the practitioners and medical informatics researchers to seek rationales behind this regime-switching and interpretations towards sepsis monitoring and caregiving. We will collect answers to a simple questionnaire with 11 essential questions in 3 domains – i) symptoms and treatment, ii) support and facilities, and iii) suspicion-driven practice. These answers will be added as a special note while providing the bi-weekly report for practitioners.

Team Composition and Capacity Building

As the proposed study is at the intersection of mHealth application development and clinical decision support system, our interdisciplinary team formed from the Department of Computer Science at Marquette University, College of Medicine at the University of Central Florida, and Department of Mathematical and Statistical Sciences at Marquette University. The PI, Dr. Sheikh Iqbal Ahamed,
is a Computer Science professor and has a considerable NIH track record in mHealth solutions and Medical Informatics research. The Co-PI, Dr. Khan, has significant experience and is one of the leading researchers and clinicians in critical care medicine. Dr. Bansal, an eminent statistician from Marquette University, will contribute to this project as a consultant and help us in statistical validation. This project will form a core capacity building initiative as follows: a) Dr. Ahamed will supervise studying the generalizability in prevalence and dichotomy (or multicollinearity) using multiple electronic medical records frequently used in the clinical studies as a part of different projects in Ubicomp Lab of Department of Computer Science at Marquette University. b) Dr. Khan will seek the rationale behind the prevalence of a vital sign in bedside monitoring from the pathophysiological and pathobiological aspects. c) Dr. Ahamed and Dr. Bansal will study the rationale behind the lack of multicollinearity between the frequent vital signs pairs, albeit being theoretically dependent on each other. This study may extend to seeking causal relations between each of the pairs. d) Dr. Khan will lead a group of practitioners at the College of Medicine, University of Central Florida, to annotate vital sign data of MIMIC-IV pertaining to sepsis. e) Dr. Ahamed and Dr. Bansal will supervise students to evaluate the contextually tailored Bayesian online change point detection algorithm in the lens of classification and clustering. f) Dr. Ahamed, Dr. Bansal, and Dr. Khan will evaluate and validate the contextually tailored bayesian online change point detection algorithm and its provision to address the varying context of the problem from both patient and facility level. g) After the development phase of SepINav, the team will study the possible biases and existing research gaps before pragmatic deployment in sepsis monitoring in the intensive care unit in multiple trials.
9.3.2 WildFire and Public Safety Power Shutoffs: Implementing Contextually-tailored Online Change Point Detection

PSPS, in other words, is a sacrifice of people for a broader good—saving people's lives, resources, and lands. PG&E’s further research aims to make this event smaller in size and shorter in length. These efforts seek to reduce the number of customers affected by a PSPS event and cut restoration times, restoring power to nearly all customers as soon as possible after severe weather condition has passed. The proposed online change point detection algorithm may help in this research endeavor. The algorithm can be implemented in designing an early PSPS warning mechanism. The structural changes in critical factors have implications in modulating the degree of alertness. Investigation shows the impact of each element may not be the same in this complex decision-making. That is why the structural changes in one parameter may not indicate the same severity or alert compared to other factors. However, this understanding is crucial. We plan to use 'soft voting' to make this combined decision making. This voting may also help in developing subjective interpretations of the event probability and structural changes in factors. Besides, this voting component will be allocated a weight depending on its relevance to PSPS early warning. The entire model will be realized using a graphical user interface and may implement as an open source software tool in PSPS monitoring.
Bibliography


[30] Nazmus Sakib, Devansh Saxena, Lin He, Paul M Griffin, Sheikh Iqbal Ahamed, and Munirul Haque. Unpacking prevalence and dichotomy in


214


[94] Marja Hynninen, Matti Valtonen, Helene Markkanen, Martti Vaara, Pentti Kuusela, Irma Jousela, Anneli Piilonen, and Olli Takkunen. Interleukin 1


219


[158] Anton A Igonin, Victor W Armstrong, Maria Shipkova, Vladimir G Kukes, and Michael Oellerich. The monoethylglycine xylidide (megx) test as a


[188] Jen Christensen. 1 in 5 deaths globally are due to sepsis, study says, Jan 2020.


[341] Computational Sustainability Network Publications.


Figure 9.1: Time series depiction for the daily closing price and volume of Apple.Inc. Stock presented in Yahoo Finance. It captures data for a time around the year 2000, sampled at every three-time steps in an effort to reduce the length of the series. From the contextual aspect, 09-29-2000 experienced a notable drop in stock price [5].

Figure 9.2: Time series of the available amount in an anonymous person’s bank account. The data stream shows periodic attributes and significant variation with time [5].
Figure 9.3: Time series representing movement switches of Honey bee between three states: waggle, right turn, and left turn [5].

Figure 9.4: Bitcoin Price in US Dollar, data collected from bitcoin.com [5].

Figure 9.5: Price of Brent Crude Oil per barrel in US Dollar. Data was collected from EIA (Energy Information Administration) of the United States and then sampled at every ten observations to reduce the series’s breadth [5].
Figure 9.6: Time series representing monthly total business inventories in US dollars. From the contextual aspect, the effects of the financial crisis are apparent. Data were collected from the US Census Bureau [5].

Figure 9.7: Time series depicts the population of Centralia, the mining town of Pennsylvania. The context is that a mine fire burning since 1962 has a significant impact, and in 1992, it experienced condemning all buildings in the town [5].

Figure 9.8: Number of children born per women on average globally (Source: GapMinder) [5].
Figure 9.9: Emissions of $CO_2$ per person in Canada. (Source: GapMinder) [5]

Figure 9.10: Time series representing total construction spending in the private sector in the United States over time. (Source: United States Census Bureau) [5]

Figure 9.11: The debt ratio of Government in Ireland over time. It captures the context of the visible financial crisis in 2007-2008. (Source: EuroStat) [5]
Figure 9.12: GDP over time in constant local currency in Argentina. (Source: World Bank) [5]

Figure 9.13: GDP over time in constant local currency in Croatia. (Source: World Bank) [5]

Figure 9.14: GDP over time in constant local currency in Iran. (Source: World Bank) [5]
Figure 9.15: GDP over time in constant local currency in Japan. (Source: World Bank) [5]

Figure 9.16: Global $CO_2$ level per month over time. To reduce the length of the time series, collected data were sampled every four years. [5]

Figure 9.17: Time series representing the number of home runs each year in the American League of baseball. (Source: Data retrieved from Baseball Databank). From the contextual aspect, Major League Baseball was expanded, indicating a potentially significant impact of World War-II [5].
Figure 9.18: The number of visitor entries per month through Keflavik airport, Iceland (Data Source: Icelandic Tourist Board) [5].

Figure 9.19: Monthly arrival and departure count of passengers at John F. Kennedy Airport, New York city (Data Source: Port Authority of New York and New Jersey) [5].

Figure 9.20: Monthly arrival and departure count of passengers at LaGuardia Airport, New York city (Data Source: Port Authority of New York and New Jersey) [5].
Figure 9.21: Monthly measles cases in England and Wales for a certain period of time [5].

Figure 9.22: Time series representing the annual volume of the Nile River at Aswan, Egypt. This data contextualizes the change point that happened as an effect of a dam built in 1898 [5].

Figure 9.23: Time series representing room occupancy associated parameters by Candanedo and Feldheim in 2016. To abridge the length of the series, collected data was sampled, while preparing the time-series, at every 16 observations. This multidimensional time-series captures four measurements: temperature, relative humidity, light, and \( \text{CO}_2 \), respectively (Data Source: UCI Repository) [5].
Figure 9.24: Ozone-depleting substances' level in the atmosphere over time, contextualizing the impact of the Montreal Protocol enforced in September 1989 [5].

Figure 9.25: Time series representing total available rail-lines in the world in Kilometers (Data Source: World Bank) [5].

Figure 9.26: Ratner Group’s stock price time-series, sampled at every 3 original observations of the collected data, in the early 1990s (Data Source: Yahoo Finance) [5].
Figure 9.27: Time series representing the number of automated phone calls in the United States over time. In context, these numbers are being subjected to vary with the degrees of regulation imposed by the government [5].

Figure 9.28: Time series data stream representing pace and total distance covered by a runner while following an interval training program (pace indicates the runner’s alteration between running and walking in an interval) [5].

Figure 9.29: A horizontal scan line of an image (Data collected from the BSDS, image no. 126007) [5].
Figure 9.30: A horizontal scan line of an image (Data collected from the BSDS, image no. 42049) [5].

Figure 9.31: Time series representing the impact of introducing seat-belt mandate in the UK as the number of drivers killed or seriously injured experienced a gradual decline. In context, seat-belts were compulsory in the UK for the new cars starting in 1972 and were mandatory to be worn for the driver from 1983 onward [5].

Figure 9.32: Number of license plate applicants over time in Shanghai [5].
Figure 9.33: Number of workers employed in coal mines in a spectrum of hundred years in Britain [5].

Figure 9.34: Time series representing the unemployment rates in The Netherlands [5].

Figure 9.35: Fluctuation in the conversion rate between USD and Icelandic Krona, representing a sharp drop as the impact of the financial crisis (Data Source: EuroStat) [5].
Figure 9.36: The population of the United States in the spectrum of seventy years (Data Source: United States Census Bureau) [5].

Figure 9.37: Well-log data, after sampling the original observations in an interval of six to truncate the breadth of the series. [6].
Figure 9.38: A simulated series with a predefined change point. A data stream with Gaussian noise and a small trend, and a series with an offset and uniform noise are partitioned at index 146 [5].

Figure 9.39: A simulated series with a predefined change point, having a constant noise and a mean shift change point at index 97 [5].

Figure 9.40: A simulated series with a predefined change point. A series having the noise of $N(0,1)$ and a series having the noise of $N(2,2)$ are partitioned at index 179. This simulated series also has an outlier at index 42 [5].
Figure 9.41: A simulated series with multiple periodic components of different amplitude and a predefined change point at index 341. Contextually, the series reveals a change point since the mean shifts [5].

Figure 9.42: A simulated series without having a change point follows $N(0, 1)$ [5].