UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Jessica Sena de Souza

Intelligent ICU Monitoring: Investigating the Role of Accelerometers

Belo Horizonte
2024

Jessica Sena de Souza

Intelligent ICU Monitoring: Investigating the Role of Accelerometers

Final Version

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor of Computer Science.

Advisor: Prof. Dr. William Robson Schwartz
Co-Advisor: Prof. Dr. Parisa Rashidi

Belo Horizonte
2024

Souza, Jéssica Sena de.

S729i Intelligent ICU monitoring: [recurso eletrônico] investigating the / role of accelerometers / Jéssica Sena de Souza – 2024.
1 recurso online (131 f. il, color.) : pdf.

Orientador: William Robson Schwartz
Coorientadora: Parisa Rashidi.
Tese (Doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciências da Computação.
Referências: f. 88-101

1. Computação – Teses. 2. inteligência artificial – Teses.
3. Unidade de terapia intensiva – Acelerometria -Teses.
4. Monitoramento de paciente – Técnicas digitais – Teses.
5. Dor – Classificação – Teses. I. Schwartz, William Robson.
II. Rashidi, Parisa. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação.
IV. Título.

CDU 519.6*82(043)

Ficha catalográfica elaborada pela bibliotecária Irenquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Intelligent ICU Monitoring: Investigating the Role of Accelerometers

JÉSSICA SENA DE SOUZA

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. WILLIAM ROBSON SCHWARTZ - Orientador
Departamento de Ciência da Computação - UFMG

PROFA. PARISA RASHIDI - Coorientadora
Departamento de Engenharia Biomédica - University of Florida

PROF. ADRIANO ALONSO VELOSO
Departamento de Ciência da Computação - UFMG

PROF. LEONARDO ANTÔNIO BORGES TORRES
Departamento de Engenharia Eletrônica - UFMG

PROF. MOACIR ANTONELLI PONTI
Departamento de Ciência da Computação - USP

PROF. FLORENC DEMROZI
Department of Electrical Engineering and Computer Science - University of Stavanger

Belo Horizonte, 5 de junho de 2024.

Acknowledgments

Completing this journey would have been impossible without the unwavering support and encouragement of those closest to me.

I am profoundly grateful to my beloved Alex, whose unwavering support, love, and understanding have sustained me throughout this journey. Your encouragement and belief in me have guided me through the challenges and triumphs of pursuing this Ph.D. I am endlessly thankful for your patience, sacrifice, and constant reassurance.

To my dear mother, your love and sacrifices have shaped me into the person I am today; this accomplishment is as much yours as it is mine. Thank you for always standing by me and instilling in me the values of perseverance and determination.

To my dear friends, your camaraderie, laughter, and unwavering encouragement provided solace during the darkest hours and made the brightest moments even more joyful. Your belief in me never wavered; I am deeply grateful for that.

I am indebted to my esteemed advisors for their invaluable guidance, mentorship, and scholarly wisdom throughout this endeavor. Your dedication to excellence and insightful feedback have shaped this dissertation and my academic growth and intellectual development.

I would like to extend my heartfelt gratitude to the Fulbright Program for their generous support, which made my research possible. The Fulbright scholarship provided me with the invaluable opportunity to spend nine months at the University of Florida, where I was able to collaborate closely with esteemed researchers from the iHeal Laboratory and the Intelligent Clinical Care Center. This collaboration greatly enriched my work, offering critical insights and expertise that were instrumental in the successful completion of my thesis. I am profoundly thankful for this unique and transformative opportunity. I would like to thank the National Council for Scientific and Technological Development – CNPq (Grant 312565/2023-2).

I sincerely appreciate my esteemed colleagues at the SSIG, Sense, UAI, and IHeal laboratories. Your camaraderie, collaboration, and shared passion for discovery have enriched my academic experience beyond measure. Together, we have overcome challenges, celebrated victories, and forged lifelong bonds.

This dissertation stands as a testament to the collective efforts of many, and I am deeply grateful for each and every one of you.

"*Everyone in academia is smart, distinguish yourself by being kind.*"
(Vidita Vaidya)

Resumo

O monitoramento inteligente de pacientes, utilizando tecnologias como inteligência artificial e análise de dados, oferece *insights* em tempo real sobre os riscos à saúde do paciente, melhorando o estado do paciente. A Unidade de Terapia Intensiva (UTI) tem se digitalizado intensamente, acumulando vastos dados de pacientes a partir de registros eletrônicos de saúde. No entanto, métricas essenciais como acuidade do paciente e níveis de dor muitas vezes são negligenciadas devido à observação limitada da equipe. Com a equipe da UTI sob grande estresse e muitos enfermeiros enfrentando esgotamento, o desafio de monitorar manualmente extensos dados destaca a necessidade de técnicas avançadas de monitoramento. Acelerômetros surgiram como uma solução potencial, oferecendo monitoramento contínuo dos movimentos dos pacientes, qualidade do sono e detecção precoce de condições como sepse. Apesar de seu potencial, seu uso na avaliação da dor e acuidade permanece pouco explorado. Esta tese visa preencher essa lacuna, examinando se os acelerômetros podem fornecer indicadores precisos de dor e prever a deterioração do paciente na UTI, levando potencialmente a intervenções oportunas e melhores resultados para o paciente.

Palavras-chave: acelerômetro; UTI; classificação de dor; acuidade.

Abstract

Intelligent patient monitoring, using technologies like artificial intelligence and data analytics, provides real-time insights into patient health risks, improving care outcomes. The Intensive Care Unit (ICU) has greatly digitalized, accumulating vast patient data from electronic health records. Yet, essential metrics like patient acuity and pain levels are often overlooked due to limited staff observations. With ICU staff under significant stress the challenge of manually monitoring extensive data emphasizes the need for advanced monitoring techniques. Wearable accelerometers have emerged as a potential solution, offering continuous monitoring of patient movements, sleep quality, and early detection of conditions like sepsis. Despite their potential, their use in assessing pain and acuity remains underexplored. This thesis aims to fill this gap, examining whether accelerometers can provide accurate pain indicators and predict patient deterioration in the ICU, potentially leading to timely interventions and better patient outcomes.

Keywords: accelerometers; ICU; pain classification; acuity assessment.

List of Figures

1.1 Intelligent ICU system 15

3.1 Data types 35
3.2 Defense and Veterans Pain Rating Scale (DVPRS). Figure extracted from Polomano et al. [100]. 41

4.1 Distribution of Self-Reported Pain Levels. This figure illustrates the distribution of pain levels within the dataset. An imbalance in the distribution is evident, characterized by a concentration of samples at pain level 0 (no pain). This may reflect pain management via medication. 45
4.2 Distribution of Self-Reported Pain Levels. 46
4.3 Histograms of the probability density of acceleration magnitudes across different pain levels 47
4.4 Histograms of the probability density of acceleration magnitudes for ICU patients, categorized by gender. 48
4.5 Histograms of the probability density of acceleration magnitudes for ICU patients, categorized by age group. 48
4.6 Histograms of the probability density of acceleration magnitudes for ICU patients, categorized by health status. 49
4.7 Histograms of the probability density of acceleration magnitudes for ICU patients, categorized by disease. 51
4.8 Comparison of different feature separability 52
4.10 Comparison of different features 53
4.11 t-SNE plot with highlighted samples from four selected patients at pain level 6 55
4.12 Accelerometer signal before and after low pass filter 56

5.1 The proposed approach is an end-to-end neural network system that leverages accelerometer and EHR data to assess patient acuity, discerning between stable and unstable states. 76
5.2 Transformer-based architecture for the acuity assessment 77
5.3 Acuity assessment evaluation protocol 78
5.4 Evaluation protocol and distribution of samples and patients. 79
5.5 Shap analysis for acuity results 83

List of Tables

2.1 Overview of the recent literature surveyed studies and their use of accelerometer 21
2.2 Comparative analysis of clinical features, accelerometer data, targeted clinical outcome, and learning models across the reviewed studies. 23
2.3 Data Processing and Machine Learning algorithms used by recent papers . . . 25
2.4 Features extracted by recent studies in the literature 28

3.1 Pain cohort characteristics 36
3.2 Acuity cohort characteristics 37
3.3 Distribution of demographic variables of encounters (recorded every four hours) stratified by class labels (stable, unstable) 38

4.1 Demographic Information of Four Selected Patients. 54
4.2 List of features divided by domains 57
4.3 List of Models with Descriptions 59
4.4 Performance metrics of the best model for pain classification using different undersampling methods. 64
4.5 DVPRS 11 classes classification 65
4.6 Pain vs No Pain Classification 67
4.7 Mild vs Moderate 68
4.8 Moderate vs Severe 70
4.9 Mild x Severe 71
4.10 Pain Variation 72
4.11 Classifier Performance 73
4.12 Best Inputs for Different Pain Classification Experiments 74

5.1 Acuity assessment results. 81
5.2 Best hyperparameters for each scenario. 82
5.3 Overview of the hyperparameters and their respective values explored in the hyperparameter optimization 83

A.1 DVPRS 11 Classes - Day - Accelerometer 102
A.2 DVPRS 11 Classes - Day -Accelerometer and Clinical 103
A.3 DVPRS 11 Classes - Day -Accelerometer and Demographics 104
A.4 DVPRS 11 Classes - Day -Accelerometer and Diseases 105
A.5 DVPRS 11 Classes - Day -Accelerometer and Medications 106

A.6 DVPRS 11 Classes - Day -Accelerometer, Medications, Demographics and Clinical . . . 107
A.7 DVPRS 11 Classes - Day -Accelerometer, Medications, Demographics, Clinical and Diseases . . . 108
A.8 DVPRS 11 Classes - Day -Accelerometer, Medications, Demographics . . . 109
A.9 DVPRS 11 Classes - Day -Clinical . . . 110
A.10 DVPRS 11 Classes - Day -Demographics . . . 111
A.11 DVPRS 11 Classes - Day -Diseases . . . 112
A.12 DVPRS 11 Classes - Day -Medications . . . 113
A.13 DVPRS 11 Classes - Night -Accelerometer . . . 114
A.14 DVPRS 11 Classes - Night -Accelerometer and Clinical . . . 115
A.15 DVPRS 11 Classes - Night -Accelerometer and Demographics . . . 116
A.16 DVPRS 11 Classes - Night -Accelerometer and Diseases . . . 117
A.17 DVPRS 11 Classes - Night -Accelerometer and Medications . . . 118
A.18 DVPRS 11 Classes - Night -Accelerometer, Medications, Demographics and Clinical . . . 119
A.19 DVPRS 11 Classes - Night -Accelerometer, Medications, Demographics, Clinical and Diseases . . . 120
A.20 DVPRS 11 Classes - Night -Accelerometer, Medications and Demographics . . . 121
A.21 DVPRS 11 Classes - Night -Clinical . . . 122
A.22 DVPRS 11 Classes - Night -Demographics . . . 123
A.23 DVPRS 11 Classes - Night -Diseases . . . 124
A.24 DVPRS 11 Classes - Night -Diseases . . . 125
A.25 DVPRS 11 Classes - Night -Medication . . . 126
A.26 Mild vs Moderate - Day -Accelerometer . . . 127
A.27 Mild vs Moderate - Day -Accelerometer and Clinical . . . 128
A.28 Mild vs Moderate - Day -Accelerometer and Demographics . . . 129
A.29 Mild vs Moderate - Day -Accelerometer and Diseases . . . 130
A.30 Mild vs Moderate - Day -Accelerometer and Medications . . . 131

Contents

1 Introduction **14**

1.1 Motivation . . . 15
1.2 Research Questions . . . 16
1.3 Contributions . . . 17
1.4 Outline . . . 18

2 Literature Review **19**

2.1 Patient Monitoring Using Accelerometers . . . 19
2.1.1 Pain Classification . . . 21
2.1.2 Acuity Assessment . . . 22
2.2 Accelerometer Data Pre-processing . . . 24
2.3 Artificial Intelligence . . . 27
2.3.1 Classical Learning Models . . . 27
2.3.2 Deep Learning Models . . . 30
2.4 Interpretability . . . 32

3 Intelligent ICU System **34**

3.1 Patient Recruitment . . . 35
3.2 Data Collection . . . 37
3.3 Data Curation and Normalization . . . 39
3.4 Data Labeling and Windows Segmentation . . . 40
3.4.1 Pain Classification . . . 40
3.4.2 Acuity Assessment . . . 43

4 Pain Classification **44**

4.1 Data Analysis . . . 44
4.1.1 Distribution of pain levels . . . 44
4.1.2 Distribution of acceleration values . . . 46
4.1.3 Data separability analysis . . . 50
4.2 Methodology . . . 55
4.2.1 Feature extraction and selection . . . 55
4.2.2 Machine Learning Methods . . . 58
4.2.2.1 Linear Models . . . 58
4.2.2.2 Naive Bayes Models . . . 60

4.2.2.3 Gradient Boosting Methods . . . 61
4.2.2.4 Meta-Estimators . . . 61
4.2.2.5 Non-Parametric Methods . . . 62
4.3 Evaluation Protocol . . . 63
4.4 Experimental Results . . . 64
4.4.1 Multiclass classification of pain scores . . . 64
4.4.2 Pain vs No Pain Classification . . . 66
4.4.3 Mild vs Moderate . . . 66
4.4.4 Moderate vs Severe . . . 69
4.4.5 Mild vs Severe . . . 69
4.4.6 Pain Variation . . . 69
4.5 Discussion . . . 73

5 Acuity Assessment **75**
5.1 Methodology . . . 75
5.1.1 Deep Learning Models . . . 76
5.2 Evaluation Protocol . . . 78
5.3 Experiment Results . . . 80
5.4 Discussion . . . 84

6 Conclusions **86**

Bibliography **88**

Appendix A Pain Classification Results **102**

Chapter 1

Introduction

Intelligent patient monitoring refers to integrating advanced technologies, such as artificial intelligence (AI), machine learning, and data analytics, into monitoring and managing patient health. By leveraging real-time data from various sources, including medical devices, electronic health records, cameras, environment sensors, and wearable sensors [90], intelligent patient monitoring systems can analyze and interpret patient data to provide valuable insights and early warning signs of potential health risks or deteriorations [29]. These systems can employ algorithms to detect patterns, anomalies, and trends, enabling healthcare professionals to make timely and informed decisions regarding patient care. With intelligent patient monitoring, healthcare providers can proactively identify and address health issues, optimize treatment plans, and improve patient outcomes. This technology has the potential to revolutionize healthcare by improving patient outcomes, reducing healthcare costs, and enhancing the overall quality of care.

The Intensive Care Unit (ICU) is a specialized department within a hospital that plays a vital role in saving lives and improving patient outcomes [2]. The primary goal of the ICU is to deliver specialized and continuous care to stabilize and treat critically ill patients, closely monitor their condition, and respond promptly to any changes [2]. The ICU provides critical care and close monitoring for patients with severe or life-threatening conditions [2]. An illustration of an intelligent ICU system is shown in Figure 1.1 extracted from the work of Davoudi et al. [29].

The digitalization of the ICU through the widespread implementation of electronic health records (EHRs) led to an increase in the information documented for each patient in the ICU [19], including detailed physiological data, a variety of lab tests, and comprehensive medical history stored in EHRs [64]. However, certain crucial aspects of patient care remain untracked through automated means. For instance, factors such as dynamic assessment of patient acuity, pain assessment, patient sleep patterns, body positioning, physical activity, mobility, and functional status are not continuously and comprehensively monitored, necessitating either self-reporting or repetitive observations by ICU nurses [97, 127], therefore being limited by the time constraints imposed on healthcare providers.

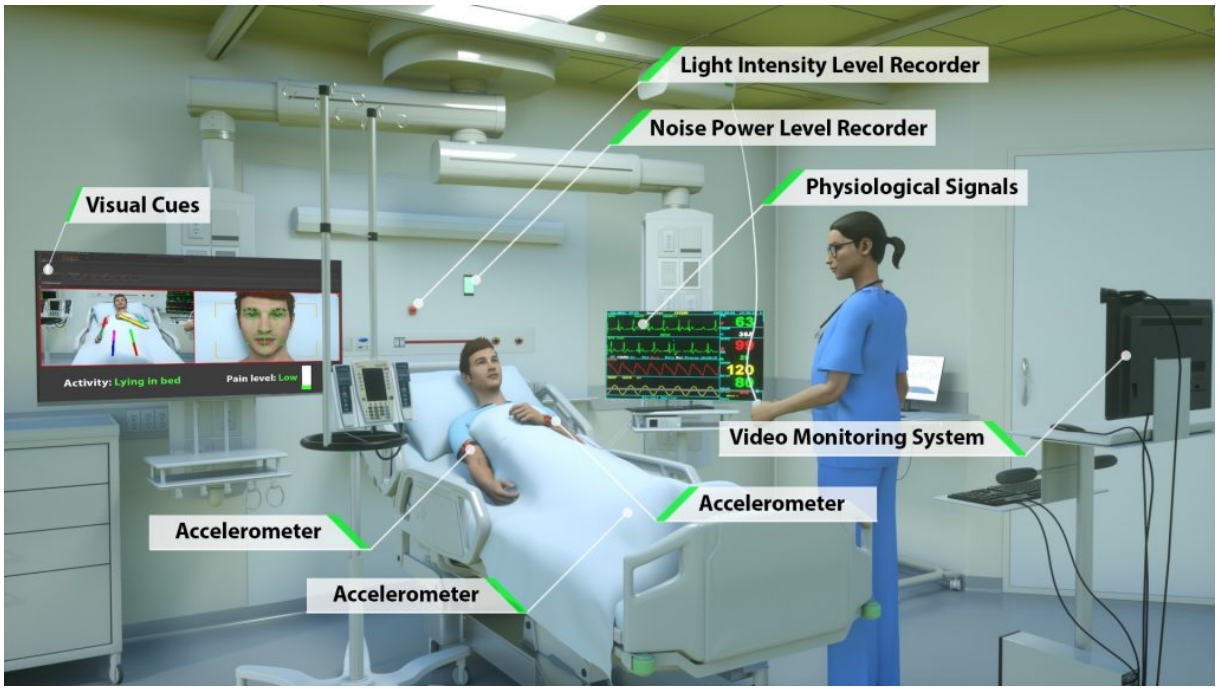


Figure 1.1: Figure extracted from the work of Davoudi et al. [29]. The Intelligent ICU system uses pervasive sensing to collect patient data and environmental data. The system includes wearable accelerometer sensors, a video monitoring system, a light sensor, and a sound sensor.

1.1 Motivation

ICU physicians spend only 9.4% of their clinical time in direct patient contact [85]. Similarly, most ICU nurses spend only 10% of their time on direct patient assessments of pain and mobility [136]. Actually, physicians or nurses may not directly observe patients for 80% of their stay in an ICU. Additionally, research has demonstrated that self-reports and manual observations can be influenced by subjectivity, have limited recall, are conducted infrequently each day, and place a heavy workload on staff [40]. This absence of thorough and ongoing monitoring can hinder the timely implementation of intervention strategies [15, 36, 72, 110, 135]. Therefore, the comprehensive monitoring and assessment of patients present an intricate challenge for healthcare professionals. With patients often exhibiting multiple coexisting medical conditions, an array of interconnected physiological variables, and the potential for rapid clinical deterioration, the demand for vigilant observation is ceaseless. The result is personnel shortages and burnout.

Critical care teams are under significant work pressure [27]. Almost a third of ICU nursing teams suffer from burnout [132]. High nursing workload is one factor in the occurrence of life-threatening adverse events in the ICU [16, 75, 95, 129]. Additionally, the human capacity to simultaneously monitor and interpret many data points is limited, and cognitive fatigue can impede the ability to detect subtle changes in a patient's condition [8]. In this context, adopting pervasive sensing techniques to monitor patients'

functional and behavioral aspects within the ICU presents an opportunity for a more comprehensive and continuous evaluation of their condition [8]. More importantly, these systems can provide timely alerts when deviations from the norm are detected, allowing healthcare providers to focus their expertise on critical decision-making and intervention [8].

An important avenue of research in pervasive sensing in the ICU involves using wearable accelerometers. These devices, designed often in the form of wristwatches, offer several advantages, including their lightweight nature, non-invasive application, and ease of use. Furthermore, they facilitate various computational analyses, pose no safety or comfort concerns for the patient, and do not interfere with the clinical procedures conducted within the ICU [102].

Accelerometer devices enable continuous monitoring of a patient's movements, allowing healthcare professionals to assess sleep quality and time [5, 33, 34, 87, 125], to measure physical activity levels and intensity [52, 67, 74, 109], to recognize patients' activities and duration [37, 44], to detect changes in patient positioning [31, 128]. Importantly, accelerometers have shown promise in detecting sepsis [17], delirium [3], and assessing acuity [121]. Additionally, accelerometers can be used for ambulation assessment and rehabilitation progress monitoring in the ICU.

While accelerometers have shown potential in predicting clinical outcomes, there's a noticeable gap in recent studies exploring their use for outcomes like pain and acuity. Additionally, there's a significant demand for systems that avoid the privacy concerns associated with vision-based monitoring [26].

1.2 Research Questions

Taking into account the gap in the literature regarding the employment of accelerometer data in the prediction of pain or acuity assessment, the research questions for this dissertation are:

- **What potential do accelerometers have in assessing and quantifying pain levels in ICU patients?** This research question explores the feasibility of using accelerometers to measure and evaluate pain levels in ICU patients. By analyzing the movement data captured by these devices, the study aims to determine if there is a correlation between physical activity and pain intensity, providing a non-invasive method for pain assessment.

- **In what ways might accelerometers aid in evaluating patient acuity in the ICU setting?** This research question investigates how accelerometers can be utilized to assess the overall condition and severity of illness in ICU patients.

The study examines whether movement patterns recorded by accelerometers can provide insights into a patient's acuity, potentially assisting healthcare providers in monitoring and managing patient care more effectively.

1.3 Contributions

The goal of this Ph.D. journey is to leave contributions to the research field and the community. Below, the main contributions are outlined. The ones we have already accomplished are marked with filled symbols, while empty symbols mark our future goals.

★ Conference paper with our pain classification pipeline (Chapter 4).

[116] **Sena, J.**, Bandyopadhyay, S., Mostafiz, MT., Davidson, A., Guan, Z., Barreto, J., Ozrazgat-Baslanti, T., Tighe, P., Schwartz, WR., Bihorac, A., & Rashidi, P. (2023). *Diurnal Pain Classification in Critically Ill Patients using Machine Learning on Accelerometry and Analgesic Data.* In IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

★ Journal paper with the extension of our acuity assessment approach (Chapter 5).

[118] **Sena, J.**, Mostafiz, MT., Zhang, J., Davidson, A., Bandyopadhyay, S., Nerella, S., Yuanfang, R., Ozrazgat-Baslanti, T., Shickel, B., Schwartz, WR,. Bihorac, A., & Rashidi, P. *Wearable sensors in patient acuity assessment in critical care.* In Frontiers in Neurology.

■ Paper with the preliminary experiments and analysis of our acuity assessment approach (Chapter 5).

[117] **Sena, J.**, Mostafiz, M. T., Zhang, J., Davidson, A., Bandyopadhyay, S., Yuanfang, R., Shickel, B., Loftus, T., Schwartz, W. R., Bihorac, A., & Rashidi, P. (2023). *The Potential of Wearable Sensors for Assessing Patient Acuity in Intensive Care Unit (ICU).* ArXiv. */abs/*2311.02251

■ Poster with a proof of concept on pain classification using deep learning models.

Sena, J., Bandyopadhyay, S., Nerella, S., Rashidi, P., and Schwartz, WR. (2023). *Accelerometer-based Pain Prediction Using Transformers: A Proof Of Concept In Critically Ill Patients.* In American Medical Informatics Association Informatics Summit.

□ Journal paper surveying the recent literature regarding the use of accelerometers in the ICU (Chapter 2 is heavily based on this manuscript).

Sena, J., Lopes, L., Lacerda, MNO., Magalhaes, H., Vilas Boas, D., Perez, L., Nery, L., Bandyopadhyay, S., Rashidi, P., and Schwartz, WR. *An Exploration of Recent*

Applications of Actigraphy in the Intensive Care Unit: A Comprehensive Study. To be submitted.

Other contributions that were not directly related to this dissertation:

- [89] Nerella, S., Bandyopadhyay S., Zhang J., Contreras, M., Siegel, S., Bumin, A., Silva, B., **Sena, J.**, Shickel, B., Bihorac, A., Khezeli, K., Rashidi, P. (2023). *Transformers and Large Language Models in Healthcare: A Systematic Review.* Journal submitted to Artificial Intelligence in Medicine. **Under Review**.
- [10] Bandyopadhyay, S., Cecil, A., **Sena, J.**, Davidson, A., Guan, Z., Nerella, A., Zhang, J., Khezeli, K., Armfield, B., Bihorac, A., Rashidi, P. (2023). *Predicting risk of delirium from ambient noise and light information in the ICU.* In arXiv preprint arXiv:2303.06253.
- [51] Gonçalves, GR., **Sena, J.**, Schwartz, WR., Caetano, CA. (2022). *Pixel-level Class-Agnostic Object Detection using Texture Quantization.* In Conference on Graphics, Patterns and Images (SIBGRAPI).
- [115] **Sena, J.**, Jordão, A., Schwartz, WR. (2021). *A content-based late fusion approach applied to pedestrian detection.* In Journal of Visual Communication and Image Representation.
- [114] **Sena, J.**, Barreto, J., Caetano, C., Cramer, G., Schwartz, WR. (2020). *Human Activity Recognition based on Smartphone and Wearable Sensors Using Multiscale DCNN Ensemble.* In Neurocomputing.
- [18] Caetano, C., **Sena, J.**, Brémond, F., Dos Santos, JA., Schwartz, WR. (2019). *Skelemotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition.* In IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS).

1.4 Outline

The remainder of this dissertation is structured as follows: After this introductory chapter, Chapter 2 delves into the background concepts and provides an extensive literature review, covering topics like patient monitoring using accelerometers, data pre-processing techniques, machine learning methods, and the importance of interpretability in these models. Chapter 3 discusses the key aspects of our data collection. Following this, Chapter 4 outlines our pain classification methodology. Chapter 5 details the methodology employed for acuity assessment. Finally, Chapter 6 presents our conclusions.

Chapter 2

Literature Review

This chapter comprehensively reviews the existing literature on automated patient monitoring using accelerometers. Given the limited number of studies that align closely with our research, we have chosen to examine the entire task pipeline. In Section 2.1, we introduce the accelerometer device and its applications in healthcare. Section 2.2 discusses the data preprocessing applied to this data type. The role of artificial intelligence in the healthcare setting is explored in Section 2.3. Section 2.4 emphasizes explainability's crucial role and relevance within the healthcare domain. These sections draw comparisons with recent works focusing on the deployment of accelerometers in intensive care units.

2.1 Patient Monitoring Using Accelerometers

An accelerometer is an electromechanical device that accurately measures acceleration forces experienced by objects due to motion or gravity. Accelerometers operate on the principle of Newton's second law of motion [92], which states that, in an inertial reference frame, force (F) is directly proportional to mass (m) and acceleration (a) according to the formula $F = ma$. Accelerometers measure acceleration by assessing the force exerted on a proof mass within the device when subjected to movement. Modern accelerometers typically employ microelectromechanical systems (MEMS) technology, consisting of tiny, microfabricated structures that can measure acceleration with high precision. A basic accelerometer consists of a proof mass, a spring, and a transducer. When the accelerometer accelerates, its mass moves in response to the applied force that is proportional to the acceleration imposed on the accelerometer. This force induces stress on the spring, resulting in an elastic deformation of the spring proportional to the force. The transducer, often based on piezoelectric or capacitive principles, converts this displacement into an electrical signal proportional to the applied acceleration. Also, accelerometers provide inclination sensing in response to gravity with respect to reference planes when the accelerometers rotate.

Wearables, encompassing diminutive and unobtrusive accelerometers often taking the form of wristwatches, hold the distinct advantage of seamlessly integrating into patients' routines. This integration is achieved without compromising patient safety,

comfort, or the smooth operation of ICU care procedures. This unobtrusive nature ensures that their presence does not disrupt the intricate operations concerning patient care within the ICU environment. Additionally, their cost-effectiveness is noteworthy, offering a readily available avenue for continuous sensory data capture. Its data, in turn, provides valuable insights spanning a spectrum of domains such as mobility patterns, sleep quality, overall comfort, and levels of sedation experienced by the patients [131]. As technology advances, accelerometers will likely play an even more significant role in enhancing healthcare delivery and patient outcomes.

Table 2.1 presents a comprehensive overview of recent studies conducted in the ICU that utilize accelerometers for various monitoring and assessment purposes. The recent application of accelerometers in the ICU encompasses a wide spectrum of healthcare applications, including delirium detection, sleep monitoring, acuity assessment, posture classification, activity recognition, and stroke detection.

The sampling rate of accelerometers, as indicated in Table 2.1 column 3, varies across studies. While some studies provide specific sampling rates (e.g., 10 Hz, 30 Hz, 100 Hz), many do not. The time sampling intervals (indicated in column 4) also differ significantly, ranging from as short as 1 second to as long as 24 hours. The sampling rate and interval choice is influenced by each study's specific requirements and the task's nature.

The number of accelerometers deployed (refer to column 5) in each study also varies, ranging from one accelerometer to multiple devices. For instance, studies aiming at assessing the motor state or activity recognition often employ multiple accelerometers placed at different body parts to obtain a comprehensive view of patient movements. In contrast, studies focused on simpler tasks such as step counting may use just one accelerometer.

Column 6 specifies the body positions where accelerometers are placed for data collection. These positions include wrists, ankles, elbows, upper arms, thighs, chest, and even under the pillow. The objective task determines the choice of placement. For example, wrist placement is common for sleep monitoring and activity recognition, while ankle placement is suitable for step counting.

Table 2.1 underscores the growing importance of accelerometer technology in the ICU. These devices offer healthcare professionals valuable insights into patient behavior, mobility, and activity levels. This information can aid in the early detection of delirium [3], assessment of neurological motor states [12], identification of sedentary behavior [52], and monitoring of sleep patterns [5], among other applications. Furthermore, accelerometers provide continuous data, allowing for real-time monitoring and more informed decision-making in critical care settings.

Table 2.1: Overview of the surveyed studies and their use of accelerometer. The table includes information on the study, the specific task or purpose in which the accelerometer was employed, the sampling rate of the accelerometer device, time sampling intervals (time windows used as samples), the number of accelerometers used, and the body positions where the accelerometers were placed. INP stands for "Information not provided".

Study	Task	Sampling rate (Hz)	Time sampling interval	Number of devices	Body position
[3]	Delirium detection	INP	1 min	2	Both wrists
[5]	Sleep monitoring	INP	24 hours	1	Non-dominant wrist
[12]	Neurological motor states and functional outcomes	10	5 sec	7	Each elbow, wrist, ankle and an additional sensor placed on the bed
[31]	Sitting and lying posture classification	10	2 sec	3	On the right ankle, on the wrist, and on the upper arm
[33]	Sleep monitoring	32	30 sec	1	Wrist
[34]	Sleep monitoring	32	30 sec	1	Wrist
[37]	Activity recognition	20	1 sec	2	The left midclavicular line and the right thigh
[45]	Step count	INP	24 hours	1	Ankle
[52]	Time spent in physical activity, sedentary behaviors, and time spent in activity and inactivity	90	5 sec	1	Right ankle
[54]	Rest-activity classification	INP	30 sec	1	Wrist
[44]	Activity recognition	100	2.56 sec	2	Both the upper arm and upper leg
[59]	Motor related potential	INP	24 hours	1	Hallux bone
[67]	Activity levels, non-zero activity levels, inactivity, and wear time as a proportion of restraint/restraint alternative time	INP	30 sec	1	Wrist
[74]	Rest-activity classification	INP	1 min	1	Wrist
[77]	Activity recognition	25	10 min	2	Chest right side and thigh
[81]	Activity recognition	30	1 min	2	Torso and thigh
[86]	Stroke detection	INP	15 min	2	Both wrists
[55]	Sedentary time estimation (lying in bed or in a sitting position)	INP	1 min	1	Belt placed on thighs or waist
[87]	Nighttime total sleep time and daytime activity ratio	INP	15 second	1	Dominant-wrist
[109]	Time and intensity of physical activity	INP	INP	1	Wrist
[113]	Percentage of time spent at different levels of physical activity (inactivity, light activity and moderate activity)	100	12 hours	1	Dominant-ankle
[121]	Acuity assessment	100	24 hours	1	Dominant-wrist
[125]	Sleep monitoring	30	1 min	2	Wrist and ankle
[128]	Body position	INP	10 to 30 seconds	1	Torso
[141]	Ballistocardiogram	200	1 min	1	Under the pillow

2.1.1 Pain Classification

Critically ill patients frequently experience pain, and despite the existence of clinical scoring systems to quantify pain, its assessment in intensive care units appears to be scarce and inconsistent [98]. Further, many critically ill patients are incapable of communicating clearly. Pre-existing factors such as native language differences, history of cognitive deficit, developmental disability, or certain psychiatric disorders might prevent commu-

nication with caregivers. Furthermore, interventions such as endotracheal intubation, tracheostomies, and medical sedation, coupled with the increased prevalence of delirium and altered mental status in ICUs, exacerbate patients' communication difficulties [56].

In patients where self-reporting of pain is possible, a visual analog scale (VAS) [13], numerator rating scale (NRS) [57], or Defense and Veterans Pain Rating Scale (DVPRS) [100] can be used. In non-verbal patients, pain levels are assessed by ICU nurses based on the Behavioral Pain Scale (BPS), Critical Care Pain Observation (CPOT), and Non-Verbal Pain Scales (NVPS) [21, 48, 93]. Due to the human component involved, visual pain assessment is infrequent and often inconsistent with established guidelines [69]. Automated pain detection using artificial intelligence may obviate this problem.

Pain diagnosis is critical in the ICU as several adverse outcomes are associated with underdiagnosed pain, including increased infection rate, prolonged mechanical ventilation, hemodynamic derangements, delirium, and compromised immunity [49]. Moreover, inadequate pain management can have serious physiological and psychological effects [50]. Research indicates that appropriate pain management, adequate analgesia, and less sedation can help reduce the number of days spent on the ventilator, increase mobility, and decrease the incidence of delirium and length of stay in the ICU [14]. Therefore, continuous and computerized pain assessment in the ICU could lead to real-time analgesic adjustment [32], thus improving patient care and outcomes.

The most common models to autonomously predict patients' pain using machine learning in the ICU have used videos of patients' facial expressions [24, 88]. Despite the high model performance, privacy concerns arising from the storage of sensitive patient information present a roadblock to the widespread clinical acceptance of these models. On the other hand, some studies have investigated the feasibility of pain detection using vital signs [7, 41]. However, research indicates that vital signs are not strong indicators of pain [47]. In contrast, accelerometers worn on patients' hands, ankles, and wrists can collect rich data concerning patient mobility associated with pain [30, 35].

2.1.2 Acuity Assessment

Acuity refers to the severity of a patient's condition, concomitant with the priority assigned to patient care in a critical care setting. Patients in the intensive care unit (ICU) exhibit volatile physiological patterns and the potential for developing life-threatening conditions in a short period. Therefore, the timely recognition of evolving illness severity is of immense value in the ICU. Swift and precise assessments of illness severity can identify patients requiring the administration of immediate life-saving interventions [121]. Furthermore, these assessments can guide collaborative decision-making involving patients, healthcare providers, and families in determining care goals and optimizing resource allocation [120]. Patient acuity is a foundational concept in critical care that ensures patient

Table 2.2: Comparative analysis of clinical features, accelerometer data, targeted clinical outcome, and learning models across the reviewed studies from the recent literature.

Study	Clinical Features	Accelerometer Features	Target	Learning Model
[121]	Vitals signs	minimum, maximum, mean, variance, standard deviation, immobile count, interquartile range (IQR), root mean square of successive differences (RMSSD), and standard deviation of RMSSD.	successful or unsuccessful hospital discharge	single-layer RNN
[78]	age and gender, vital signs and lab tests	-	Mortality	LR, KNN, GaussianNB, SVC, DT, RF, Adaboost, GBDT and LightGBM.
[96]	severity of illness scores	-	death during the current admission, central line placement, and CLABSI	logistic regression, gradient boosted trees, and multi-layer NN
[82]	age, sex, admission information, physical frailty, laboratory tests, vital signs, treatment, and urine output	-	Mortality	XGboost
[60]	Vital signs, lab tests, Age, Gender and ICU admission information	-	Sepsis	neural network plus Weibull-Cox survival model

needs are met with precision, safety, and efficiency. Accurate acuity assessments are crucial for guiding clinical interventions, optimizing staffing ratios, and ensuring the presence of adequately trained personnel to address the needs of high-acuity patients [68, 91]. From management and fiscal perspectives, an accurate understanding of in-patient acuity levels permits effective budgeting and resource allocation [40].

Traditional, manual, threshold-based scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE) [73], the Simplified Acute Physiology Score (SAPS) [76], Sequential Organ Failure Assessment (SOFA) [133], Modified Early Warning Score (MEWS) [46] and others, have been developed to predict the risk of mortality in ICU patients and, by extension, gauge the complexity of their care needs [73]. These tools evaluate physiological parameters, laboratory results, and other pertinent clinical information. However, static variable thresholds and additive scores have lesser predictive accuracy for outcomes of interest, and they tend to use a few rudimentary biomarkers to represent complex disease states.

Recent studies in clinical informatics have validated the effectiveness of automated machine learning methods by utilizing comprehensive data from Electronic Health Records (EHR) systems [25, 71, 80, 103, 119]. EHR encompasses a variety of patient-level data categories, including demographic information, diagnoses, procedures, vital signs, medications, and laboratory measurements. Table 2.2 surveys each paper's features, learning

models, and clinical outcome targeted. The studies showed that advanced algorithms using machine and deep learning techniques have proven superior to conventional bedside severity evaluations in predicting in-hospital deaths, which indicates immediate patient acuity [120]. However, these systems have limitations as they solely utilize physiological data from EHRs, neglecting crucial factors that could impact the patient. This includes environmental aspects (such as noise, light, and sleep quality) and behavioral elements (such as facial expressions indicating pain, agitation, or emotional state, as well as patient mobility and functional status).

To overcome these limitations, Davoudi et al. [29] explored the benefits of augmenting traditional ICU EHR-based data with continuous and pervasive sensing technology. The study gathered detailed information on ICU patients' activity levels, environmental factors, and behaviors by combining wearable, light, sound, and camera data. This multisensor approach provided a holistic perspective on patient care and monitoring, facilitating a thorough analysis of delirium classification in critical conditions. Wearable device data significantly contributed to the study's results by offering valuable insights into patients' activity levels, movement patterns, and functional status. The study shows that integrating wearable sensor data with other modalities enables a comprehensive assessment of patients' behaviors and conditions in the ICU, potentially leading to advancements in patient care and monitoring. Inspired by the positive impact of these novel clinical data streams, Shickel et al. [121] proposed to augment EHR data with continuous activity measurements via wrist-worn accelerometer sensors to predict hospital discharge status as a proxy for acuity. The study employs deep learning techniques, specifically single-layer recurrent neural networks (RNN) with gated recurrent units (GRU), to process sequential data and predict patient illness severity. The findings suggest that integrating pervasive sensing data with conventional EHR data can enhance real-time acuity estimation for critically ill patients. Furthermore, they propose that additional investigation and integration of even more innovative data streams could offer further benefits in this regard.

2.2 Accelerometer Data Pre-processing

Accelerometer data can be noisy and complex, making it necessary to preprocess the data before extracting meaningful information. Accelerometer preprocessing involves several steps to clean, filter, and prepare the data for analysis or further processing. One significant advantage of preprocessing is noise reduction. Accelerometer data often contains noise due to various factors, such as environmental vibrations or device malfunctions, which can obscure the true signal. By applying filters and smoothing techniques, preprocessing can enhance the signal-to-noise ratio, thereby improving the accuracy of subsequent analyses. Another advantage is the standardization and normalization of data. Different accelerometer devices might have varying ranges and sensitivities, and prepro-

Table 2.3: Data Processing and Machine Learning algorithms used by recent papers

Study	Data Processing	Machine Learning
[3]	×	Random Forests, SVM, XGBoost
[12]	Butterworth high-pass filter and missing data imputation	Logistic regression
[31]	Reduced sampling rate and data imbalance treatment (undersampling and SMOTE)	Random forest model
[33]	×	Linear regression
[34]	×	Linear regression
[37]	Three-point median filter, low pass infinite impulse response filter, integral of the body accelerations	Decision tree
[54]	Logarithmic transformation	Linear regression
[44]	High-pass filter	Artificial neural network
[67]	×	×
[63]	×	Linear mixed effects and linear regression
[77]	Low-pass filter	×
[81]	×	Logistic Regression, Bagging Decision Tree, and SVM
[86]	Magnitude of first-derivative of acceleration	×
[55]	×	Multivariable linear regression model
[87]	×	×
[109]	Energy expenditure	Linear regression
[121]	Signal magnitude	RNN with GRU units
[125]	×	Repeated measure mixed model
[128]	×	×
[141]	Butterworth bandpass filter	×

cessing ensures that the data from multiple sources is comparable. Additionally, preprocessing can involve feature extraction, simplifying the data by focusing on the most relevant aspects, reducing the computational burden, and enhancing model performance.

However, there are also disadvantages to consider. Preprocessing can be time-consuming and computationally intensive, especially with large datasets. The process requires careful parameter selection, and inappropriate choices can lead to losing important information or introducing biases. For instance, overly aggressive noise reduction might remove subtle but significant patterns in the data. Moreover, preprocessing steps such as filtering and normalization can vary widely between studies, making comparing results across different research projects challenging. Despite these challenges, effective preprocessing is essential for maximizing the utility of accelerometer data and ensuring robust and reliable outcomes in research and practical applications. Here is an overview

of the key steps involved in accelerometer data preprocessing:

- **Data Collection** The first step in accelerometer data preprocessing is data collection. Accelerometers can measure acceleration in three axes (X, Y, and Z), and data are typically collected as a time series. The sampling rate and sensor placement are important considerations during data collection. Table 2.1 shows a comprehensive overview of sampling rate and sensor placement used in recent works.

- **Filtering** Filtering is crucial in accelerometer preprocessing. Low-pass filters are often applied to remove high-frequency noise, while high-pass filters can eliminate low-frequency drift. Bandpass filters may be used to isolate specific frequency ranges of interest. The choice of filter type and cutoff frequencies depends on the specific application. As shown in Table 2.3, Butterworth (both high-pass and bandpass) and low-pass filters have been employed in the recent literature [12, 44, 77, 141]) to process the data.

- **Normalization** Normalization is the process of scaling the accelerometer data to ensure that it is on a consistent scale. This step can be essential when comparing data from different sensors or experiments [9]. For deep learning, a common practice is normalizing data between 0 and 1, improving the performance and convergence of deep learning.

- **Integration** Dikkema et al. [37] employed integration over body accelerations to obtain velocity information from acceleration data. The typical approach is to perform numerical integration, which involves cumulatively summing the acceleration values over time to calculate velocity or displacement. However, this approach is highly prone to drift.

- **Segmentation** In some applications, it may be necessary to segment the accelerometer data into meaningful chunks or events. For example, one might want to segment data into activities such as walking, running, or sitting in activity recognition.

- **Feature Extraction** Feature extraction involves computing features from the preprocessed data to capture relevant information. Common features include mean, standard deviation, peak values, and spectral characteristics. Table 2.4 shows a careful survey of the features used by recent works.

- **Data Alignment** When working with data from multiple sensors or time series from different sources, data alignment ensures that the data is synchronized and corresponds to the same time intervals.

- **Data Imbalance** Data imbalance refers to a dataset's disproportionate representation of classes. It can be a challenge in machine learning, as models might be biased

towards the dominant class. Davouldi et al. [31] treated data imbalance using the techniques of undersampling and the Synthetic Minority Over-sampling Technique (SMOTE).

2.3 Artificial Intelligence

Artificial intelligence can potentially revolutionize the healthcare field, particularly in using accelerometers [65]. AI aims to mimic human cognitive functions and is powered by the increasing availability of healthcare data and rapid progress in analytics techniques [65]. The use of AI in healthcare applications, including the use of accelerometers, offers practical benefits and replicates human intellectual functions [1].

AI has the ability to transform various aspects of patient care, administrative processes, and nursing information systems. It can assist in monitoring and documenting patient data, managing quality, enhancing care efficacy, and performing interventions appropriately [70]. AI applications in nursing information systems can be cost-effective time-saving, and improve the documentation of patient information [70].

In the field of clinical medicine and medical research, AI can help minimize the scarcity of human resources and broaden the role of humans in healthcare. It can potentially improve patient outcomes by enabling faster prognosis and diagnosis of diseases [137]. AI algorithms can analyze large volumes of unstructured data and provide rapid, actionable deductions, leading to faster decision-making in emergency medicine and other healthcare-related applications [94]. However, there are certain gaps and challenges in the application of AI in healthcare. Published research in real-world settings often lacks a prominent role in supporting care-dependent individuals, patient education, and the health of caregivers or nurses [23]. Additionally, there is a need for more studies that assess the effects of AI on clinical and organizational outcomes [23]. Despite these challenges, the field of AI in healthcare is rapidly growing. A bibliometric analysis of healthcare-related AI publications revealed a significant increase in the literature, highlighting the critical driving power of AI in promoting healthcare [53]. Patients' perceptions toward human-AI interaction in healthcare also play a crucial role in the widespread use of clinical AI [42].

2.3.1 Classical Learning Models

Classical machine learning models offer several advantages, such as decision trees, support vector machines, and logistic regression. They are often easier to interpret and understand than complex models like deep neural networks, making them suitable for applications where model transparency is essential. These models typically require less computational

Table 2.4: Features extracted by recent studies in the literature

Study	Features
[3]	Minutes at rest and within-patient dynamic time warping
[12]	The proportion of dynamic activity, signal magnitude area, median of high-pass filtered signal, median of low-pass filtered signal, median frequency according to Fourier transform coefficients, frequency-domain entropy, band power, and level 2–6 detail coefficients of the 5th-order Daubechies wavelet transform
[31]	Vector of magnitude (mean, standard deviation, covariance, skewness, kurtosis, entropy, coefficient of variation, dominant frequency, percentage of the power that is in 0.6-2.5 Hz, dominant frequency/sum of moduli at each frequency), mean angle of acceleration relative to vertical on the device, standard deviation of the angle of acceleration relative to vertical on the device, and correlation between axis
[33]	Mean and median of activity counts
[34]	Mean, standard deviation, range, median and interquartile range of activity counts
[37]	Horizontal and vertical body angles and a signal magnitude area
[54]	Time of day of highest activity, mean activity level of the fitted curve, the difference between mesor and peak activity, and goodness-of-fit (R2) of the regression model
[44]	Mean, Minimum, Maximum, Median, Standard Deviation, Coefficients of variation, peak-to-peak amplitude, percentiles, interquartile range, pitch angle, roll angle, median crossings, skewness, kurtosis, signal power, root mean square, peak intensity, Pearson correlation coefficient, inter-axis cross-correlation, autocorrelation, trapezoidal numerical integration, signal magnitude area, signal vector magnitude, power spectral density
[67]	Activity levels, non-zero activity levels, inactivity, and wear time as a proportion of restraint time.
[63]	Total nighttime and maximum nighttime activity counts, sleep duration, average sleep bout length, total number of sleep bouts, total daytime and maximum daytime activity counts, and daytime sleep.
[77]	Mean acceleration, standard deviation, inclination of the x-axis, forward/ backward angle of the thigh
[81]	Metrics (mean, maximum, minimum, standard deviation, median, and entropy) for signal magnitude, magnitude of low-frequency components, high-frequency components, and the derivation of each high-frequency component
[86]	Kolmogorov-Smirnov statistic
[55]	Time spent in mobility positions (standing, sitting, lying), time spent in activity intensities (sedentary, moderate, light, vigorous), step counts, steps per minute, and kcal per day consumption
[87]	Nighttime total sleep time, sleep efficiency, daytime activity, hourly activity counts
[109]	Frequency analysis, counts and percentages, means and standard deviations
[121]	Minimum, maximum, mean, variance, standard deviation, immobile count, interquartile range, root mean square of successive differences (RMSSD), and standard deviation of RMSSD
[125]	Total sleep duration, amount of time spent sleeping in minutes, and the number of transitions from sleep to waking
[141]	2nd degree polynomial regression of normed x and z axis

power and training time, making them accessible for use in environments with limited resources. Additionally, classical models can perform well on smaller datasets, whereas deep learning models might overfit or fail to generalize effectively. However, classical machine learning models also have their disadvantages. They may struggle to capture intricate patterns in large and high-dimensional datasets, where more sophisticated techniques might excel. Furthermore, their performance can be limited by the need for extensive feature engineering, requiring domain expertise to extract relevant features from raw data.

In the recent literature review, linear regression appears to be the dominant modeling approach, evidenced by its adoption in numerous studies [33, 54, 109]. Furthermore, there is a noticeable inclination toward logistic regression and multivariable linear regression models, which is explained since the primary aim of these works revolves around predicting continuous values or classifying data into binary categories.

Diving deeper into the modeling techniques, a subset of recent research has embraced tree-based models. Studies such as Ahmed et al. [3], Davoudi et al. [31], and Dikkema et al. [37] have incorporated algorithms like random forests, decision trees, and XGBoost. These models are celebrated for their transparency, making them favorable choices when deducing the significance of individual features is imperative.

Logistic Regression [28] is one of the foundational methods for binary and multiclass classification. It derives its name from the logistic function used to model the probability that a given instance belongs to a particular class. This linear approach assumes that there is a linear relationship between the input features and the log-odds of the output. The coefficients of the linear equation are optimized using techniques like Maximum Likelihood Estimation. It is a straightforward and interpretable algorithm, making it a first choice for many simple classification tasks and for instances when a basic understanding of the relationship between variables is sought. However, it may struggle with non-linear data and can be outperformed by more sophisticated algorithms in complex scenarios.

Moving from linear to ensemble methods, XGBoost [22] stands out. XGBoost, which stands for eXtreme Gradient Boosting, is an optimized gradient boosting library. Gradient boosting involves building trees sequentially, where each new tree tries to correct the errors of its predecessor. XGBoost takes this approach to another level by focusing on computational speed and model performance. It handles missing data, provides facilities for regularization to avoid overfitting, and can be used for both classification and regression tasks. The flexibility and efficiency of XGBoost have made it a dominant choice in machine-learning competitions and real-world applications.

CatBoost [39], short for Categorical Boosting, is a relatively newer entrant in the gradient boosting arena. As the name suggests, CatBoost shines when dealing with categorical features. While most algorithms require categorical data to be preprocessed into a numerical format, CatBoost handles categorical variables natively, reducing the

need for extensive preprocessing and mitigating the risk of data leakage. Additionally, it employs an ordered boosting approach, which helps in reducing overfitting. Its built-in support for visualization also allows users to understand and interpret model performance and feature importance easily.

2.3.2 Deep Learning Models

Deep learning models, particularly neural networks, offer significant advantages in handling complex and high-dimensional data. They have the capability to automatically learn features from raw data through multiple layers of abstraction, reducing the need for extensive manual feature engineering. This makes them highly effective in tasks such as image and speech recognition, natural language processing, and complex pattern recognition, often achieving state-of-the-art performance. Deep learning models can also scale well with large datasets, improving their performance as more data becomes available. However, these models come with notable disadvantages. They require substantial computational resources, including powerful GPUs and large amounts of labeled data for training, which can be costly and time-consuming to obtain. Deep learning models are often seen as black boxes due to their complexity, making them difficult to interpret and understand, which can be a limitation in applications requiring transparency and explainability. Additionally, they are prone to overfitting if not properly regularized and can be challenging to fine-tune, necessitating expertise in hyperparameter optimization and architecture design. Despite these challenges, deep learning remains a powerful tool in the modern data science toolkit, driving advancements in various fields.

Shickel et al. [121], for example, utilizes recurrent neural networks equipped with GRU units since they work with time series data. Similarly, FR et al. [44] exploits artificial neural networks, highlighting their capability to delineate intricate data relationships. Transitioning to more nuanced modeling approaches, Jaiswal et al. [63] and Smichenko et al. [125] have adopted mixed-effect models. These models excel in situations where both fixed and random effects are at play. They are especially favored in longitudinal analyses, wherein data samples are hierarchically structured, such as patients grouped within hospitals.

In our acuity assessment approach, we evaluated five different neural network architectures, namely, VGG, ResNet, MobileNet, SqueezeNet, and a custom Transformers network, as both Convolutional Neural Networks (CNNs) and Transformers architectures are well-researched in the sensor-based human activity recognition field [38, 66, 89, 104, 105, 138]. Due to their architectural advantages, CNNs are particularly adept at extracting features from accelerometer data. Their design promotes local connectivity, making them proficient in recognizing short sequences of time-series data and specific motion patterns. The shared weights in CNNs enable them to detect

patterns regardless of their position in the sequence, while the hierarchical structure allows for extracting both simple and complex movement patterns. On the other hand, Transformers are advantageous for processing accelerometer data due to their self-attention mechanism, which aptly captures dependencies in time-series data. It allows the model to weigh the importance of different elements in a sequence when generating outputs [126]. Consequently, each patient's movement can be contextualized in relation to the other movements within a time window, directing the network's attention to the key movement patterns for assessing the patient's condition. In the following, we briefly discuss each of the evaluated architectures.

Developed by the Visual Geometry Group at Oxford University, the VGG architecture [124], specifically VGG-16 and VGG-19, relies heavily on deep layers consisting of small (3×3) convolutional filters. This design increases the depth of the network, allowing it to learn more intricate features without drastically increasing computational requirements compared to larger filter sizes. While VGG has demonstrated excellent performance on several benchmarks, it is computationally expensive and memory-intensive, making it less suitable for deployment on devices with limited resources.

The Residual Network (ResNet) was introduced by He et al. [58], addressing the vanishing gradient problem that plagued deeper networks. The architecture's uniqueness lies in its skip connections or residual connections, which bypass one or more layers, allowing the network to learn identity functions. As a result, ResNet can be trained deeper without the diminishing returns on accuracy often observed in other deep networks. These deep networks have achieved state-of-the-art results on numerous benchmarks. However, despite its depth, the computational cost might be prohibitive for real-time applications or edge devices.

Squeezenet [62] is a lightweight deep neural network architecture designed for efficient inference on resource-constrained devices. It achieves high accuracy with a significantly smaller model size than other architectures. Squeezenet reduces the number of parameters by using 1×1 convolutions to squeeze the input channels and expand them back using 1×1 and 3×3 convolutions. This approach results in a drastically reduced model size, making SqueezeNet particularly appealing for deployment on devices with limited computational capabilities or storage.

Mobilenet [61] is another lightweight deep neural network architecture that is specifically designed for mobile and embedded vision applications. It uses depthwise separable convolutions to reduce the number of parameters and computations. Depthwise separable convolutions split the standard convolution into a depthwise convolution and a pointwise convolution, reducing the computational cost while maintaining accuracy. MobileNet has various versions, with the later versions optimizing accuracy and efficiency trade-offs, making them ideal for real-time applications on resource-constrained devices.

Transformer [130] is a revolutionary architecture that has gained significant at-

tention in natural language processing tasks, particularly machine translation. Unlike convolutional neural networks, which rely on convolutional and pooling layers, the Transformer architecture is based on self-attention mechanisms. Self-attention allows the model to weigh the importance of different words in a sentence when generating translations. Its versatility and performance have driven its adoption in various applications, from machine translation to image classification. However, its high computational cost, especially regarding memory usage for longer sequences, can be limiting.

2.4 Interpretability

As artificial intelligence systems are increasingly used in healthcare delivery, it is important to understand and interpret the decisions made by these systems [79]. Explainability refers to the ability to provide understandable and transparent explanations for the decisions and predictions made by AI models [112].

In the context of healthcare, explainability is essential. It helps build trust and acceptance among healthcare professionals and patients, as they can understand the reasoning behind AI-driven recommendations or diagnoses [99]. This is particularly important in critical areas such as diagnosis and treatment planning, where the consequences of AI decisions can significantly impact patient outcomes [43].

Explainability also plays a crucial role in ensuring accountability and ethical considerations in the application of AI in healthcare [101]. It allows healthcare professionals to assess the reliability and validity of AI models, identify potential biases or errors, and make informed decisions based on AI-generated insights [139]. Additionally, explainability enables regulatory bodies to evaluate the safety and effectiveness of AI systems and ensure compliance with ethical standards [106]. However, achieving explainability in AI models can be challenging, especially in complex deep learning models that operate as black boxes [112]. Researchers have been exploring various methods and techniques to enhance the interpretability of AI models, such as generating explanations based on feature importance, rule extraction, or visualization techniques [79]. These approaches aim to provide understandable explanations that healthcare professionals can easily interpret and validate. Examples of these models are SHAP analysis, DeepLIFT, and Gradient SHAP.

SHAP (SHapley Additive exPlanations) [83] analysis is a method that provides a unified framework for explaining the output of any machine learning model. It is based on the concept of Shapley values from cooperative game theory, which assigns a value to each feature in a prediction based on its contribution to the prediction outcome. SHAP analysis provides a global explanation of the model by quantifying the importance of each feature in the overall prediction. It can help identify which features significantly impact the model's decision-making process [20].

DeepLIFT (Deep Learning Important FeaTures) [123] works by attributing the difference between the model's output for a given input and a reference output to the input features. DeepLIFT assigns importance scores to each feature, indicating their contribution to the model's prediction. It provides a local explanation for individual predictions, allowing users to understand why a particular prediction was made [134].

Gradient SHAP [83] is a variant of SHAP analysis that combines the concept of SHAP values with the gradient-based approach. It calculates the SHAP values by approximating the gradients of the model's output with respect to the input features. Gradient SHAP provides a more efficient and scalable way to compute SHAP values for deep learning models. It allows for both global and local explanations, enabling users to understand the overall behavior of the model as well as the reasoning behind individual predictions [134].

The aforementioned explainability techniques have been applied in various healthcare domains. For example, in a study on predicting long-term mortality in critically ill ventilated patients, SHAP analysis was used to provide explanations of the entire model and individual features, aiding in the interpretation of the deep learning model's predictions [20]. Another study employed Gradient SHAP to interpret the correlation between risk features and outcomes in an ICU mortality prediction model [107]. However, it is important to note that explainability techniques are not without limitations. Deep learning models are highly complex, and their explanations may not always be intuitive or easily understandable to non-experts. Interpreting SHAP values, DeepLIFT scores, or Gradient SHAP may require domain expertise and further validation. Additionally, the explanations provided by these techniques are post-hoc and do not guarantee a complete understanding of the model's internal workings [111].

Chapter 3

Intelligent ICU System

This chapter presents the Intelligent ICU system and the data processing pipeline. The system is responsible for capturing and saving the data used in this work. The author of this thesis developed and executed the subsequent steps, including, but not limited to, curating, cleaning, normalizing, labeling, and segmenting the data.

The Intelligent ICU system was created by Iheal Laboratory[1] and was first described in the work of Davoudi et al. [29]. The system, as illustrated in Figure 1.1, collects clinical, imaging, wearable, environmental, and physiological data from the ICU. The clinical data includes details such as patient drugs, laboratory findings, age, gender, and key care metrics. Imaging data captures visual information such as RGB videos of patients' faces and depth images with a view of the entire ICU room. Physiological data comprises vital signs, heart rate, blood pressure, temperature, and respiratory rate. Wearable data includes accelerometers, gyroscopes, and electromyography (EMG) sensors. Finally, environmental factors detail the ICU's ambient noise, luminosity, and air purity. This dissertation focused on the accelerometer, physiological and clinical data.

The data captured and saved by the Intelligent ICU system was raw and heterogeneous, originating from various devices with different configurations and saving methods, accumulated by different personnel over the years. This variability necessitated a careful and comprehensive understanding of the data. During her sandwich doctorate, the author spent nine months working closely with the clinical staff at the University of Florida. This proximity allowed her to engage in detailed discussions with the clinical staff who recruited the patients. She gained a comprehensive understanding of the entire data collection pipeline, including how the accelerometers were applied to the patients and the specifics of data storage. These conversations were crucial for understanding how device configurations could impact the final data. Subsequently, the author meticulously curated, cleaned, and pre-processed the data, identifying and addressing any inconsistencies or errors introduced during the data collection process. In the next sections, we delve into the steps of patient recruitment (Section 3.1), data collection (Section 3.2), data curation and normalization (Section 3.3), and data labeling and segmentation (Section 3.4).

[1] https://www.bme.ufl.edu/labs/rashidi/

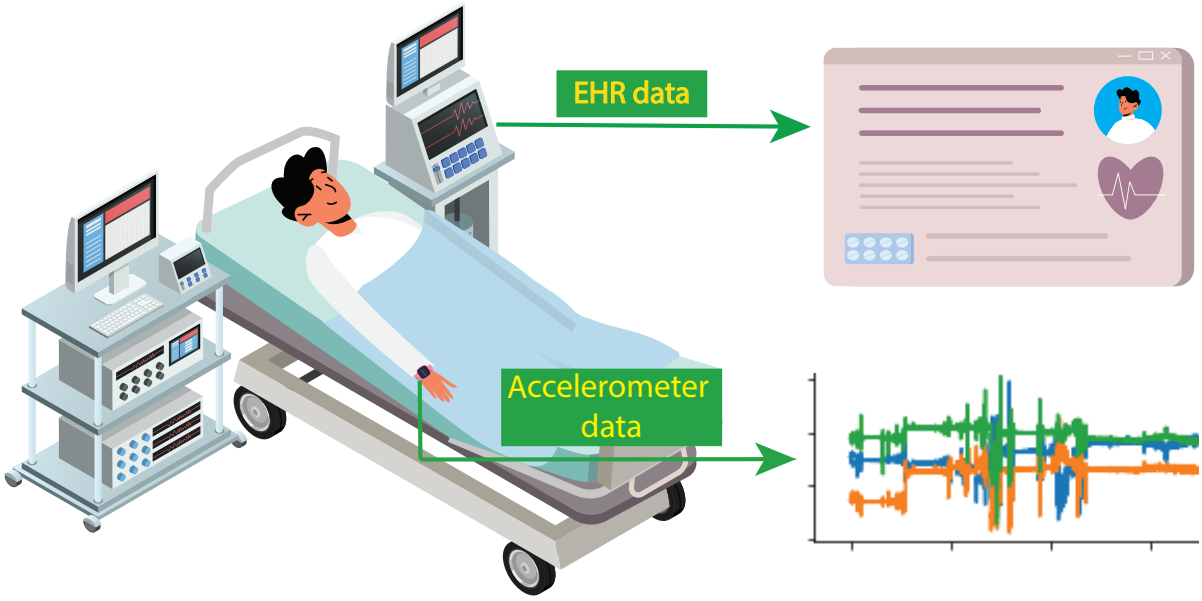


Figure 3.1: Data used in this dissertation consisted of physiological, clinical, and accelerometer data collected from ICU patients of the University of Florida Shands Hospital.

3.1 Patient Recruitment

The data used in this research was sourced from adult patients admitted to the surgical ICUs at Gainesville's University of Florida Shands Hospital, following compliance with all relevant federal, state, and local laws and regulations. Approval for the study was granted by the University of Florida Institutional Review Board (IRB) under the following numbers: IRB201900354 and IRB202101013. Before enrolling patients in the study, written informed consent was obtained from all participants. In cases where patients could not provide informed consent, consent was secured from a legally authorized representative acting on their behalf. Eligible participants were individuals aged 18 and above admitted to the ICU and expected to remain there for at least 24 hours. Exclusion criteria encompassed patient transfers, discharges, deaths occurring within 24 hours of recruitment, and individuals necessitating isolation, contact precautions, or lacking informed consent, whether from the patient or their legal representative.

IRB201900354: Data and informed consent were collected from 71 patients between June 2021 and December 2021. These patients wore Shimmer3 accelerometers on their wrists, which collected data from 100 Hz to 512 Hz sampling rates. The average data collection length was 4.61 days. Their biological sex composition was 59% male and 41% female, with a mean age of 57.5 years.

IRB202101013: Data and informed consent were collected from 190 patients between January 2022 and February 2023. These patients wore Shimmer3 with 30 Hz, 100 Hz, or 512 Hz sampling rates. The mean data collection duration was 4.46 days. Patient

Table 3.1: Clinical characteristics of the patient cohort.

Variables	Patients (N=128)
Female sex, N(%)	41 (32.03)
Age in years, mean (SD)	59.31 (16.22)
Height in cm, mean (SD)	172.97 (10.64)
Weight in kgs, mean (SD)	87.09 (25.43)
Length of stay in days, mean (SD)	22.47 (27.28)
Race, White(%)/ Black (%)/ Other (%)	83%/ 11%/ 6%
Cancer, N(%)	17 (13.28)
Cerebro-vascular, N(%)	15 (11.71)
Dementia, N(%)	6 (4.68)
Paraplegia Hemiplegia, N(%)	5 (3.90)
Congestive Heart Failure, N(%)	22 (17.18)
Chronic Obstructive Pulmonary Disease, N(%)	20 (15.62)
Diabetes, N(%)	35 (27.34)
Metastatic Carcinoma, N(%)	7 (5.46)
Liver, N(%)	28 (21.87)
Peptic Ulcer, N(%)	7 (5.46)
Renal, N(%)	30 (23.43)
Rheumatologic, N(%)	2 (1.56)

recruitment is ongoing. The biological sex composition of these participants was 58% male and 42% female, with a mean age of 59.0 years.

In our pain classification analysis, we used data from 128 patients. Table 3.1 presents the demographic and clinical variables distribution for the patients included in these analyses. The majority of study participants were elderly, white individuals with comorbidities. Our acuity assessment analysis contained data from 87 patients. The demographic and clinical characteristics of the patients analyzed are detailed in Table 3.2, while Table 3.3 provides a demographic breakdown categorized by stable and unstable conditions. The distribution of patients by race and gender was approximately equivalent in both the development and test sets. The average age was marginally higher in the development cohort, though not significantly. Heights were similar between the cohorts, whereas the mean weight was notably greater in the development cohort. The length of stay did not significantly differ between cohorts. Notable differences in disease prevalence included a higher incidence of cancer and diabetes in the test cohort, while liver-related diseases were more prevalent in

Table 3.2: Acuity cohort characteristics

Variables	Development Cohort (N=60)	Test Cohort (N=27)	p-value
Female sex. N (%)	22 (36.7%)	9 (33.3%)	0.76
Hispanic ethnicity, N (%)	8 (13.3%)	2 (7.4%)	0.42
Age in years, mean (SD)	58.4 (15.9)	52.2 (18.3)	0.12
Height in cm, mean (SD)	173.6 (9.1)	172.4 (8.5)	0.56
Weight in kgs, mean (SD)	87.2 (23.6)	77.8 (15.0)	0.06
Length of stay in days, median (25th, 75th percentile)	11.0 (6.0, 29.0)	13.0 (8.0, 23.0)	0.60
Race: N (%)			
White	49 (81.7%)	18 (66.7%)	0.12
African American	9 (15.0%)	3 (11.0%)	0.63
Other	2 (3.3%)	6 (22.2%)	<0.05
Comorbidities:N (%)			
Cancer	0 (0.0%)	6 (22.2%)	<0.05
Cerebrovascular disease	8 (13.3%)	4 (14.8%)	0.85
Dementia	1 (1.7%)	2 (7.4%)	0.18
Paraplegia hemiplegia	6 (10.0%)	2 (7.4%)	0.70
Congestive heart failure	7 (11.7%)	2 (7.4%)	0.55
Chronic obstructive pulmonary disease	4 (6.7%)	3 (11.1%)	0.48
Diabetes	7 (11.7%)	6 (22.2%)	0.20
Liver disease	15 (25.0%)	5 (18.5%)	0.51
Peptic ulcer	2 (3.3%)	0 (0.0%)	0.34
Renal disease	9 (15.0%)	4 (14.8%)	0.98

Abbreviation: SD, standard deviation; N, number.
Notes: Our analysis employed two distinct statistical tests to examine the differences between the development and test cohorts. We used Welch's t-test for the continuous variables, while we used the two-proportion z-test for the categorical variables.

the development cohort.

3.2 Data Collection

Figure 3.1 depicts the data sources used in this dissertation: EHR data and accelerometer readings. Patients wore either Shimmer3 [122] Inertial Measurement Units (IMU) on one

Table 3.3: Distribution of demographic variables of encounters (recorded every four hours) stratified by class labels (stable, unstable)

Variables	Stable Encounters (N=434)	Unstable Encounters (N=101)	p-value
Female sex, N (%)	187 (43.0%)	16 (15.8%)	<0.05
Hispanic ethnicity, N (%)	64 (14.8%)	4 (4.0%)	<0.05
Age in years, mean (SD)	59.2 (16.7)	57.5 (13.4)	0.34
Height in cm, mean (SD)	171.2 (9.1)	178.6 (7.2)	<0.05
Weight in kg, mean (SD)	83.6 (21.50)	97.5 (20.2)	<0.05
Length of stay in days, median (25th, 75th percentile)	16.0 (8.0, 31.0)	29.0 (12.0, 33.0)	<0.05
Race, N (%)			
White, N (%)	336 (77.4%)	85 (84.2%)	0.97
African American, N (%)	42 (9.7%)	16 (15.8%)	0.07
Other, N (%)	56 (100.0%)	0 (0.0%)	<0.05

Abbreviation: SD, standard deviation; N, number.

of their wrists. The accelerometers in the devices used in this study capture the direction and magnitude of acceleration along 3 axes. The IMU devices convey information on the patient's arm's direction and intensity of movement as well as rotational position through continuous measurement of the device's linear acceleration and angular velocity. These devices capture various aspects of movement and activity, offering insights into physical dynamics such as speed, direction, and motion intensity. These measurements enable the quantification of movement patterns and activity levels with high precision and detail. In this work, we did not include clinical information reflected at the motor level, such as assessments of muscle strength, coordination, balance, and overall mobility. Accelerometer readings were taken for a maximum of 7 days or until the patient's discharge from the ICU, whichever came first. During this time, the study team performed daily visits to ensure the device was correctly positioned on the patient's wrist and requested that the nursing staff document any time the device was removed. All known removal and reapplication times were documented as device downtimes to be excluded from the analysis. Conservative estimations were used if the exact removal time was unknown.

Using a daily pipeline, UF's Integrated Data Repository service extracted clinical data relevant to the patient's acuity state from the EHR. This information included demographics such as age, sex, race, height, weight, length of stay, medications, and physiological signals like blood pressure, heart rate, oxygen saturation (SpO2), respira-

tory device, continuous renal replacement therapy, blood transfusion, pain score, Braden score [11] and acute brain dysfunction status (whether the patient was in a coma, experiencing delirium, or had normal cognitive status) [108].

The presence of existing intravenous lines, wounds, and patient objection to wearable devices are examples of cases where motion data could not be collected. For the purposes of this analysis, we have excluded patients for whom accelerometer data collection or retrieval was not possible. A user-friendly touchscreen interface allowed nurses and caregivers to stop data collection anytime. Nurses were instructed to remove the devices for bathing and medical procedures, if necessary, and to replace the devices afterward. Data were captured on a local secure computer throughout the patient enrollment period and transferred to a secure server for analysis upon patient discharge.

3.3 Data Curation and Normalization

In accordance with best practices for data-driven models, we incorporated a data cleaning and preparation step in our processing pipeline. Initially, we filtered the data to exclude periods when the patient was not using the sensor, as determined by the downtime annotations made by the nurses. These annotations provided a reliable source for identifying non-use periods, ensuring that only valid data points were included in our analysis.

Occasionally, the sensor may not capture data accurately. We identified these instances of missing data and implemented strategies to rectify them, ensuring that our dataset remains complete and consistent. For accelerometer data, we included only data from contiguous windows without missing values (for window size, refer to Section 3.4). For EHR data, we employed the nearest neighbor method to impute missing values, with the time frame for this process varying based on the nature of the data. For example, we utilized a 2-hour window for heart rate, whereas we adopted a 12-hour window for brain status. In cases where accurate imputation was not feasible, the data was excluded from the dataset.

Standardization of the accelerometer data was necessary due to the varying sampling frequencies depending on the device or time of collection. Thus, accelerometer data were standardized to common sampling frequencies of 10 Hz for acuity analysis and 32 Hz for pain classification. The choice of 10 Hz for acuity measurements was driven by the need to manage the size of the 4-hour windows. This standardization ensures uniformity in the input data rate, facilitates more accurate analysis, and prevents excessively long sequences. The acuity sequence size was 144,000 (14,400 seconds x 10 Hz) and 28,800 (900 seconds x 32 Hz) for pain classification. Additionally, on a sample-wise basis, accelerometer values were normalized to a range of [0,1] using Equation 3.1 with min and max values calculated per sample.

$$X_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{3.1}$$

Clinical data features were encoded based on their data type. Numerical features were standardized using the Z-Score as defined in Equation 3.2.

$$z = \frac{x - \mu}{\sigma} \tag{3.2}$$

Z-score standardization involves rescaling the features with a mean of zero and a standard deviation of one. This transformation ensures that numerical features contribute equally to the model, preventing features with larger scales from dominating the learning process.

Categorical features, on the other hand, were encoded using the one-hot encoding technique. One-hot encoding transforms categorical variables into a binary matrix, with a unique binary vector representing each category. This encoding technique preserves the categorical nature of the data while making it suitable for inclusion in machine-learning models that require numerical input.

Analgesics were initially one-hot encoded by the medication name. If a patient received an analgesic, we assessed the proportion of time during a 15-minute interval when the patient was under the influence of the analgesic. This information was calculated based on the time of the last dose and the half-life of the medication in question.

Lastly, only wrist-related data was used to maintain consistency and specificity.

3.4 Data Labeling and Windows Segmentation

In this dissertation, we employed supervised machine learning algorithms. This family of algorithms learns the relationship between data and a label. In our case, the label is a patient's acuity state or the pain level in a certain time range. These labels were derived from clinical assessments documented in the ICU records, ensuring each data point was paired with a corresponding label. The data was systematically segmented into time windows to facilitate accurate labeling and analysis.

3.4.1 Pain Classification

The pain label is a self-reported score obtained using the Defense and Veterans Pain Rating Scale (DVPRS). The DVPRS is a comprehensive tool designed to assess pain levels in military and veteran populations. It consists of an 11-point numerical scale, ranging from 0 (no pain) to 10 (worst possible pain), allowing for a detailed representation of pain intensity. The DVPRS incorporates visual and verbal descriptors to enhance its accuracy and usability. Alongside the numerical ratings, the scale includes color-coded faces and

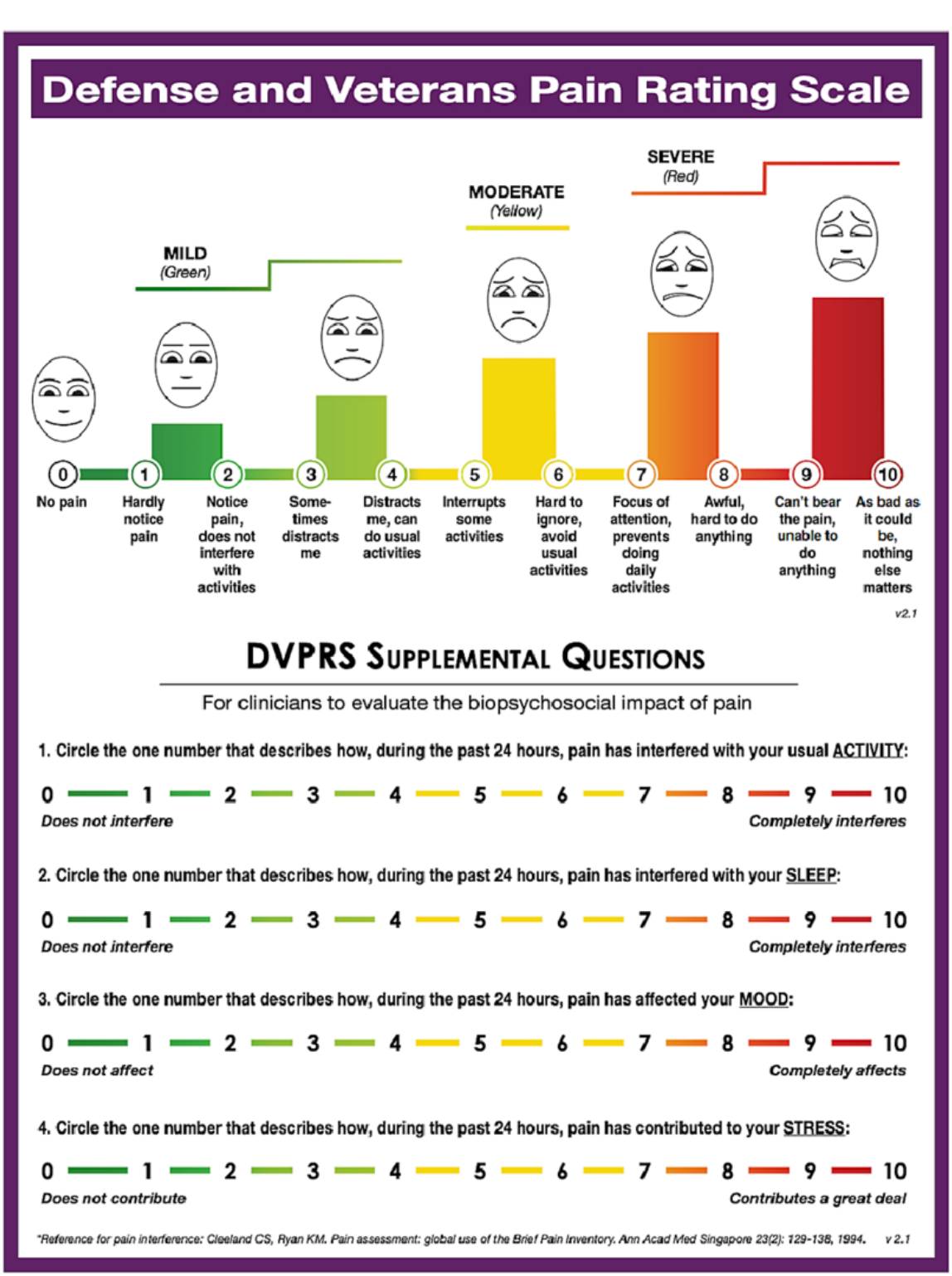


Figure 3.2: Defense and Veterans Pain Rating Scale (DVPRS). Figure extracted from Polomano et al. [100].

word descriptors correlating with pain intensity, providing a multidimensional approach to pain assessment. Additionally, the DVPRS encompasses functional impact questions that evaluate how pain affects various aspects of daily life, such as activity, sleep, mood, and stress. An illustration of the DVPRS form is shown in Figure 3.2. This holistic approach ensures that the scale measures the intensity of pain and its broader impact on the individual's quality of life. The DVPRS is typically administered through patient self-report, either verbally or in written form, allowing patients to communicate their pain levels in a clear and standardized manner.

In this work, we evaluated the viability of predicting the DVPRS pain scores, as well as groups of scores and the pain variation. Let P represent the pain score. Pain scores were divided into four groups, as shown in Equation 3.3:

$$\text{Pain Group} = \begin{cases} \text{no pain} & \text{if } P = 0 \\ \text{mild pain} & \text{if } P \in \{1, 2, 3, 4\} \\ \text{moderate pain} & \text{if } P \in \{5, 6\} \\ \text{severe pain} & \text{if } P \in \{7, 8, 9, 10\} \end{cases} \tag{3.3}$$

Pain variation was calculated by taking the current pain score, P_{current}, and subtracting the previous pain score, P_{previous}.The difference ΔP is defined as follows:

$$\Delta P = P_{\text{current}} - P_{\text{previous}} \tag{3.4}$$

The pain variation based on the value of ΔP is given by:

$$\text{Pain variation} = \begin{cases} \text{same} & \text{if } \Delta P = 0 \\ \text{decreased} & \text{if } \Delta P < 0 \\ \text{increased} & \text{if } \Delta P > 0 \end{cases} \tag{3.5}$$

We collected 15-minute windows of accelerometer data, starting 30 minutes before the pain assessment. We excluded the 15-minute window immediately before pain assessment because this contained artifacts of the patient's interaction with the caregiver. Throughout our analysis, 7 am to 7 pm was considered daytime, while 7 pm to 7 am was considered nighttime.

We collected 15-minute windows of accelerometer data, starting 30 minutes before the pain assessment. We excluded the 15-minute window immediately before the pain assessment because this contained artifacts of the patient's interaction with the caregiver. Throughout our analysis, 7 am to 7 pm was considered daytime, while 7 pm to 7 am was considered nighttime.

3.4.2 Acuity Assessment

To phenotype the patient acuity state as stable or unstable, we applied the method devised by Ren et al. [108], determining transitions in acuity status within the ICU. To capture the relevant data (accelerometer and clinical data) leading up to each assessment, we established a consecutive and non-overlapping 4-hour segmentation window that concluded immediately before the acuity evaluation to reflect patients' status. For every 4 hours leading up to the assessment, patients—excluding those who had passed away or were already discharged alive—were identified as unstable or stable. A patient was labeled as unstable if they required any of the following life-supportive therapies: vasopressors (epinephrine, vasopressin, phenylephrine, norepinephrine, droxidopa, or ephedrine), mechanical ventilation, continuous renal replacement therapy, or a massive blood transfusion (defined as at least ten units in the previous 24 hours), as previously described. If none of these conditions were met, the patient was considered stable.

Chapter 4

Pain Classification

Critically ill patients in ICUs often suffer from pain, with inconsistent and scarce pain assessment, exacerbated by communication barriers due to factors like medical interventions, cognitive deficits, and linguistic differences. While there are tools such as the visual analog scale (VAS) [13], numerator rating scale (NRS) [57], or Defense and Veterans Pain Rating Scale (DVPRS) [100] for self-reporting, non-verbal patients rely on visual assessments such as the Behavioral Pain Scale (BPS) [93], Critical Care Pain Observation (CPOT) [21], and Non-Verbal Pain Scales (NVPS) [48], which can be inconsistent due to human errors. Not addressing pain adequately can lead to various adverse physiological and psychological outcomes. Automated pain detection using AI has been explored as a solution, with current models mainly utilizing patients' facial expressions in videos [24, 88]. However, privacy concerns hinder their clinical adoption. Some studies suggest using vital signs, but these are not reliable pain indicators. Alternatively, accelerometers, tracking patient mobility, show promise in detecting pain.

4.1 Data Analysis

A dataset analysis was conducted to comprehensively understand the problem and help to build an effective solution. The subsequent discussion examines various aspects of the data, including the distribution of pain levels and accelerometer values and the correlations between the different modalities considered in this study: accelerometer data, demographic information, and pain records.

4.1.1 Distribution of pain levels

Figure 4.1 illustrates an analysis of pain level distribution, revealing a bimodal pattern. A notable concentration of samples at pain level zero indicates a substantial portion of the dataset representing no pain. However, a distinct secondary peak emerges at moderate pain levels, specifically levels five and six. This bimodality strongly suggests the presence of two distinct clusters of samples within the dataset. One subgroup is characterized by minimal to no pain, while the other exhibits measurable moderate pain. We attribute this pattern to the administration of medications when patients experience pain. It's expected

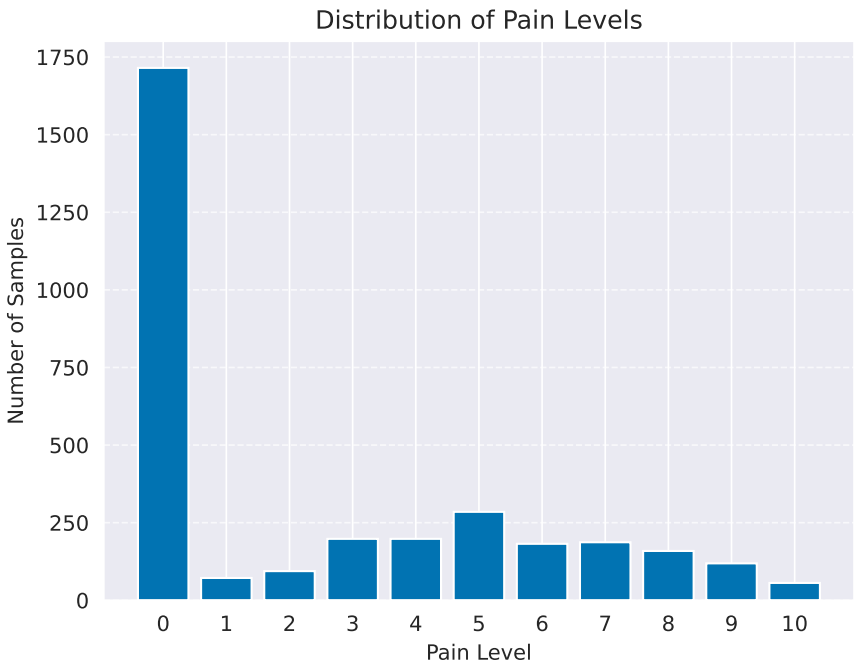


Figure 4.1: Distribution of Self-Reported Pain Levels. This figure illustrates the distribution of pain levels within the dataset. An imbalance in the distribution is evident, characterized by a concentration of samples at pain level 0 (no pain). This may reflect pain management via medication.

that even though a patient may endure significant pain, they may not experience it for the majority of the time. Each instance of pain observation triggers medication administration by nurses, resulting in pain relief and the prevalence of zero pain instances in the dataset.

Figure 4.2 presents the distribution of samples according to pain levels across various data stratifications: shift, gender, age, and health status. These four stratifications reveal a consistent data imbalance, predominantly skewed towards the 'no pain' level. The upper left figure illustrates the distribution by shift, indicating a slight predominance of samples during the day shift. Nonetheless, the dataset maintains a reasonable balance between day and night samples overall.

The upper right figure illustrates the distribution by gender, revealing a lower representation of samples from individuals assigned female at birth compared to those assigned male at birth. The lower left figure delineates the distribution across two age groups: adults (18-64 years old) and the elderly (65+ years old). Notably, the elderly group exhibits fewer samples across most pain levels than the adult group, except at pain levels 2 and 10, where the representations are similar.

Finally, the lower right figure displays the distribution by health status severity, categorized according to the Sequential Organ Failure Assessment (SOFA) score. This analysis shows a predominance of samples from individuals with medium-severity health statuses, whereas high-severity moments are underrepresented. It is particularly impor-

Figure 4.2: Distribution of Self-Reported Pain Levels.

tant to highlight that, for certain pain levels such as 1 and 10, the dataset contains only one sample representing high severity status, indicating a critical gap in the data.

4.1.2 Distribution of acceleration values

Figure 4.3 shows the histograms of the probability density of acceleration magnitudes across different pain levels. The histograms reveal a largely consistent pattern. Most pain levels' probability densities are concentrated at lower acceleration magnitudes, with smaller probabilities extending into higher accelerations. This suggests that, generally, ICU patients exhibit limited movement regardless of the pain level experienced, possibly due to the combined effects of analgesia, the controlled ICU environment, and the inherent physical limitations of patients in critical care.

However, exceptions to this trend are observed at pain levels 1 and 10, which appear to deviate from the common pattern. These levels display histograms that are not as well aligned with the others, possibly due to undersampling in the dataset. Pain level 1 shows an unusually high peak at a specific low acceleration range, which might be exaggerated due to fewer data points, potentially skewing the representation of typical patient movement at this level. Similarly, pain level 10 shows a distinct pattern, which might not accurately reflect the true movement behavior at high pain levels due to the limited sample size.

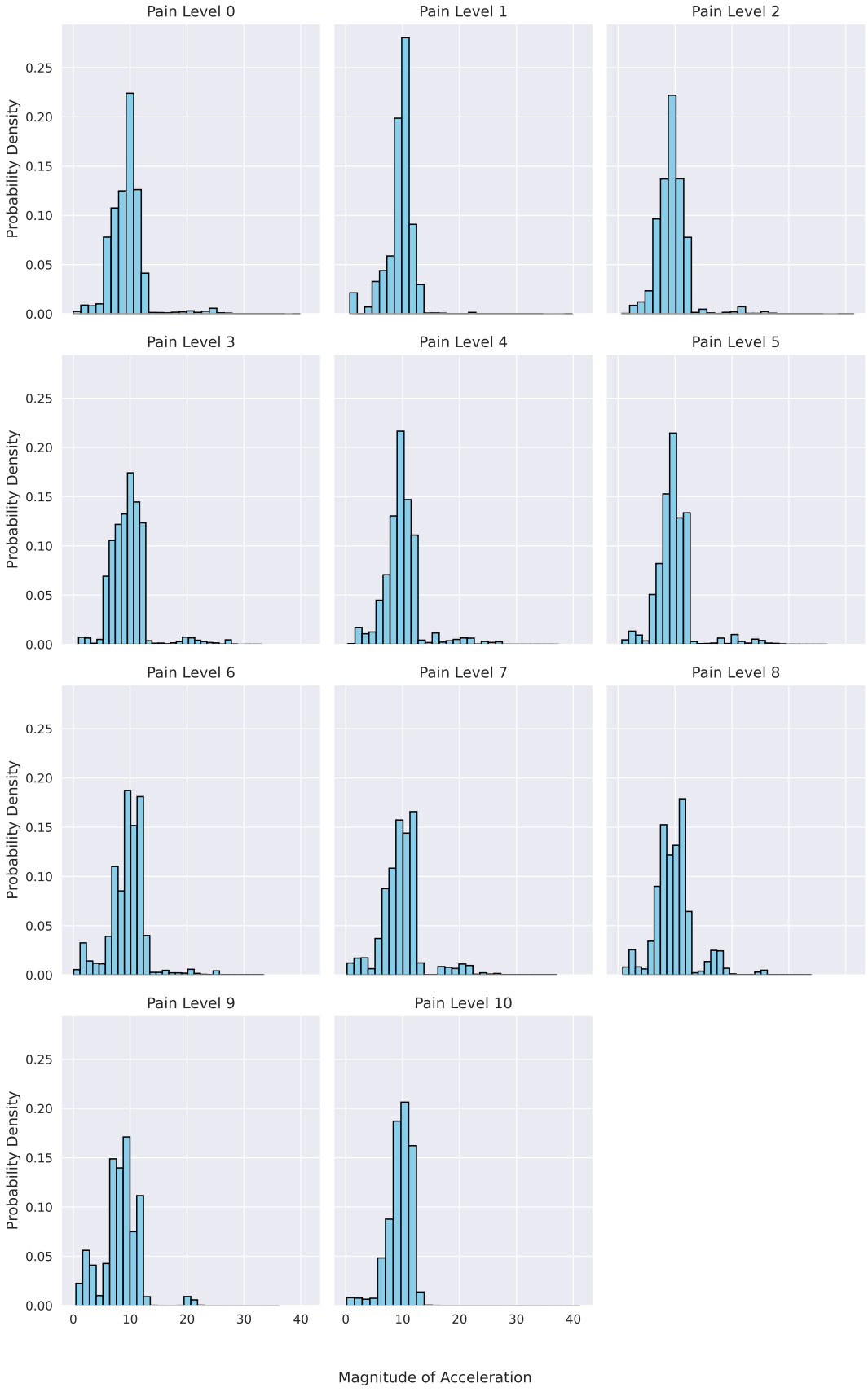


Figure 4.3: Histograms of the probability density of acceleration magnitudes across different pain levels

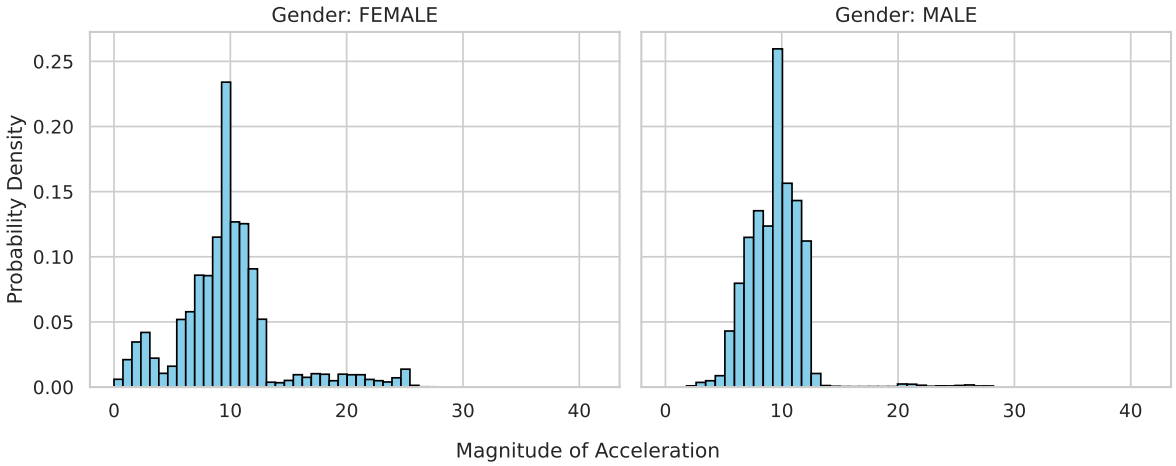


Figure 4.4: Histograms of the probability density of acceleration magnitudes for ICU patients, categorized by gender.

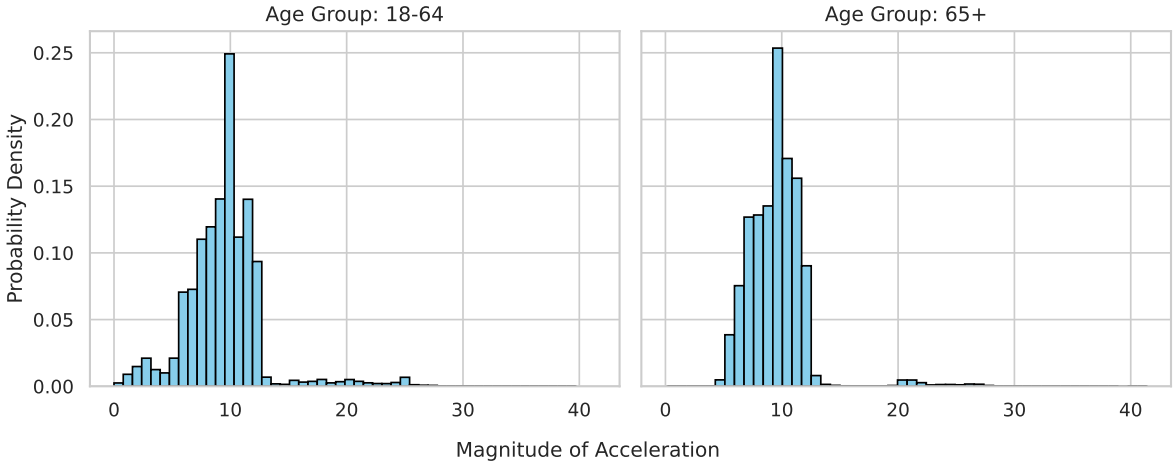


Figure 4.5: Histograms of the probability density of acceleration magnitudes for ICU patients, categorized by age group.

Figure 4.4 shows the histograms representing the probability density of acceleration magnitudes for ICU patients, categorized by gender, displaying a predominant concentration at lower acceleration magnitudes with similar peak regions for females and males. This consistent pattern across genders suggests a general trend of minimal movement within the ICU environment, likely reflecting effective pain management protocols and the controlled nature of ICU settings to limit patient movement.

While both genders show most of their probability density at lower accelerations, there are still minor differences in the data spread. The female histogram shows slightly more frequent occurrences at the lowest acceleration magnitudes compared to the male histogram, which could suggest subtle differences in movement or pain response strategies between genders.

Figure 4.5 shows the histograms comparing the probability density of acceleration magnitudes across two different age groups, 18-64 and 65+. The histograms exhibit

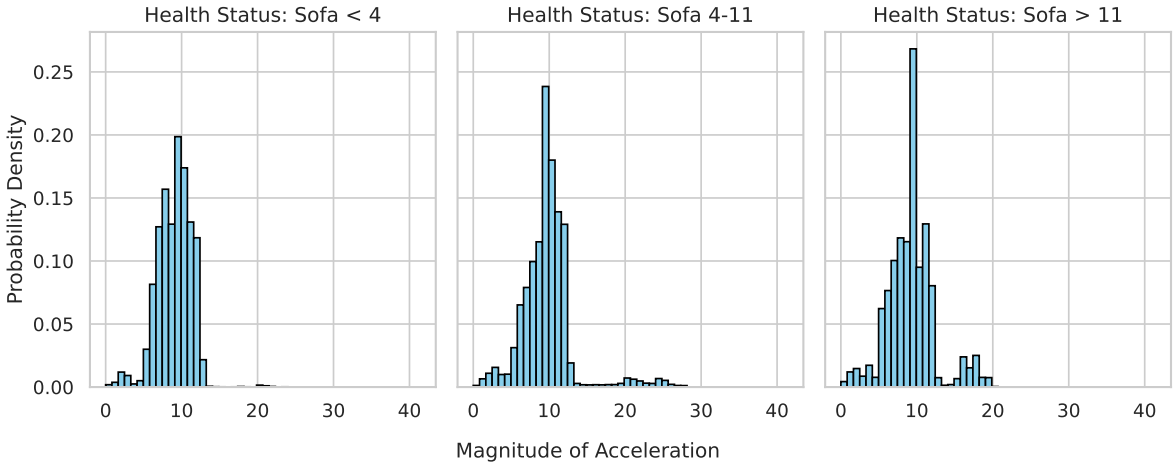


Figure 4.6: Histograms of the probability density of acceleration magnitudes for ICU patients, categorized by health status.

a strong peak at lower acceleration magnitudes for both age groups, indicating a predominant trend of minimal movement. This similarity suggests that irrespective of age, patients in the ICU typically exhibit limited mobility, which is consistent with the controlled nature of ICU environments. However, upon closer inspection, the histogram for the 18-64 age group displays a slightly wider spread towards higher acceleration magnitudes than the 65+ group. This could suggest that younger patients might have a slightly higher level of physical activity or response capability when compared to older patients. Such differences may reflect age-related variations in physical capacity or responsiveness to pain and medical interventions.

Figure 4.6 illustrates the probability density of acceleration magnitudes for ICU patients, grouped according to their health status as measured by the SOFA (Sequential Organ Failure Assessment) score in three categories: SOFA < 4 (minimal organ dysfunction), SOFA 4-11 (mid-range organ dysfunction), and SOFA > 11 (severe organ dysfunction). These histograms provide insights into how different levels of organ dysfunction impact patient movement.

The histogram for SOFA < 4 indicates a prominent peak at lower acceleration magnitudes, with minimal dispersion towards higher values. This trend suggests that individuals with minimal organ dysfunction predominantly exhibit reduced mobility. Within the SOFA 4-11 range, the distribution also centers on lower acceleration magnitudes but displays a slightly broader spread compared to the SOFA < 4 cohort. This broader distribution implies that patients experiencing mid-range organ dysfunction may demonstrate variable mobility, potentially reflecting a blend of stable periods and instances of acute distress or medical interventions prompting slight increases in movement.

Analogous to the other categories, for SOFA > 11, the prevailing acceleration remains at lower magnitudes, yet this group also exhibits some dispersion towards higher accelerations. The presence of higher magnitudes, albeit less frequent, may suggest that

individuals with severe organ dysfunction (SOFA > 11) intermittently undergo significant movements, potentially attributable to medical emergencies, interventions, or periods of considerable discomfort.

The histograms presented in Figure 4.7 illustrate the probability density of acceleration magnitudes for ICU patients across various medical conditions, reflecting differences in patient mobility associated with each specific health issue. These conditions include Cancer, Cerebrovascular Disease, Dementia, Paraplegia/Hemiplegia, Congestive Heart Failure (CHF), Chronic Obstructive Pulmonary Disease (COPD), Diabetes, Myocardial Infarction, Liver Disease, Peptic Ulcer Disease, Renal Disease, and Rheumatologic Disease.

The histograms generally show concentrated peaks at lower acceleration magnitudes, indicating limited movement across most conditions, which aligns with the restrictive nature of ICU care where patient movement is minimized to promote recovery. Conditions such as Congestive Heart Failure (CHF), Diabetes, Liver Disease, Peptic Ulcer Disease, and Renal Disease are characterized by broader distributions in their histograms. This broader spread suggests more variability in the movement levels of these patients, potentially indicative of fluctuating symptoms that might occasionally require or allow for greater mobility within the constraints of ICU care.

On the other hand, Rheumatologic Disease, Chronic Obstructive Pulmonary Disease (COPD), Paraplegia/Hemiplegia, and Cerebrovascular Disease histograms show a narrower distribution with a defined peak. This pattern implies a more uniform level of restricted movement among these patients, likely due to more consistent physical limitations or symptoms directly impacting mobility.

Notably, patients with Cancer and Dementia exhibit histograms with narrow distributions but multiple peaks. This unique pattern could indicate different mobility profiles within these patient groups, perhaps reflecting varying stages of disease progression or differing responses to treatment that affect physical activity levels.

4.1.3 Data separability analysis

Understanding the patterns and structure within the dataset is crucial for developing effective AI algorithms for ICU patient monitoring. A data separability analysis was conducted on the accelerometer data using Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) to achieve this. These techniques were selected for their ability to reduce dimensionality and visualize complex data structures, thereby providing insights into the inherent separability of the data. For a comprehensive visualization, we plot the raw accelerometer and the various feature extraction used in this work in Figure 4.9a. Additionally, we plot each set of EHR data in Figure 4.10.

PCA and t-SNE demonstrate that pain levels do not form clearly separable clus-

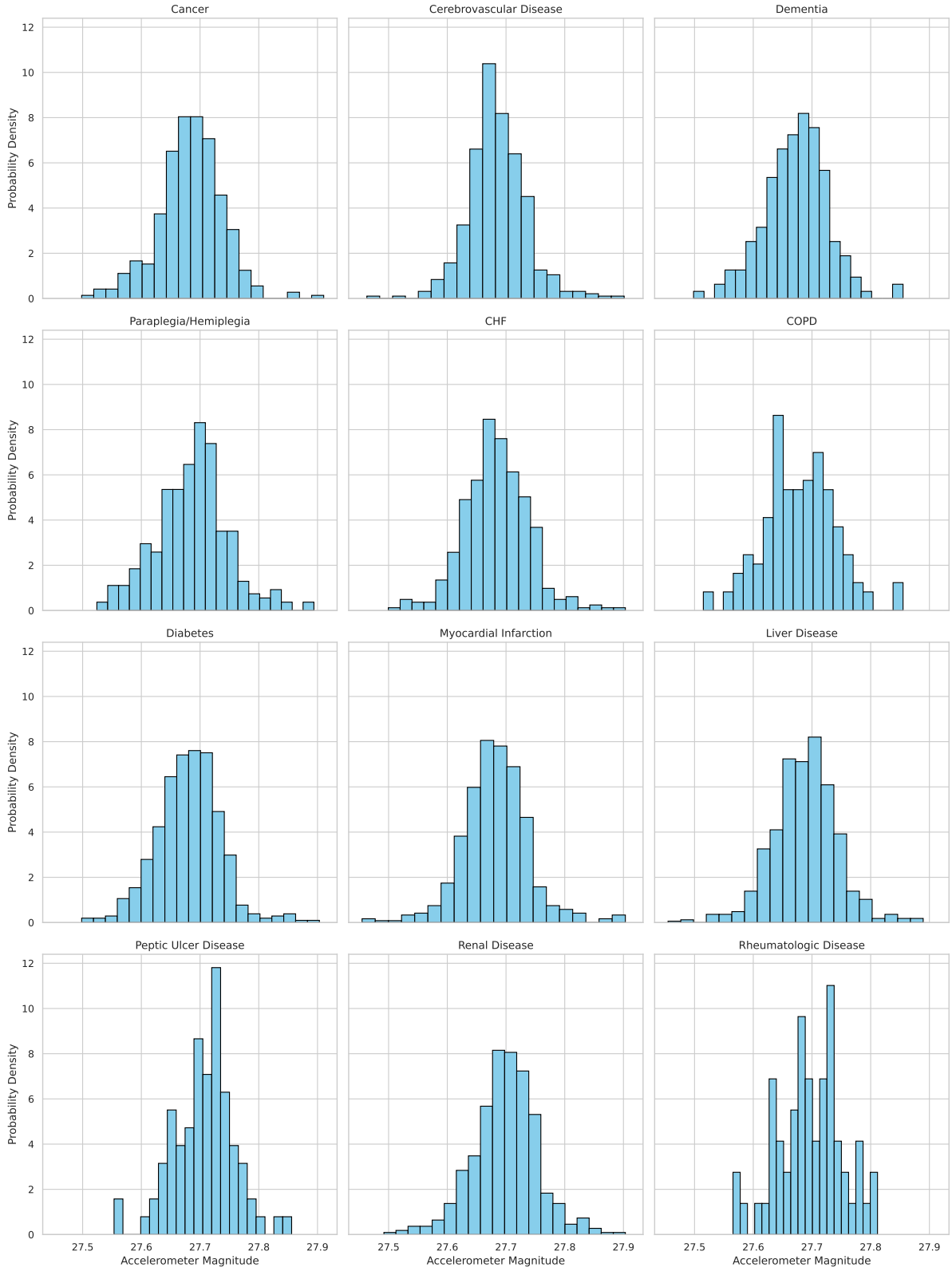
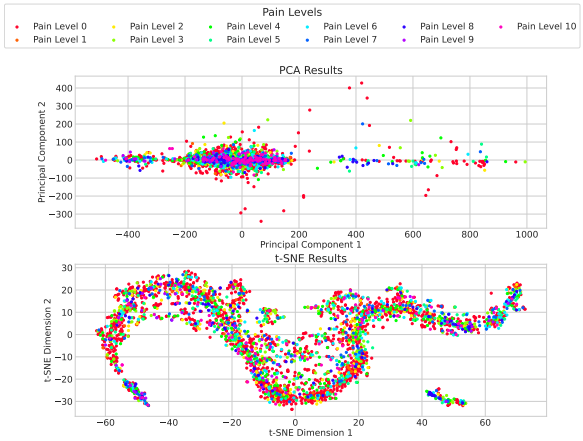
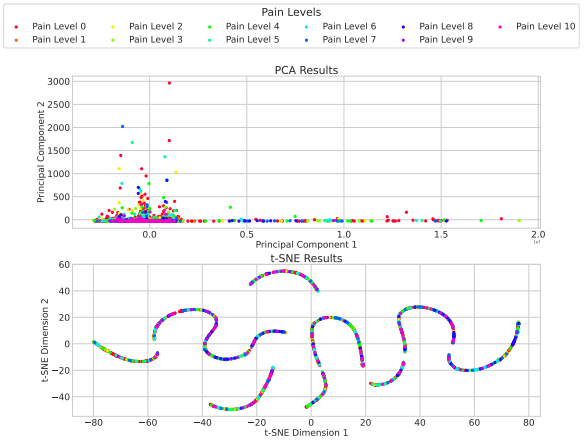


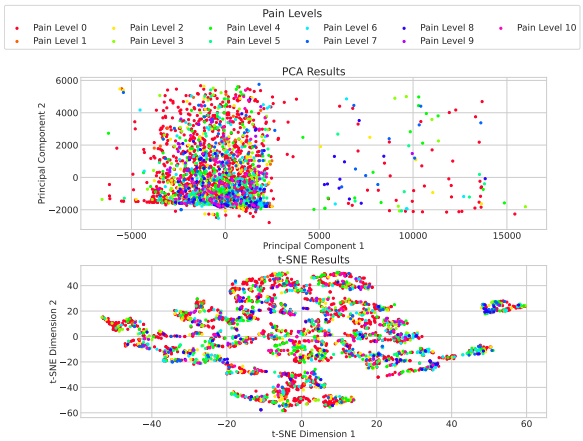
Figure 4.7: Histograms of the probability density of acceleration magnitudes for ICU patients, categorized by disease.



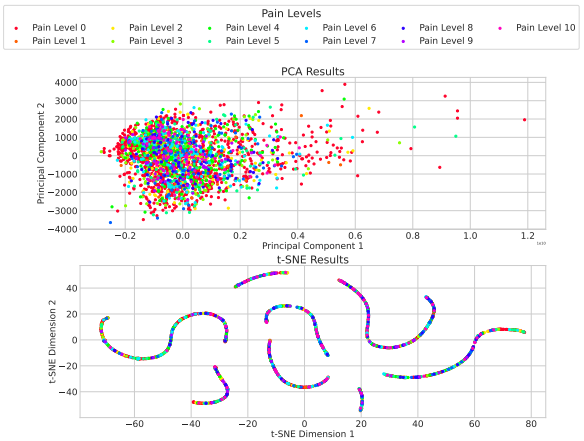
(a) Raw accelerometer data



(b) Statistical features from tsfel library

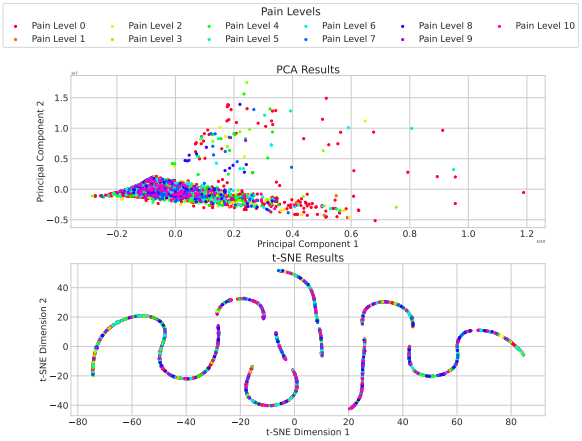


(c) Temporal features from tsfel library

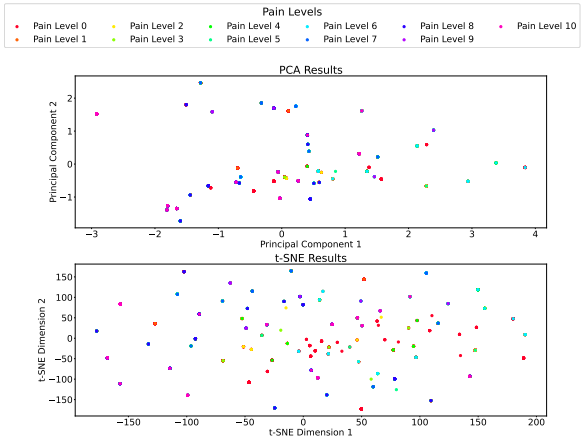


(d) Spectral features from tsfel library

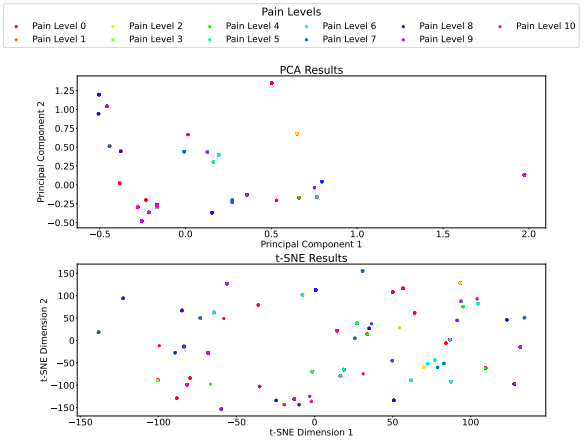
Figure 4.8: Comparison of different feature separability



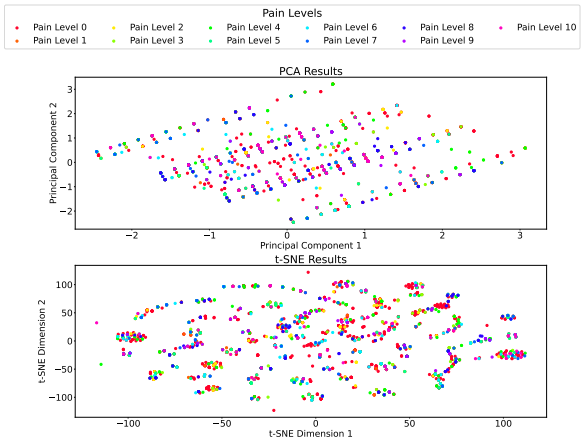
(a) All features from tsfel library



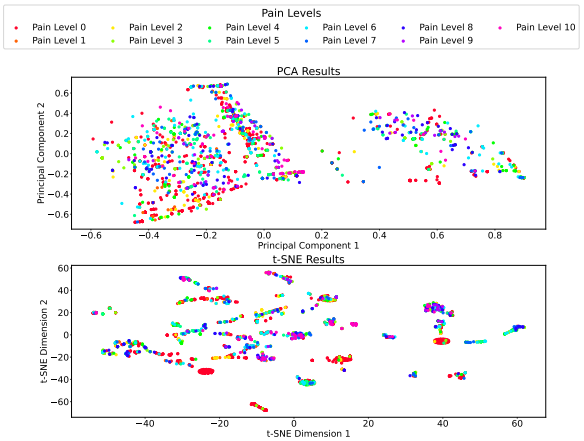
(a) Demographic Data



(b) Comorbidities Data



(c) Clinical Data



(d) Analgesics Data

Figure 4.10: Comparison of different features

ters in any of the input features considered. This is expected since pain is subjective and influenced by many factors beyond just the physical injury itself. Unlike measuring temperature or blood pressure, there's no single objective gauge for pain.

To analyze inter-individual pain variability, which is known to be significant, we examined samples from four different patients. Using Euclidean distance and demographic data, we identified the two most similar patients, referred to as S1 and S2, and the two most dissimilar patients, referred to as D1 and D2. Table 4.1 provides the demographic information for these four patients. We employed t-SNE to reduce dimensionality from a combination of medication, clinical, and disease data. Figure 4.11 shows the samples from our dataset and highlights the samples from these patients for a commonly experienced pain level of 6.

Table 4.1: Demographic Information of Four Selected Patients. This table presents the demographic details of four patients analyzed for inter-individual pain variability. Patients S1 and S2 were identified as the most similar based on Euclidean distance and demographic data, while D1 and D2 were the most dissimilar.

	sex	**race**	**height cm**	**age**	**weight kgs**	**length stay**	**aids**	**cancer**	**cerebro-vascular disease**	**dementia**	**paraplegia hemiplegia**	**diabetes**	**liver disease**
S1	MALE	WHITE	177.8	50	86.58	31	0	0	0	0	0	0	1
S2	MALE	WHITE	180.34	53	81.65	3	0	0	0	0	0	0	1
D1	FEMALE	WHITE	157.48	77	54.12	4321	0	0	0	0	0	0	0
D2	MALE	WHITE	182.9	61	188.7	4	0	0	1	0	0	1	0

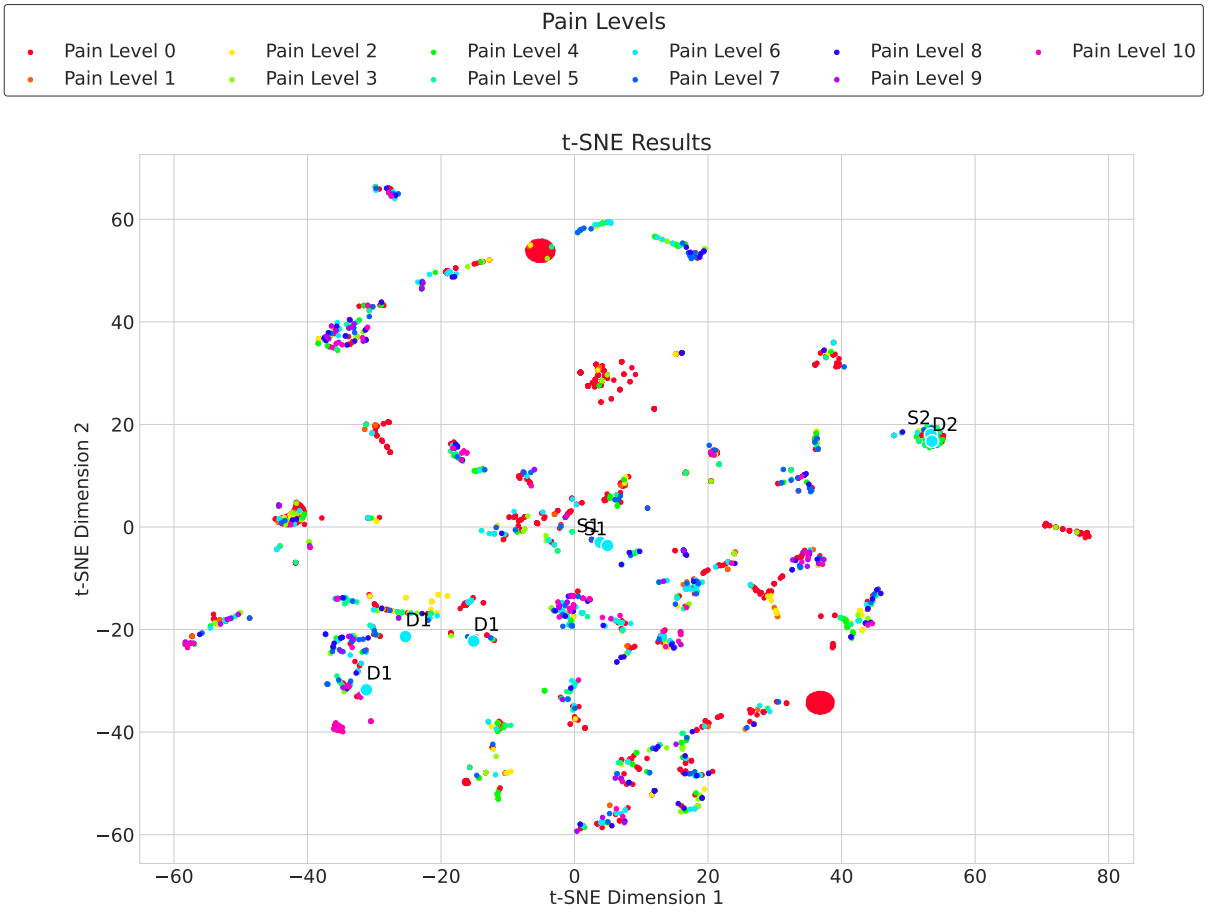


Figure 4.11: t-SNE visualization of all pain levels for all patients, reduced from a combination of medication, clinical, and disease data. Points corresponding to the four selected patients (S1, S2, D1, and D2) at pain level 6 are specifically indicated, highlighting the most similar and most dissimilar patients based on Euclidean distance from demographic data.

4.2 Methodology

4.2.1 Feature extraction and selection

The feature extraction process for accelerometer data involved important steps to ensure the relevance and quality of the features used for subsequent analysis.

Initially, a low-pass filter was applied to the raw accelerometer data. This filtering step was important to remove high-frequency noise, which could otherwise obscure the meaningful patterns within the data. The low pass filter helped preserve the signal's integrity corresponding to human movements by attenuating the frequencies above a certain threshold. Figure 4.12 shows an accelerometer excerpt from our dataset before and after applying the low-pass filter. We empirically choose an order of 6 and a cutoff of 3.

Following the filtering step, the magnitude vector of the accelerometer signal was calculated. The magnitude vector provided a single comprehensive measure of movement

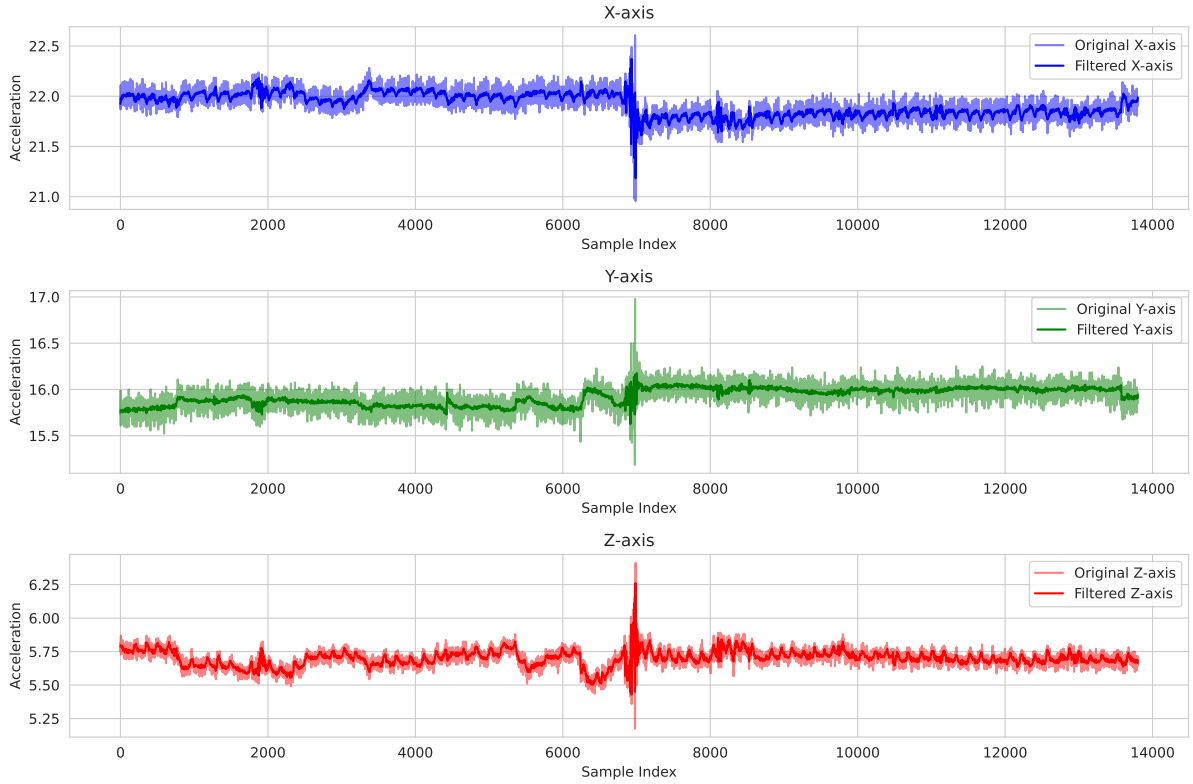


Figure 4.12: Accelerometer signal before and after low pass filter

intensity, combining the three-dimensional accelerometer data into one scalar value for each time point. This transformation facilitated the extraction of features that are more representative of the overall movement.

Next, the TSFEL (Time Series Feature Extraction Library) was employed to extract a wide array of features from the filtered accelerometer data. TSFEL is a powerful tool that automates the extraction of relevant features from time series data, encompassing statistical, temporal, and spectral domains. This comprehensive feature extraction step yielded a rich feature set. The list of features extracted is shown in Table 4.2.

To ensure the robustness of the feature set, features with NaN (Not a Number) and infinite values were identified and removed. Retaining such values would adversely affect the reliability of the machine-learning models.

Subsequently, highly correlated features were removed to mitigate redundancy and multicollinearity within the feature set. Highly correlated features do not contribute to additional information and can lead to overfitting in machine learning models. Calculating the correlation matrix and applying a threshold systematically eliminated redundant features.

Additionally, features exhibiting low variance were excluded from the feature set. Low variance features carry minimal discriminatory information and do not contribute significantly to the model's predictive power. Their removal streamlined the feature set, enhancing the computational efficiency of the subsequent analysis. We used the scikit-

Table 4.2: List of features divided by domains

Statistical Domain	**Temporal Domain**	**Spectral Domain**
Absolute energy	Area under the curve	FFT mean coefficient
Average power	Autocorrelation	Fundamental frequency
ECDF	Centroid	Human range energy
ECDF Percentile	Lempel-Ziv-Complexity	LPCC
ECDF Percentile Count	Mean absolute diff	MFCC
Entropy	Mean diff	Max power spectrum
Histogram	Median absolute diff	Maximum frequency
Interquartile range	Median diff	Median frequency
Kurtosis	Negative turning points	Power bandwidth
Max	Peak to peak distance	Spectral centroid
Mean	Positive turning points	Spectral decrease
Mean absolute deviation	Signal distance	Spectral distance
Median	Slope	Spectral entropy
Median absolute deviation	Sum absolute diff	Spectral kurtosis
Min	Zero crossing rate	Spectral positive turning points
Root mean square	Neighbourhood peaks	Spectral roll-off
Skewness		Spectral roll-on
Standard deviation		Spectral skewness
Variance		Spectral slope
		Spectral spread
		Spectral variation
		Wavelet absolute mean
		Wavelet energy
		Wavelet standard deviation
		Wavelet entropy
		Wavelet variance

learn library default threshold. The default is to keep all features with non-zero variance, i.e., remove the features with the same value in all samples.

Finally, the Infinite Feature Selection (Inf-FS) technique was applied to the refined feature set. Inf-FS is a graph-based feature filtering method designed to identify the most relevant features for a given dataset. The core idea behind Inf-FS is to model the relationships between features using a weighted graph, where nodes represent features and edges represent the similarity or relevance between them. This graph-based approach enables the method to consider the global structure of the feature space rather than evaluating features independently. By leveraging the concept of random walks on this graph, Inf-FS iteratively computes a score for each feature, reflecting its importance and connectivity within the feature network. The scores are then used to rank the features, allowing for selecting the most informative subset. This method is particularly effective in handling high-dimensional data, as it efficiently filters out redundant or irrelevant features, thereby improving the performance of subsequent machine learning models.

4.2.2 Machine Learning Methods

Since this is the first study to utilize accelerometers and clinical data to classify pain in the ICU, it is important to evaluate a wide range of classifiers to establish a robust reference point. Table 4.3 lists the comprehensive array of classifiers considered in this study, along with brief explanations. This evaluation covers various linear models, Naive Bayes models, ensemble methods, gradient boosting methods, meta-estimators, and non-parametric methods. Additionally, we evaluated a dummy classifier that makes predictions that ignore the input features. This classifier will serve as a simple baseline to compare against the other more complex classifiers. Below, we provide a detailed description of these classification algorithms and a comparative analysis highlighting their strengths and weaknesses.

4.2.2.1 Linear Models

Linear Discriminant Analysis (LDA) is a statistical method for classifying data into different classes. It works by projecting the data onto a lower-dimensional space to maximize class separation. LDA is particularly effective when the assumption that the classes have similar covariance matrices holds true. This method is computationally efficient and easy to interpret, making it a popular choice for many classification problems. However, one of its main limitations is that it assumes linear decision boundaries between classes, which can be a significant drawback when dealing with datasets with nonlinear boundaries. This limitation can lead to suboptimal performance on complex datasets, where more flexible models might be required to capture the underlying patterns in the data.

Table 4.3: List of Models with Descriptions

Model	Description
Linear Discriminant Analysis (LDA)	Linear classification algorithm that projects data to maximize class separation.
Bernoulli NB	Naive Bayes classifier for binary/boolean features.
Logistic Regression	Regression model for binary classification, estimating probabilities using a logistic function.
Calibrated Classifier CV	Meta-estimator for probability calibration using cross-validation.
Dummy Classifier	Baseline classifier that makes predictions using simple rules.
Extra Trees Classifier	Ensemble of randomized decision trees for classification.
Random Forest Classifier	Ensemble of decision trees with bagging and feature randomness.
CatBoost Classifier	Gradient boosting on decision trees with categorical feature support.
LGBM Classifier	LightGBM classifier with efficient training for large datasets.
XGBoost Classifier	Extreme Gradient Boosting with high performance and scalability.
Bagging Classifier	Ensemble method using bootstrap aggregating for various classifiers.
KNN Classifier	Classification based on the k-nearest neighbors algorithm.
AdaBoost Classifier	Adaptive boosting ensemble method for improving weak classifiers.
Decision Tree Classifier	Non-parametric classifier based on decision tree structures.
Extra Tree Classifier	Extremely Randomized Trees classifier for more randomness.
Quadratic Discriminant Analysis (QDA)	Classification algorithm assuming quadratic decision boundaries.
Gaussian NB	Naive Bayes classifier assuming Gaussian distribution of features.
Multi-Layer Perceptron (MLP)	A neural network with multiple layers that capture complex, non-linear relationships in data.

Logistic Regression is a widely used model for binary classification tasks. It estimates the probability that a given instance belongs to a particular class by applying the logistic function to a linear combination of the input features. The output is a probability value between 0 and 1, which can be thresholded to make a binary decision. Logistic Regression is highly interpretable, as it allows us to understand the influence of each feature on the outcome. It is also computationally efficient and works well when the data is linearly separable. However, its performance diminishes when the relationship between the features and the target variable is not linear. In such cases, Logistic Regression might fail to capture the complex patterns in the data, making it less effective than non-linear models. Despite this, its simplicity and ease of implementation make it a baseline method that is often used before exploring more complex models.

4.2.2.2 Naive Bayes Models

Bernoulli Naive Bayes (Bernoulli NB) is a variant of the Naive Bayes classifier specifically designed for binary or boolean features. This method operates on the principle that each feature in the dataset is binary (either present or absent) and assumes that the presence or absence of a feature is independent of the presence or absence of any other feature, given the class label. This independence assumption simplifies the computation of the probabilities needed for classification, making Bernoulli NB computationally efficient and easy to implement. It is particularly effective for text classification tasks where the data can be represented as binary word vectors (e.g., word present or not in a document). However, the independence assumption can be a significant limitation in practice, as it often does not hold true for real-world data where features can be interdependent. This can lead to suboptimal performance if the relationships between features are crucial for accurate classification.

Gaussian Naive Bayes (GaussianNB) is another variant of the Naive Bayes classifier that assumes a Gaussian (normal) distribution for the continuous features. This method calculates the likelihood of the data given the class label by assuming that the features follow a Gaussian distribution. GaussianNB is particularly efficient and works well with continuous data, making it suitable for various applications, including medical diagnosis and financial predictions. Its main advantages are simplicity and speed, as it only requires the features' mean and variance to compute the probabilities. However, assuming that the features follow a Gaussian distribution can be a significant drawback when the actual data distribution deviates significantly from normality. The model may produce inaccurate probability estimates in such cases, leading to poor classification performance. Despite this limitation, GaussianNB remains a popular choice due to its ease of implementation and computational efficiency, especially for large datasets where more complex models might be impractical.

4.2.2.3 Gradient Boosting Methods

CatBoost Classifier is a sophisticated gradient-boosting algorithm that handles categorical features efficiently. It uses an innovative technique called ordered boosting to reduce overfitting, and it supports categorical feature encoding internally, making it highly suitable for datasets with categorical variables. CatBoost achieves high accuracy and often outperforms other gradient-boosting methods regarding prediction quality. However, it requires careful parameter tuning to achieve optimal performance and can be resource-intensive in terms of both computational power and memory usage. This makes it a powerful yet demanding tool, particularly effective for complex problems where categorical data plays a significant role.

LightGBM Classifier is an efficient implementation of gradient boosting that excels at handling large datasets. It uses a histogram-based approach to bin continuous features into discrete bins, which speeds up training and reduces memory usage. LightGBM also supports the leaf-wise growth of trees, allowing it to handle large datasets with many features more effectively than other gradient-boosting methods. While LightGBM is highly efficient and scalable, it may struggle with smaller datasets where the binning process can lead to loss of information. Additionally, it requires careful handling of categorical features to ensure optimal performance, making it less straightforward compared to other methods.

XGBoost Classifier is a highly popular gradient-boosting algorithm known for its performance and scalability. It incorporates regularization techniques to prevent overfitting, which enhances its generalization ability. XGBoost supports parallel processing, enabling faster training on large datasets, and includes features like tree pruning, which helps avoid overfitting. Despite its strengths, XGBoost can be complex to tune, requiring extensive parameter adjustments to achieve the best results. Moreover, its computational intensity and longer training times for very large datasets can be a drawback, making it less suitable for applications where quick turnaround is essential. However, due to its robustness and versatility, XGBoost remains a go-to method for high-stakes and large-scale applications.

4.2.2.4 Meta-Estimators

Calibrated Classifier CV is a meta-estimator designed to improve the probability calibration of base classifiers, ensuring that a model's predicted probabilities accurately reflect the true likelihood of an event. This is particularly crucial in applications involving risk assessment or decision-making under uncertainty, where well-calibrated probabilities are essential. Calibrated Classifier CV uses cross-validation to enhance the reliability of probability estimates, making it valuable in fields such as finance, healthcare, and weather forecasting. While it can significantly improve the accuracy of probability predictions, it

also adds computational complexity and processing time compared to using the base classifier alone. Additionally, it inherits the strengths and weaknesses of the base classifier, so issues like overfitting may still persist despite improved calibration. Overall, Calibrated Classifier CV is a powerful tool for enhancing probability estimates, especially in high-stakes applications, but its computational demands and the characteristics of the base classifier must be carefully considered.

4.2.2.5 Non-Parametric Methods

K-Nearest Neighbors Classifier (KNN) is a simple yet effective non-parametric method used for classification. The core idea behind KNN is to classify a data point based on the majority class among its k-nearest neighbors in the feature space. This approach makes KNN highly intuitive and easy to implement. It is particularly effective for small datasets and situations where the decision boundary is not necessarily linear. However, KNN can be computationally expensive for large datasets because it requires calculating the distance between the query point and all other data points. Additionally, KNN is sensitive to the choice of k and the distance metric, which can significantly impact its performance. It also lacks interpretability, as the model does not provide explicit rules for classification.

Decision Tree Classifier is a non-parametric method that splits the data into subsets based on feature values, forming a tree-like structure. Each node in the tree represents a decision rule based on a feature, and each branch represents the outcome of the rule, leading to a leaf node representing a class label. Decision Trees are easy to interpret and visualize, making them valuable for understanding the decision-making process. However, they are prone to overfitting, especially when the tree is deep and captures noise in the data. To mitigate overfitting, techniques like pruning, which removes less important nodes, can be applied. Despite this, Decision Trees may still struggle with capturing complex patterns compared to ensemble methods like Random Forests.

Extra Tree Classifier is similar to the standard Decision Tree Classifier but introduces additional randomness into the tree-building process. Unlike standard decision trees, which choose the best split based on a criterion like Gini impurity or information gain, Extra Trees randomly select the split points. This added randomness helps to improve generalization and reduce overfitting, making Extra Trees more robust on various datasets. However, this increased randomness can also make the model less interpretable and more challenging to tune. While it often results in better generalization than a single decision tree, it requires more computational resources and may not always outperform other ensemble methods like Random Forests.

Quadratic Discriminant Analysis (QDA) is a classification algorithm that assumes quadratic decision boundaries between classes. QDA generalizes LDA by allowing each class to have its own covariance matrix, which provides more flexibility in modeling the

data. This flexibility enables QDA to handle datasets where the relationships between features and class labels are more complex and non-linear. However, QDA requires a large amount of data to accurately estimate the parameters of each class's covariance matrix, which can be a limitation for smaller datasets. Additionally, the model's performance can be sensitive to the accuracy of the estimated parameters, making it crucial to have sufficient and well-distributed data.

Multilayer Perceptron (MLP) is an artificial neural network consisting of multiple layers of nodes, or neurons, connected feedforward. MLPs can capture complex, non-linear relationships in the data, making them highly versatile and powerful for a wide range of classification tasks. Each neuron applies a weighted sum of its inputs and passes the result through a non-linear activation function, allowing the network to learn intricate patterns. The strength of MLP lies in its ability to model highly non-linear decision boundaries and its applicability to various types of data, including time series and image data. However, MLPs require extensive training data to achieve good performance and can be computationally intensive due to the need for multiple iterations during training (backpropagation). Additionally, they require careful tuning of hyperparameters, such as the number of layers, the number of neurons per layer, and the learning rate to prevent overfitting and ensure convergence.

4.3 Evaluation Protocol

To evaluate the performance of our machine learning models, we employed a hold-out test approach, splitting the data into 80% for training and 20% for testing. This method was chosen due to the large number of experiments, making cross-validation impractical. The hold-out approach efficiently assesses model performance by training and testing on separate data sets, ensuring effective evaluation without the computational overhead of cross-validation.

Additionally, we conducted a comparative analysis of three different undersampling methods against a baseline with no undersampling. In each experiment, we assessed the impact of undersampling on classifying DVPRS 11 pain levels using only accelerometer data across 17 different classifiers (refer to 4.3). The analysis focused on performance metrics such as ROC AUC, F1 Score, Recall, and Precision, which are crucial for evaluating the classifiers' effectiveness in distinguishing between multiple classes.

The results shown in Table 4.4, demonstrate that applying different undersampling methods minimally impacts improving the performance metrics for the given models in the context of 11-class classification. The ROC AUC values indicate that all models are essentially performing at the level of random guessing, with none surpassing the threshold of 0.5. Among the undersampling approaches, Tomek Links shows slight improvements

Table 4.4: Performance metrics of the best model for pain classification using different undersampling methods.

Undersampling Method	Model	ROC AUC	F1 Score	Recall	Precision
None	QDA	0.49	0.39	0.55	0.31
Random	Dummy	0.50	0.39	0.55	0.30
Near Miss	Dummy	0.50	0.39	0.55	0.30
Tomek Links	**CatBoost**	**0.50**	**0.40**	**0.52**	**0.39**

Abbreviation: QDA - Quadratic Discriminant Analysis, ROC AUC - Area Under the Receiver Operating Characteristic Curve

in F1 Score and Precision, which leads us to employ this method in the remaining experiments of this Chapter.

4.4 Experimental Results

In this section, we present the results of our experiments in pain classification, focusing on various labeling strategies for the data. We evaluated the performance of our models across multiclass classification tasks, including pain score and pain variation, as well as binary classification tasks, such as grouped pain levels (mild vs. moderate, moderate vs. severe, severe vs. mild) and pain vs. no pain. For all experiments, both day and night timeframes were considered. In the following a summary of the results for each experiment will be shown. The Harmonic Mean was the metric used to determine the best classifier. For detailed results with the performance of each classifier for each input, please refer to Appendix A.

4.4.1 Multiclass classification of pain scores

The experiments aimed to measure the performance of various inputs and their combinations in classifying self-reported pain levels from the DVPRS pain scale composed of 11 pain levels.

Table 4.5 demonstrates that analgesic medication yielded the highest performance metrics for both day and night classifications when used as the sole input. During the day, the Logistic classifier achieved the best results with an ROC AUC of 0.63, an F1 Score of 0.42, and a Harmonic Mean of 0.48. At night, XGBoost was the top performer for the analgesic medication input, with an ROC AUC of 0.66, an F1 Score of 0.44, and a Harmonic Mean of 0.48. These outcomes indicate that analgesic medication information is a strong indicator of pain levels in ICU patients, which aligns with expectations.

Table 4.5: DVPRS 11 classes classification

Input	Best Classifier	ROC AUC	F1 Score	Recall	Precision	Harmonic Mean
Day						
	Dummy	0.50	0.38	0.54	0.29	0.40
accel	Extra Trees	0.52	0.38	0.52	0.34	0.42
meds	**Logistic**	**0.63**	**0.42**	**0.55**	**0.39**	**0.48**
demo	Bagging	0.61	0.42	0.54	0.34	0.45
diseases	Dummy	0.50	0.38	0.54	0.29	0.41
clinical	Extra Trees	0.48	0.38	0.44	0.35	0.41
accel + meds	MLP	0.59	0.42	0.48	0.41	0.47
accel + demo	Bagging	0.57	0.42	0.54	0.39	0.47
accel + diseases	Bernoulli NB	0.46	0.38	0.49	0.32	0.40
accel + clinical	Calibrated	0.48	0.38	0.54	0.29	0.40
accel + meds + demo	MLP	0.61	0.43	0.49	0.40	0.47
accel + meds + demo + clin	**Logistic**	**0.57**	**0.45**	**0.50**	**0.41**	**0.48**
accel + meds + demo + clin + diseases	Logistic	0.57	0.42	0.50	0.38	0.46
Night						
	Dummy	0.50	0.35	0.51	0.26	0.38
accel	MLP	0.49	0.36	0.52	0.32	0.41
meds	**XGBoost**	**0.66**	**0.44**	**0.48**	**0.41**	**0.48**
demo	Extra Trees	0.58	0.37	0.52	0.29	0.41
diseases	CatBoost	0.51	0.37	0.45	0.31	0.40
clinical	Extra Trees	0.52	0.37	0.47	0.39	0.43
accel + meds	MLP	0.69	0.41	0.50	0.39	0.47
accel + demo	LDA	0.45	0.38	0.53	0.37	0.42
accel + diseases	Bernoulli NB	0.61	0.38	0.40	0.37	0.42
accel + clinical	KNN	0.53	0.40	0.51	0.36	0.44
accel + meds + demo	Decision Tree	0.60	0.44	0.45	0.44	0.47
accel + meds + demo + clin	XGBoost	0.62	0.40	0.46	0.37	0.44
accel + meds + demo + clin + diseases	**Decision Tree**	**0.60**	**0.44**	**0.40**	**0.51**	**0.48**

When analyzing inputs that included accelerometer data, some notes are worth mentioning. For the daytime, the combination of accelerometer data with analgesic medication data ("accel + meds") was particularly effective, with the MLP classifier achieving a ROC AUC of 0.59 and an F1 Score of 0.42. However, adding demographic data to this combination ("accel + meds + demo") improved the performance, with MLP again showing strong results with an ROC AUC of 0.61 and an F1 Score of 0.43. At night, the combination of accelerometer and analgesic medication data ("accel + meds") also performed well, with MLP reaching a ROC AUC of 0.69 and an F1 Score of 0.41. Adding demographic data to this combination ("accel + meds + demo") led to a slight decrease in performance, with the Decision Tree classifier achieving an ROC AUC of 0.60 and an F1 Score of 0.44. Interestingly, the best night-time performance involving accelerometer data was observed when all available inputs were combined ("accel + meds + demo + clin + diseases"), where the Decision Tree classifier yielded a ROC AUC of 0.60 and an F1 Score of 0.44.

When used as the sole input, it is noteworthy that the accelerometer, demographics, diseases, and clinical data yield results similar to those of a dummy classifier. Similarly, combinations of accelerometer data with the other mentioned data types also produce results akin to random guessing. However, the results improve significantly when more than three of these data types are combined, underscoring the value of integrating multiple data types.

4.4.2 Pain vs No Pain Classification

The experiments aimed to classify pain versus no pain in ICU patients using various inputs and their combinations. The results, detailed in Table 4.6, show that the combination of accelerometer and analgesic medication data ("accel + meds") produced the highest performance metrics for both day and night classifications. During the day, the CatBoost classifier achieved the best results with a Harmonic Mean of 0.76, an ROC AUC of 0.76, and an F1 Score of 0.75. At night, the Calibrated classifier was the top performer for the "accel + meds" input, with a Harmonic Mean of 0.80, an ROC AUC of 0.79, and an F1 Score of 0.77.

However, the improvements over using only analgesic medication were marginal. This suggests that analgesic medication alone is a robust predictor of pain versus no pain in ICU patients, and adding other variables does not significantly enhance the predictive accuracy.

4.4.3 Mild vs Moderate

The results for the binary classification of Mild vs Moderate, detailed in Table 4.7, show that the combination of accelerometer, medication, demographic, and clinical data ("accel

Table 4.6: Pain vs No Pain Classification

Input	Best Classifier	ROC AUC	F1 Score	Recall	Precision	Harmonic Mean
			Day			
	Dummy	0.50	0.30	0.47	0.22	0.33
accel	CatBoost	0.60	0.57	0.59	0.63	0.60
meds	Calibrated	0.75	0.75	0.75	0.76	0.75
demo	AdaBoost	0.52	0.52	0.52	0.52	0.52
diseases	Gaussian NB	0.62	0.6	0.64	0.7	0.64
clinical	QDA	0.57	0.57	0.57	0.57	0.57
accel + meds	**CatBoost**	**0.76**	**0.75**	**0.75**	**0.78**	**0.76**
accel + demo	Extra Tree	0.54	0.54	0.54	0.54	0.54
accel + diseases	XGBoost	0.57	0.57	0.57	0.58	0.57
accel + clinical	QDA	0.61	0.59	0.60	0.64	0.60
accel + meds + demo	LDA	0.67	0.67	0.67	0.67	0.67
accel + meds + demo + clin	Bagging	0.71	0.71	0.71	0.71	0.71
accel + meds + demo + clin + diseases	Logistic	0.69	0.69	0.70	0.72	0.70
			Night			
	Dummy	0.50	0.32	0.49	0.24	0.36
accel	Extra Tree	0.56	0.55	0.56	0.57	0.56
meds	Calibrated	0.78	0.77	0.78	0.83	0.79
demo	QDA	0.74	0.71	0.73	0.83	0.75
diseases	LGBM	0.64	0.64	0.64	0.65	0.64
clinical	Calibrated	0.56	0.5	0.57	0.62	0.56
accel + meds	**Calibrated**	**0.79**	**0.77**	**0.78**	**0.84**	**0.80**
accel + demo	Calibrated	0.59	0.57	0.59	0.61	0.59
accel + diseases	Logistic	0.60	0.59	0.60	0.61	0.60
accel + clinical	KNN	0.50	0.49	0.50	0.50	0.50
accel + meds + demo	CatBoost	0.74	0.73	0.74	0.79	0.75
accel + meds + demo + clin	Random Forest	0.77	0.76	0.77	0.84	0.78
accel + meds + demo + clin + diseases	**Random Forest**	**0.79**	**0.77**	**0.78**	**0.84**	**0.80**

Table 4.7: Mild vs Moderate

Input	Best Classifier	ROC AUC	F1 Score	Recall	Precision	Harmonic Mean
			Day			
	Dummy	0.50	0.54	0.67	0.45	0.53
accel	MLP	0.52	0.56	0.68	0.67	0.60
meds	AdaBoost	0.58	0.64	0.66	0.64	0.63
demo	QDA	0.59	0.63	0.63	0.64	0.62
diseases	CatBoost	0.50	0.54	0.67	0.45	0.53
clinical	KNN	0.57	0.63	0.70	0.68	0.64
accel + meds	Random Forest	0.64	0.66	0.66	0.68	0.66
accel + demo	Random Forest	0.62	0.59	0.58	0.67	0.61
accel + diseases	XGBoost	0.52	0.54	0.53	0.57	0.54
accel + clinical	CatBoost	0.58	0.64	0.66	0.63	0.62
accel + meds + demo	CatBoost	0.61	0.68	0.71	0.69	0.67
accel + meds + demo + clin	**AdaBoost**	**0.68**	**0.70**	**0.70**	**0.71**	**0.70**
accel + meds + demo + clin + diseases	Random Forest	0.67	0.69	0.68	0.70	0.68
			Night			
	Dummy	0.50	0.46	0.61	0.37	0.47
accel	Bagging	0.54	0.53	0.63	0.68	0.59
meds	XGBoost	0.63	0.63	0.62	0.65	0.63
demo	XGBoost	0.61	0.63	0.68	0.72	0.66
diseases	XGBoost	0.49	0.49	0.59	0.50	0.51
clinical	**LGBM**	**0.63**	**0.66**	**0.70**	**0.70**	**0.67**
accel + meds	MLP	0.54	0.55	0.63	0.65	0.59
accel + demo	AdaBoost	0.57	0.59	0.63	0.62	0.60
accel + diseases	Bagging	0.50	0.50	0.59	0.52	0.52
accel + clinical	LGBM	0.58	0.53	0.54	0.62	0.56
accel + meds + demo	QDA	0.61	0.64	0.66	0.65	0.64
accel + meds + demo + clin	QDA	0.62	0.65	0.67	0.66	0.65
accel + meds + demo + clin + diseases	Gaussian NB	0.50	0.51	0.59	0.53	0.53

+ meds + demo + clin") produced the highest performance metrics for daytime classifications. During the day, the AdaBoost classifier achieved the best results with a Harmonic Mean of 0.70, an ROC AUC of 0.68, and an F1 Score of 0.70. Notably, there was a considerable gain in this scenario compared to using only analgesic medication, which is different from the previous experiments. At night, the clinical data input alone ("clinical"), with the LGBM classifier as the best performer, resulted in the highest performance with a Harmonic Mean of 0.67, an ROC AUC of 0.63, and an F1 Score of 0.66.

4.4.4 Moderate vs Severe

The results, detailed in Table 4.8, show that the combination of accelerometer, medication, and demographic data ("accel + meds + demo") produced the highest performance metrics for daytime classifications. In this configuration, the AdaBoost classifier achieved the best results with a Harmonic Mean of 0.75, an ROC AUC of 0.74, and an F1 Score of 0.74.

At night, the demographic data input alone ("demo") with the XGBoost classifier resulted in the highest performance, with a Harmonic Mean of 0.73, an ROC AUC of 0.73, and an F1 Score of 0.73.

4.4.5 Mild vs Severe

The results, shown in Table 4.9, indicate that the combination of accelerometer, medication, and demographic data ("accel + meds + demo") produced the highest performance metrics for daytime classifications. The Random Forest classifier achieved the best results for this input with a Harmonic Mean of 0.68, an ROC AUC of 0.63, and an F1 Score of 0.69.

At night, the clinical data input alone ("clinical") with the Logistic classifier resulted in the highest performance, with a Harmonic Mean of 0.73, an ROC AUC of 0.69, and an F1 Score of 0.71.

4.4.6 Pain Variation

The results in Table 4.10 indicate that the best performance during the day was achieved by the demographic data input alone ("demo") with the Decision Tree classifier. This combination yielded a Harmonic Mean of 0.51, a ROC AUC of 0.49, and an F1 Score of 0.52. This highlights the effectiveness of using demographic data to differentiate between mild and severe pain levels in ICU patients during the day.

At night, the combination of medication data ("meds") with the LGBM classifier resulted in the highest performance, with a Harmonic Mean of 0.62, a ROC AUC of 0.72,

Table 4.8: Moderate vs Severe

Input	Best Classifier	ROC AUC	F1 Score	Recall	Precision	Harmonic Mean
Day						
	Dummy	0.50	0.30	0.47	0.22	0.34
accel	QDA	0.52	0.41	0.49	0.54	0.49
meds	LDA	0.73	0.74	0.74	0.74	0.74
demo	Logistic	0.63	0.63	0.64	0.66	0.64
diseases	Extra Tree	0.62	0.62	0.63	0.64	0.63
clinical	Decision Tree	0.55	0.54	0.54	0.55	0.54
accel + meds	AdaBoost	0.70	0.70	0.70	0.70	0.70
accel + demo	MLP	0.64	0.64	0.64	0.64	0.64
accel + diseases	Gaussian NB	0.60	0.59	0.61	0.61	0.60
accel + clinical	Extra Tree	0.53	0.53	0.53	0.53	0.53
accel + meds + demo	**AdaBoost**	**0.74**	**0.74**	**0.75**	**0.75**	**0.75**
accel + meds + demo + clin	AdaBoost	0.74	0.74	0.74	0.74	0.74
accel + meds + demo + clin + diseases	LDA	0.67	0.67	0.67	0.67	0.67
Night						
	Dummy	0.50	0.28	0.45	0.20	0.31
accel	Calibrated	0.60	0.57	0.58	0.63	0.59
meds	Extra Tree	0.61	0.56	0.58	0.64	0.60
demo	**XGBoost**	**0.73**	**0.73**	**0.73**	**0.74**	**0.73**
diseases	Bernoulli NB	0.54	0.52	0.57	0.56	0.55
clinical	Bagging	0.47	0.46	0.46	0.48	0.47
accel + meds	AdaBoost	0.58	0.58	0.60	0.59	0.59
accel + demo	XGBoost	0.66	0.66	0.66	0.67	0.66
accel + diseases	Bagging	0.58	0.56	0.57	0.59	0.57
accel + clinical	AdaBoost	0.55	0.53	0.58	0.58	0.56
accel + meds + demo	XGBoost	0.66	0.66	0.66	0.67	0.66
accel + meds + demo + clin	MLP	0.66	0.64	0.64	0.68	0.65
accel + meds + demo + clin + diseases	LDA	0.68	0.68	0.69	0.69	0.68

Table 4.9: Mild x Severe

Input	Best Classifier	ROC AUC	F1 Score	Recall	Precision	Harmonic Mean
			Day			
	Dummy	0.50	0.52	0.66	0.43	0.51
accel	XGBoost	0.58	0.64	0.66	0.63	0.63
meds	CatBoost	0.62	0.67	0.69	0.67	0.66
demo	Random Forest	0.60	0.66	0.69	0.68	0.66
diseases	Gaussian NB	0.61	0.64	0.64	0.64	0.63
clinical	KNN	0.57	0.62	0.64	0.62	0.61
accel + meds	LGBM	0.60	0.66	0.68	0.66	0.65
accel + demo	LDA	0.61	0.66	0.68	0.66	0.65
accel + diseases	LDA	0.56	0.62	0.67	0.64	0.62
accel + clinical	KNN	0.56	0.62	0.62	0.61	0.60
accel + meds + demo	**Random Forest**	**0.63**	**0.69**	**0.71**	**0.69**	**0.68**
accel + meds + demo + clin	Gaussian NB	0.65	0.67	0.67	0.68	0.67
accel + meds + demo + clin + diseases	CatBoost	0.60	0.66	0.69	0.68	0.66
			Night			
	Dummy	0.50	0.45	0.60	0.36	0.46
accel	Bernoulli NB	0.56	0.58	0.59	0.58	0.58
meds	AdaBoost	0.69	0.63	0.65	0.76	0.68
demo	QDA	0.64	0.65	0.65	0.65	0.65
diseases	GaussianNB	0.52	0.53	0.60	0.57	0.55
clinical	**Logistic**	**0.69**	**0.71**	**0.74**	**0.80**	**0.73**
accel + meds	Bernoulli NB	0.66	0.62	0.62	0.70	0.65
accel + demo	Bernoulli NB	0.52	0.53	0.60	0.57	0.55
accel + diseases	Decision Tree	0.58	0.56	0.56	0.60	0.57
accel + clinical	Extra Tree	0.60	0.59	0.59	0.62	0.60
accel + meds + demo	Decision Tree	0.62	0.6	0.6	0.65	0.62
accel + meds + demo + clin	Extra Tree	0.61	0.62	0.62	0.63	0.62
accel + meds + demo + clin + diseases	Bernoulli NB	0.56	0.57	0.63	0.64	0.60

Table 4.10: Pain Variation

Input	Best Classifier	ROC AUC	F1 Score	Recall	Precision	Harmonic Mean
			Day			
	Dummy	0.50	0.31	0.48	0.23	0.34
accel	LGBM	0.50	0.40	0.43	0.40	0.43
meds	Categorical NB	0.55	0.43	0.49	0.42	0.47
demo	**Decision Tree**	**0.49**	**0.52**	**0.53**	**0.51**	**0.51**
diseases	Gaussian NB	0.54	0.41	0.48	0.51	0.48
clinical	KNN	0.56	0.42	0.42	0.42	0.45
accel + meds	CatBoost	0.54	0.39	0.39	0.45	0.44
accel + demo	Extra Tree	0.49	0.36	0.37	0.35	0.38
accel + diseases	Gaussian NB	0.54	0.40	0.46	0.53	0.48
accel + clinical	Extra Trees	0.54	0.41	0.44	0.43	0.45
accel + meds + demo	CatBoost	0.49	0.38	0.38	0.46	0.42
accel + meds + demo + clin	CatBoost	0.50	0.39	0.38	0.46	0.43
accel + meds + demo + clin + diseases	Decision Tree	0.57	0.45	0.43	0.54	0.49
			Night			
	Dummy	0.50	0.32	0.49	0.24	0.36
accel	Gaussian NB	0.53	0.39	0.49	0.51	0.47
meds	**LGBM**	**0.72**	**0.58**	**0.59**	**0.61**	**0.62**
demo	Bagging	0.61	0.53	0.55	0.53	0.55
diseases	Bernoulli NB	0.64	0.50	0.54	0.52	0.55
clinical	Bagging	0.44	0.38	0.41	0.38	0.40
accel + meds	CatBoost	0.73	0.57	0.57	0.57	0.60
accel + demo	Extra Tree	0.55	0.49	0.49	0.51	0.51
accel + diseases	Bernoulli NB	0.59	0.46	0.48	0.45	0.49
accel + clinical	KNN	0.47	0.42	0.45	0.41	0.44
accel + meds + demo	AdaBoost	0.65	0.57	0.60	0.60	0.60
accel + meds + demo + clin	Decision Tree	0.65	0.57	0.57	0.58	0.59
accel + meds + demo + clin + diseases	Bernoulli NB	0.67	0.55	0.54	0.58	0.58

and an F1 Score of 0.58. This outcome suggests that medication data alone can be a strong predictor for classifying pain severity during nighttime.

4.5 Discussion

Table 4.11: Classifier Performance

Classifier	Appearances
CatBoost	15
Bernoulli NB	13
XGBoost	13
AdaBoost	13
Bagging	12
Decision Tree	11
QDA	11
Gaussian NB	11
MLP	9
Extra Tree	9
KNN	9
Logistic	6
Calibrated	6
Extra Trees	6
LDA	7
Random Forest	7
LGBM	7
Dummy	2
Categorical NB	1

As shown in Table 4.11, CatBoost, Bernoulli NB, XGBoost, AdaBoost and Bagging are the top 5 most frequently best-performing classifiers across various input scenarios we evaluated. These classifiers share several similarities that may contribute to their performance. Firstly, CatBoost, XGBoost, and AdaBoost are all boosting algorithms, which work by combining the predictions of several base estimators to improve robustness and accuracy. This boosting approach helps capture complex patterns in the data by focusing on difficult-to-predict instances, thus enhancing overall model performance. Secondly, Bernoulli NB and Bagging represent probabilistic and ensemble methods, respectively.

Table 4.12: Best Inputs for Different Pain Classification Experiments

Experiment	Day	Night
11 pain levels	meds	meds
Pain vs No pain	accel + meds	accel + meds
Mild vs Moderate	accel + meds + demo + clin	clinical
Moderate vs Severe	accel + meds + demo	demo
Mild vs Severe	accel + meds + demo	clinical
Pain Variation	demo	meds

Bernoulli NB, a type of Naive Bayes classifier, is particularly effective when features are binary and conditionally independent, which may suit the structured and diverse nature of the input data. Bagging, short for Bootstrap Aggregating, reduces variance by averaging multiple decision trees, thus providing stability and improving predictive accuracy.

Overall, the adaptability of these classifiers to different data characteristics and their inherent mechanisms to enhance prediction accuracy likely contribute to their consistent top performance.

Table 4.12 provides insights into the optimal inputs for the various pain classification experiments conducted. Overall, analgesic medications (meds) consistently emerged as a key input for multiple pain classification scenarios, both during the day and night. This suggests that medication data is crucial for understanding and predicting pain levels, likely due to its direct impact on pain modulation.

The combination of accelerometer data and medications (accel + meds) was particularly effective for distinguishing between pain and no pain across different times. This combination indicates that incorporating both physiological data and medication usage provides a comprehensive perspective on the patient's pain status.

During the day, more complex combinations of inputs, such as accelerometer data, medications, demographics, and clinical data (accel + meds + demo + clin), often provided the best results. This trend suggests that a multifaceted approach, integrating various types of data, is beneficial in capturing the nuances of pain experiences when patients are more active and exposed to a wider range of influences.

At night, single data sources like clinical data or demographic data alone were frequently sufficient for effective pain classification. This may indicate that during nighttime, when activity levels are lower and more stable, the complexity of input combinations is less critical for accurate pain assessment.

Chapter 5

Acuity Assessment

In critical care, acuity indicates the seriousness of a patient's illness and the priority of care required. Traditional methods of assessing acuity have relied on manual, threshold-based scoring systems such as Acute Physiology and Chronic Health Evaluation (APACHE) [73], Simplified Acute Physiology Score (SAPS) [76], Sequential Organ Failure Assessment (SOFA) [133] and Modified Early Warning Score (MEWS) [46], which analyze physiological parameters, laboratory results, and relevant clinical data to predict patient mortality and determine their care needs. These assessments aid clinical decision-making, resource allocation, staffing decisions, and efficient budget management. Recently, machine learning, especially deep learning techniques, has shown promise in improving acuity assessments by analyzing comprehensive data from electronic health records (EHR) [25, 71, 80, 103, 119]. However, they are restricted to EHR's physiological data and often overlook behavioral factors. To address this, recent studies have proposed augmenting EHR data with continuous sensing technology like wrist-worn accelerometers, offering more holistic insights into patient discharge as a proxy for acuity.

5.1 Methodology

In this work, we differ from the current literature by proposing to evaluate the viability of using accelerometer and EHR data to assess patients' acuity directly instead of using patient discharge status as a proxy. Following the same acuity phenotyping approach proposed by Ren et al. [108], the goal is to discern the patient's state as stable or unstable. To achieve this, we have developed an end-to-end deep learning pipeline based on accelerometer and EHR data (Figure 5.1).

We evaluated five different neural network architectures, namely, VGG [124], ResNet [58], MobileNet [61], SqueezeNet (SENet) [62], and a custom Transformer-based network [130] since both convolutional neural networks (CNNs) and Transformers architectures are well-accepted in the sensor-based human activity recognition field [66, 84, 89, 104, 105]. The CNN architecture can detect patterns regardless of their position in the sequence and extract both simple and complex movement patterns due to its hierarchical structure. On the other hand, Transformers are advantageous for pro-

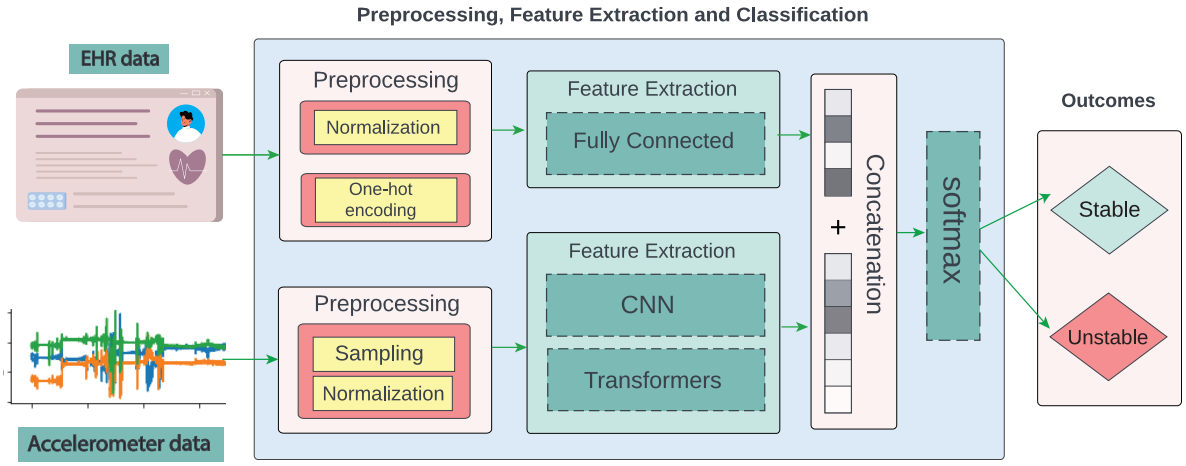


Figure 5.1: The proposed approach is an end-to-end neural network system that leverages accelerometer and EHR data to assess patient acuity, discerning between stable and unstable states.

cessing accelerometer data because their self-attention mechanism can capture long-term dependencies and weigh the importance of different elements in a temporal sequence [89]. Consequently, each patient's movement can be contextualized in relation to the other movements within a time window, directing the network's attention to the key movement patterns for assessing the patient's condition.

5.1.1 Deep Learning Models

The selection of models was grounded in their capabilities: VGG and ResNet for their depth, MobileNet, and SENet for their small number of parameters compared to ResNet and VGG, thus making them a suitable choice for edge deployment and for reducing the decision-making latency, which is crucial if deployed in the ICU setting. The transformer was selected for its unique attention mechanism, which enables modeling long-range dependencies in input signals. VGG, ResNet, MobileNet, and SENet were initially designed for image classification and required an architecture adaptation to suit accelerometer data. We tailored the original models to process 1D time series while preserving the fundamental layer-wise structure and defining characteristics. It entailed replacing 2D convolution, average pooling, and max pooling layers with their 1D counterparts and adjusting input channel configurations to match our data dimensions. For ResNet, SqueezeNet, and MobileNet, we retained essential components such as residual blocks (in ResNet), Squeeze-and-Excitation blocks (in SqueezeNet), and Depthwise Separable Convolution blocks (in MobileNet), with modifications primarily consisting of substituting 2D convolution and pooling filters with their 1D counterparts and updating channel parameters. The fully connected layers were kept unchanged. To further aggregate clinical and demographic features into the classification pipeline, we concatenated them with the dense features

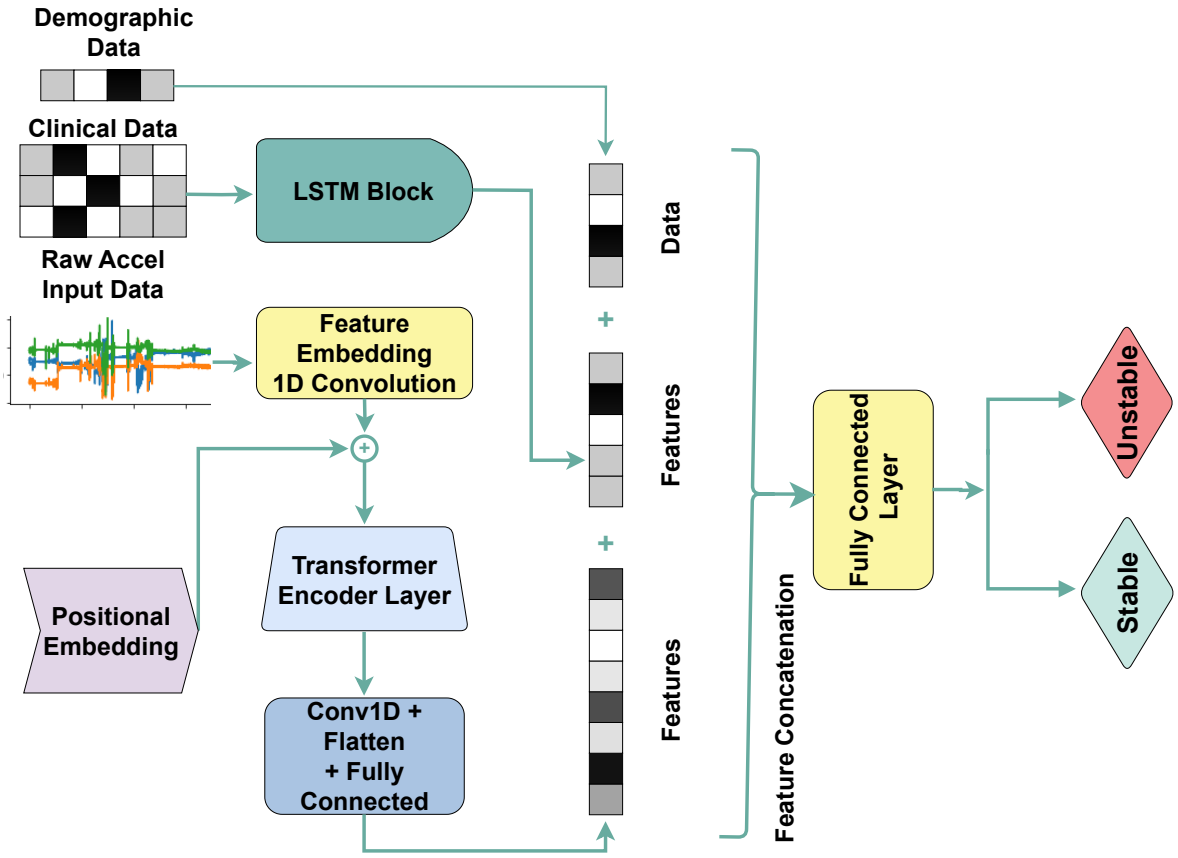


Figure 5.2: Transformer-based architecture for the acuity assessment task. "Accel" stands for Accelerometer Sensor, "Conv1D" stands for 1D Convolution, "Flatten" stands for 2D to 1D flattening Layer, and "Fully Connected" stands for fully connected layer.

extracted from the fully connected layer.

In contrast, Transformer architectures are innately suited for sequence data processing due to their self-attention mechanism and parallel processing capabilities. In our methodology, we extracted sequential feature embeddings from raw accelerometer sensor data using a feature embedding convolution layer with a kernel size of 5 and 64 channels. We then provided the extracted features, followed by adding positional encodings to capture the temporal order of the data into a Transformer encoding layer. We further processed the extracted contextual features through another set of convolution and fully connected layers to enable our downstream classification tasks. We concatenated clinical and demographic features with the dense features extracted from the fully connected layer, like the approach adopted in earlier models. The model architecture is demonstrated in Figure 5.2.

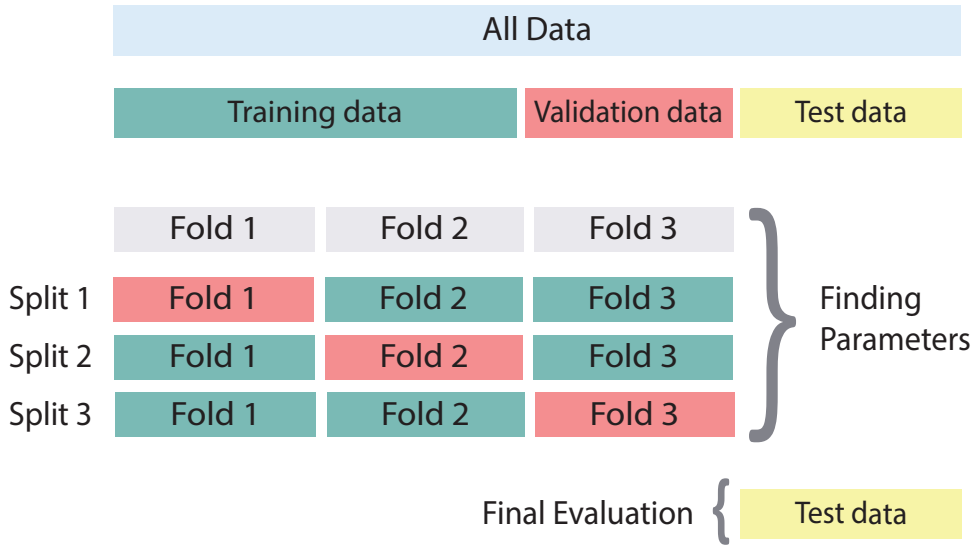


Figure 5.3: Evaluation protocol consists of a hold-out test set (30% of the data) for evaluation and a training set (70% of the data) split using 3-fold cross-validation for hyperparameter tuning.

5.2 Evaluation Protocol

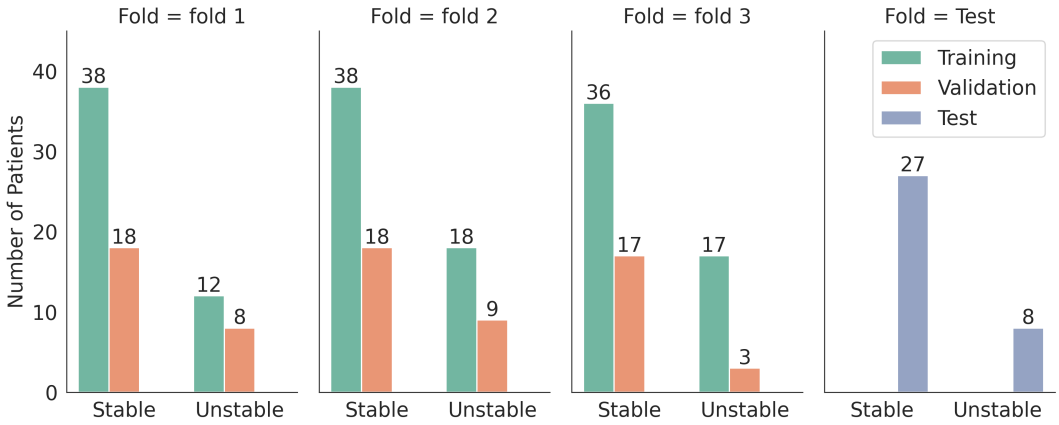
In assessing the deep learning models, we implemented a thorough evaluation protocol to ensure reliability and transparency, emphasizing subject independence. As shown in Figure 5.3, this protocol combined two established methods: 3-fold cross-validation and the holdout approach.

Initially, the holdout method divided the dataset into a development set (70%) and a separate holdout test set (30%), adhering to subject independence principles. The 3-fold cross-validation was then applied within the development set to facilitate robust hyperparameter optimization and guard against potential overfitting. This step was crucial for obtaining a reliable performance estimate, especially given our dataset's limited size. Within each fold, distinct training and validation datasets were created, ensuring that each patient's data was exclusively assigned to either the training or validation set used in that fold. This approach maintained the integrity of the evaluation process and upheld the principle of subject independence throughout.

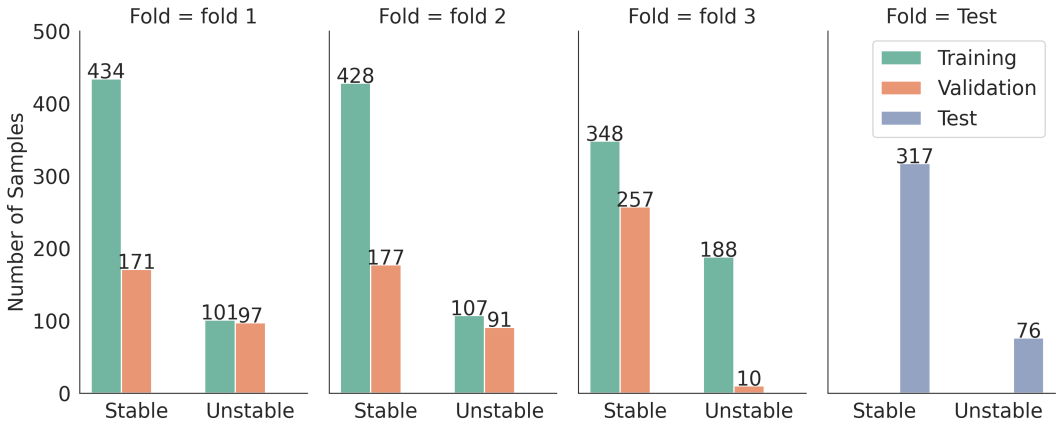
The models underwent training and validation using the development dataset to determine the most effective hyperparameters. After completing this step, they were assessed using the holdout test set to gauge their ability to generalize to unseen data.

Figure 5.4 shows our dataset's patient and sample distribution.

During the 3-fold cross-validation process, we utilized Optuna [4] to search over the hyperparameters rather than a traditional grid search. Optuna reduces the runtime by pruning fewer promising trials during runtime. For every set of hyperparameters, we maximized the area under the ROC curve (AUC) for each fold. After deriving the AUC for every fold, we calculated the mean and standard deviation of these values over all folds



(a) Patient Distribution



(b) Sample Distribution

Figure 5.4: Hold-out and 3-fold cross-validation distribution of patients and samples

of the 3-fold cross-validation. The hyperparameters yielding the highest mean validation AUC across all folds were deemed optimal and were used to train the final model.

In addition to using deep neural networks, we also incorporated the SOFA score as a rule-based scoring system into our evaluation process as a baseline. The SOFA score, well-established in assessing patients in ICUs, provides an objective and standardized means of tracking a patient's condition over time. These properties make the SOFA score an indicator of the acuity state assessment task. To measure the acuity states, we scaled the SOFA scores within the range of [0, 1] using min-max normalization. We treated these normalized scores as probability values and utilized the Youden index [140] to determine the optimal threshold for classifying the normalized scores and generating predictions.

Once the models were trained using optimized hyperparameters over the entire training cohort, we assessed their performance using bootstrapping with a replacement on a holdout test set. We created 100 synthetic bootstrapped versions of the holdout test set samples. These bootstrapped test sets were of the same length as the original test set. The model's performance was then calculated on all bootstraps. We reported the median and 95% confidence interval (CI) of several performance metrics: AUC, precision, sensitivity, specificity, and F1 score. The p-value was calculated to assess the statistical significance of the observed performance metrics values against a null hypothesis that was no better than the previous setups [6].

Finally, we performed SHAP (SHapley Additive exPlanations) [83] analysis on the best-performing models to interpret relative feature importance, providing insights into how various features contribute to model predictions. This analysis aids in understanding the model's decision-making process and can guide further refinement or feature engineering efforts.

5.3 Experiment Results

We evaluated the performance of five deep learning models on different combinations of feature sets: accelerometer data only (Accel), accelerometer data with demographics (Accel + Demo), accelerometer data with clinical information (Accel + Clinical), and a combination of accelerometer data, demographics, and clinical information (Accel + Demo + Clinical). We refer to demographics as the features of age, sex, race, height, and weight and as clinical data, the length of stay, blood pressure, heart rate, SpO2, pain score, Braden score, and cognitive status. In addition, we used the SOFA score as a baseline to compare performances across the rule-based and deep learning-based methods. We also evaluate the combination of demographics and clinical data (Demo + Clinical) to evaluate the accelerometer's contribution to the acuity status assessment. The results are summarized in Table 5.1.

Table 5.1: The best results reported as average and 95% confidence interval in each scenario.

	Model	AUC (95% CI, p-value)	Precision (95% CI, p-value)	Sensitivity (95% CI, p-value)	Specificity (95% CI, p-value)	F1-score (95% CI, p-value)
SOFA score	-	0.53 (0.48-0.58)	0.23 (0.19-0.28)	0.30 (0.22-0.38)	0.76 (0.69-0.82)	0.66 (0.61-0.72)
Demo + Clinical*	XGBoost	0.51 (0.45-0.57, 0.63)	0.65 (0.59-0.70, <0.05)	0.14 (0.06-0.21, <0.05)	0.74 (0.69-0.79, 0.65)	0.64 (0.59-0.68, 0.59)
Accel*	Squeezenet	0.62 (0.53-0.70, 0.07)	0.75 (0.71-0.79, <0.05)	0.47 (0.35-0.57, <0.01)	0.76 (0.71-0.81, 1.00)	0.72 (0.68-0.76, 0.08)
Accel + Demo**	Resnet	0.62 (0.52-0.69, 1.00)	0.76 (0.71-0.80, 0.76)	0.52 (0.40-0.63, 0.55)	0.74 (0.70-0.78, 0.55)	0.72 (0.68-0.76, 1.00)
Accel + Clinical**	Squeezenet	0.62 (0.52-0.69, 1.00)	0.75 (0.70-0.79, 1.00)	0.49 (0.37-0.57, 0.80)	0.74 (0.70-0.78, 0.55)	0.72 (0.68-0.75, 1.00)
Accel + Demo + Clinical***	Resnet	0.73 (0.63-0.78, 0.06)	0.80 (0.75-0.84, 0.12)	0.60 (0.48-0.70, 0.33)	0.79 (0.74-0.82, 0.08)	0.77 (0.73-0.80, <0.05)

Abbreviations: Accel - accelerometer data, demo - demographics (age, sex, race, height, weight, and length of stay), clinical - the clinical set of features (blood pressure, heart rate, spo2, pain score, Braden score, and acute brain dysfunction status)
Notes. * Indicates that the p-values for the setups were calculated by comparison with the SOFA score baseline, ** Indicates that the p-values for the setups were calculated by comparison with the Accel-only setup, *** Indicates that the p-values for the setups were calculated by comparison with the Accel + Clinical setup.

Table 5.2: Best hyperparameters for each scenario.

	Model	Number of Parameters (Million)	Batch size	Learning Rate	Weight Decay	Accelerometer Downsampling Factor
Accel	Squeezenet	4.33	16	2.11 x 10-4	9.23 x 10-6	1
Accel + Demo	Resnet	3.90	16	1.16 x 10-4	2.77 x 10-6	2
Accel + Clinical Data	Squeezenet	5.61	16	2.14 x 10-4	9.88 x 10-4	1
Accel + Demo + Clinical Data	Resnet	4.21	16	9.37 x 10-5	2.13 x 10-4	2

The performance of our baseline SOFA score-based predictor is notably limited, with suboptimal AUC (0.53), precision (0.23), sensitivity (0.30), and F1 score (0.66). However, the model demonstrates a relatively high specificity of 0.76.

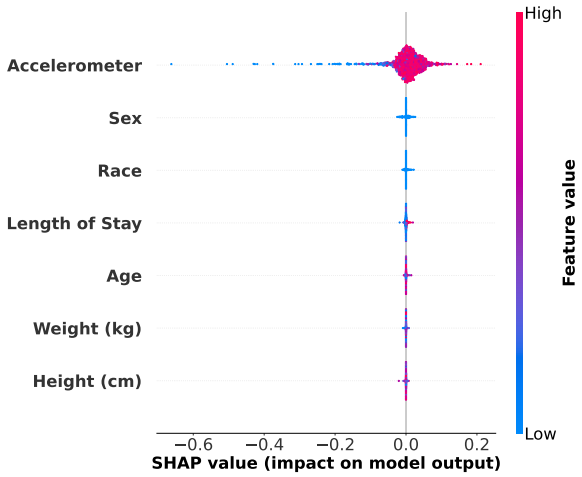
Incorporating accelerometer data (Accel) alone or combined with demographic and clinical variables (Accel + Demo, Accel + Clinical, Accel + Demo + Clinical) significantly improved the model's performance across all metrics. Adding accelerometer data improves AUC, precision, sensitivity, specificity, and F1-score compared to the SOFA score baseline.

Combining accelerometer data with demographic and clinical variables (Accel + Demo + Clinical) yields the best overall performance among the scenarios involving accelerometer data. This model achieves the highest AUC of 0.73, indicating superior discriminative ability compared to other scenarios. Moreover, it exhibits the highest precision (0.80), sensitivity (0.60), specificity (0.79), and F1 score (0.77). Our best setup demonstrated a relatively lower p-value.

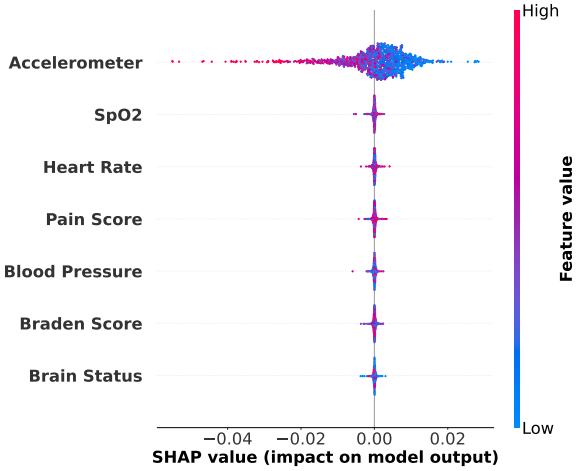
Optuna provided us with detailed information and hyperparameter selection suggestions. Table 5.2 outlines the best hyperparameters found by the search for each combination of feature sets. Table 5.3 comprehensively overviews the hyperparameters and their corresponding values. For the scenario where only accelerometer data was utilized (Accel), SqueezeNet architecture with a batch size 16, learning rate of 2.11 x 10-4, and weight decay of 9.23 x 10-6 yielded the best results. The accelerometer downsampling factor was set to 1. Incorporating demographic data and accelerometer data (Accel + Demo) led to selecting Resnet architecture with similar hyperparameters, except for a slightly lower learning rate of 1.16 x 10-4 and weight decay of 2.77 x 10 6. The downsampling factor was adjusted to 2 in this scenario. When clinical data was added to accelerometer data (Accel + Clinical Data), SqueezeNet architecture was again favored, with hyperparameters akin to the Accel scenario, except for a higher weight decay of 9.88 x 10-4 Finally, combining accelerometer, demographic, and clinical data (Accel + Demo + Clinical Data) led to the choice of Resnet architecture with a batch size of 16, a learning rate of 9.37 x 10-5, and weight decay of 2.13 x 10-4. The downsampling factor remained consistent with the Accel + Demo scenario at 2. We achieved the best Accel + Demo + Clinical scenario performance AUC of 0.73 (0.63-0.78).

Table 5.3: Overview of the hyperparameters and their respective values explored in the hyperparameter optimization

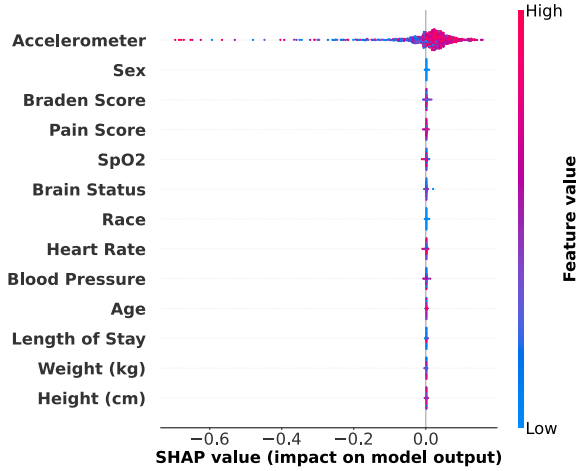
Hyperparameter	Values
Model	VGG, ResNet, MobileNet, SqueezeNet, and Transformers
Batch size	8, 16, 24 and 32
Learning rate	ranging from 10-5 to 10-1
Weight decay	ranging from 10-10 to 10-3
Accelerometer downsampling factor	1,2 and 4



(a) Accelerometer and Demo (Accel + Demo) Features.



(b) Accelerometer and Clinical (Accel + Clinical) Features.



(c) Accelerometer, Demo and Clinical (Accel + Demo + Clinical) Features

Figure 5.5: SHAP bee swarm plot illustrating feature importance for different types of feature combinations.

Figure 5.5 illustrates the application of SHAP interpretability analysis in detecting relative feature importance for three specific feature combinations: accelerometer and demographic features, accelerometer and clinical features, and accelerometer clinical and demographic features. This analysis is conducted on the best models obtained for each feature combination scenario. The significance of these features can aid in potential feature selection or assessing their impact on patient diagnosis.

5.4 Discussion

This study explored the potential of accelerometry and EHR data in directly determining patients' acuity state as an alternative to depending exclusively on rule-based scoring systems like the SOFA score. Our analysis revealed that the SOFA score-based predictor exhibited notable limitations, with suboptimal precision, sensitivity, and F1 score, reflecting its inadequacy in effectively evaluating patient conditions. Although the model demonstrated relatively high specificity, its AUC did not significantly surpass random chance, indicating the need for more sophisticated predictive models in clinical practice.

In contrast, incorporating accelerometer data alone or combined with demographic and clinical variables significantly enhanced model performance across all metrics. Notably, adding accelerometer data improved AUC, precision, sensitivity, specificity, and F1 score compared to the SOFA score baseline. These findings underscored the importance of integrating additional features beyond traditional clinical variables for accurate predictive modeling in medical settings. We believe that the additional features encompass aspects of patient physiology and functional status that are not effectively captured by SOFA inputs (or inputs for other traditional models such as APACHE and MEWS). The ability of accelerometer data to capture patient mobility and range of motion continuously can augment the current practice of hourly assessments that are subject to individual bias and are limited to observations of the bedside nurse. Therefore, we are enhancing predictive performance and adding nuance to patient assessment, enriching the overall assessment process. Among the scenarios involving accelerometer data, the model incorporating accelerometer data with demographics and clinical information (Accel + Demo + Clinical) demonstrated the best overall performance. This comprehensive approach yielded the highest AUC, precision, sensitivity, specificity, and F1 score, emphasizing the synergistic benefits of integrating multiple data types for predictive modeling. The robust performance of this model, with highly significant p-values, validated its effectiveness in predicting patient outcomes.

All models performed best with maintaining the original frequency of the accelerometer (downsampling scale of 1) which indicated that the long sequence was not a problem for the employed architectures. Notably, a bigger batch size was necessary for the models when clinical data was included in the model. It could indicate that the

added complexity introduced by the clinical data required more samples to be processed simultaneously for the model to effectively identify patterns, optimize the gradients, and achieve better convergence during training.

While our study sheds light on the use of accelerometers for acuity assessment, it is crucial to acknowledge limitations. Firstly, the generalizability of our findings may be constrained by the size and patient population of mainly white people studied at a single center, warranting validation on diverse datasets to enhance applicability. Despite the clinical research team's daily checks to ensure proper placement of accelerometer devices and requests to the nursing staff to document the times of device removal and application, it is probable that a small amount of data included in this study's analyses were recorded while the device was not placed on the patient. The exclusion of patients who died within 24 hours of recruitment, coupled with the inability to place study devices on the arms of patients with numerous intravenous and/or intraarterial lines or other equipment (i.e., wrist restraints), may have introduced bias through the exclusion of these high acuity patients from our cohort. Furthermore, the collection of accelerometry data and use of a motion-monitoring system may be unsuited for the acuity assessments of intubated and sedated patients since the active mobility in these patients is extremely limited.

Finally, it is essential to note that SHAP feature importance is correlated with model performance and may be vulnerable to misclassification due to overfitting, potentially leading to erroneous feature interpretations.

In Figures 5.5a, 5.5b, and 5.5c, it is evident that in the combinations of accelerometer with demographic features, accelerometer with clinical features, and all of them together, the accelerometer features exhibit higher importance compared to other features. The accelerometer features demonstrate a broad range of values in positive and negative directions, suggesting its strong indicative nature for acuity analysis, which aligns with our best model results.

Across scenarios utilizing only accelerometer data, accelerometer with demographic data, and accelerometer with clinical data, similar performance was observed on our test data, with an AUC of 0.62 for each combination. It suggests that clinical or demographic features alone, when combined with accelerometer data, do not significantly enhance the models' ability to classify our dataset. It underscores the critical role of accelerometer data in acuity assessment tasks.

Furthermore, combining accelerometer data with clinical and demographic data improved the AUC from 0.62 to 0.73, indicating an inter-feature dependency among these variables, which benefits our model.

Chapter 6

Conclusions

This study demonstrates the viability of using accelerometer data combined with electronic health records to predict pain in ICU patients. Analgesic medication data consistently emerged as a critical input for pain classification across different times of the day, underscoring the importance of medication data in understanding and predicting pain levels due to its direct impact on pain modulation.

Our findings reveal that multiclass classification for pain score and pain variation has demonstrated poor performance. In contrast, binary classification between groups of pain has yielded satisfactory results. Among the classifiers evaluated, boosting algorithms have consistently been the top three most frequently best-performing methods. Additionally, the inclusion of analgesic information has proven to be critical in the accurate prediction of pain. Our findings indicate that during the day, the integration of multiple input sources typically produced the most reliable results. However, at night, single data sources were often sufficient to achieve effective pain classification. In addressing our first research question, *"What potential do accelerometers have in assessing and quantifying pain levels in ICU patients?"*, it was found that accelerometer data did not significantly contribute to improving pain prediction.

While this study provides some insights, there remains room for improvement. Future work will focus on intra-patient pain prediction to better tailor pain management strategies to individual patients. Additionally, ensemble methods will be implemented to combine the best-performing classifiers, potentially enhancing predictive accuracy. Exploring deep learning methods, particularly those suited for processing temporal data, will also be a key area of investigation to further refine our pain prediction models.

Regarding acuity, our analysis revealed limitations in the SOFA-based predictor, highlighting the need for more sophisticated models in clinical practice. In addressing our first research question, *"In what ways might accelerometers aid in evaluating patient acuity in the ICU setting?"*, it was found that integrating accelerometer data, either alone or with demographic and clinical variables, significantly enhanced model performance, underscoring the importance of diverse data sources in predictive modeling. The model combining accelerometer data with demographics and clinical information exhibited the highest performance, validating its efficacy in predicting patient acuity. This underscores

the importance of a comprehensive approach to patient acuity assessment in critical care settings.

Accelerometer data emerges as an area of high potential for future research endeavors. Its utility extends to evaluating patient mobility, i.e., measuring the ability to change and control body position. Expanding this research to include the integration of additional clinical features, such as medication history, laboratory test results, and admission information, holds potential for further advancements. Moreover, utilizing multimodal models incorporating various pervasive sensing data like depth images, color RGB images, electromyography, sound pressure, and light levels offers opportunities to enhance model performance.

It is important to acknowledge that the observational studies for which this data was collected were conducted to be unobtrusive to patient care, and patients or their proxies were always allowed to opt out of or discontinue accelerometer data collection. Additional research is required to ascertain the reliability of mobility data for evaluating intubated and sedated patients. Moreover, further investigation is warranted to evaluate their seamless integration into clinical workflows, ensuring they don't add to the nursing workload or cause physician information overload. Additionally, thoughtful consideration needs to be given to how the outputs and assessments of these models can be communicated effectively, ensuring they offer actionable insights for healthcare providers.

Bibliography

[1] R. Abdullah and B. Fakieh. Health care employees' perceptions of the use of artificial intelligence applications: Survey study. *Journal of Medical Internet Research*, 2020. doi: 10.2196/17620.

[2] R. Abe, N. Bunya, T. Endo, Y. Fujino, K. Fujita, K. Fujizuka, Y. Hagiwara, J. Hamaguchi, Y. Hara, E. Hashiba, S. Hashimoto, N. Hattori, K. Hoshino, S. Ijuin, T. Ikeyama, S. Ichiba, W. Iwanaga, Y. Iwashita, M. Kanamoto, H. Kaneko, K. Kawamae, T. Kotani, Y. Koyama, K. Liu, T. Masuno, N. Morimura, T. Nakamura, M. Nakane, M. Nasu, O. Nishida, M. Nishimura, K. Ochiai, T. Ogura, S. Ohshimo, K. Oyama, J. Sasaki, R. Seo, T. Shimazu, K. Shimizu, H. Suzuki, S. Takauji, S. Takeda, I. Takeuchi, M. Takita, H. Taniguchi, and N. Shime. Save the icu and save lives during the covid-19 pandemic. *Journal of Intensive Care*, 8, 2020. doi: 10.1186/s40560-020-00456-1.

[3] A. Ahmed, A. Garcia-Agundez, I. Petrovic, F. Radaei, J. Fife, J. Zhou, H. Karas, S. Moody, J. Drake, R. N. Jones, et al. Delirium detection using wearable sensors and machine learning in patients with intracerebral hemorrhage. *Frontiers in Neurology*, 2023.

[4] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. *ArXiv*, 2019. arXiv:1907.10902.

[5] L. Alegria, P. Brockmann, P. Repetto, D. Leonard, R. Cadiz, F. Paredes, I. Rojas, A. Moya, V. Oviedo, P. García, et al. Improve sleep in critically ill patients: Study protocol for a randomized controlled trial for a multi-component intervention of environment control in the icu. *Plos one*, 2023.

[6] D. G. Altman and J. M. Bland. How to obtain the p value from a confidence interval. *BMJ*, 343, 2011. ISSN 0959-8138. doi: 10.1136/bmj.d2304. URL https://www.bmj.com/content/343/bmj.d2304.

[7] C. Arbour, M. Choinière, J. Topolovec-Vranic, C. G. Loiselle, and C. Gélinas. Can fluctuations in vital signs be used for pain assessment in critically ill patients with a traumatic brain injury? *Pain Research and Treatment*, 2014.

[8] B. Arthurs, V. Mohan, K. McGrath, G. Scholl, and J. Gold. Impact of passive laboratory alerts on navigating electronic health records in intensive care simulations. *Sage Open*, 8:215824401877438, 2018. doi: 10.1177/2158244018774388.

[9] J. Bai, B. He, H. Shou, V. Zipunnikov, T. A. Glass, and C. M. Crainiceanu. Normalization and extraction of interpretable metrics from raw accelerometry data. *Biostatistics*, 15(1):102–116, 2014.

[10] S. Bandyopadhyay, A. Cecil, J. Sena, A. Davidson, Z. Guan, S. Nerella, J. Zhang, K. Khezeli, B. Armfield, A. Bihorac, et al. Predicting risk of delirium from ambient noise and light information in the icu. *arXiv preprint arXiv:2303.06253*, 2023.

[11] N. Bergstrom, B. J. Braden, A. Laguzza, and V. Holman. The braden scale for predicting pressure sore risk. *Nursing research*, 1987.

[12] S. Bhattacharyay, J. Rattray, M. Wang, P. H. Dziedzic, E. Calvillo, H. B. Kim, E. Joshi, P. Kudela, R. Etienne-Cummings, and R. D. Stevens. Decoding accelerometry for classification and prediction of critically ill patients with severe brain injury. *Scientific reports*, 11(1):23654, 2021.

[13] P. E. Bijur, W. Silver, and E. J. Gallagher. Reliability of the visual analog scale for measurement of acute pain. *Academic emergency medicine*, 8(12):1153–1157, 2001.

[14] F. F. Bourbonnais, S. Malone-Tucker, and D. Dalton-Kischel. Intensive care nurses' assessment of pain in patients who are mechanically ventilated: How a pilot study helped to influence practice. *Canadian Journal of Critical Care Nursing*, 2016.

[15] H. Brown, J. Terrence, P. Vasquez, D. W. Bates, and E. Zimlichman. Continuous monitoring in an inpatient medical-surgical unit: a controlled clinical trial. *The American journal of medicine*, 127(3):226–232, 2014.

[16] T. K. Bucknall. Medical error and decision making: Learning from the past and present in intensive care. *Australian critical care*, 23(3):150–156, 2010.

[17] T. Bui, M. Grayson, K. Hofer, K. McGuire, M. Morrow, N. Rodammer, A. Farooque, L. E. Barnes, and S. Patek. Remote patient monitoring for improving outpatient care of patients at risk for sepsis. In *2016 IEEE Systems and Information Engineering Design Symposium (SIEDS)*, pages 136–141, 2016. doi: 10.1109/SIEDS.2016.7489286.

[18] C. Caetano, J. Sena, F. Brémond, J. A. Dos Santos, and W. R. Schwartz. Skelemotion: A New Representation of Skeleton Joint Sequences based on Motion Information for 3D Action Recognition. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019.

[19] G. Carra, J. I. Salluh, F. J. da Silva Ramos, and G. Meyfroidt. Data-driven icu management: Using big data and algorithms to improve outcomes. *Journal of Critical Care*, 2020.

[20] M. Chan, K. Pai, S. Su, M. Wang, C. Wu, and W. Chao. Explainable machine learning to predict long-term mortality in critically ill ventilated patients: a retrospective study in central taiwan. *BMC Medical Informatics and Decision Making*, 22, 2022. doi: 10.1186/s12911-022-01817-6.

[21] G. Chanques, E. Viel, J.-M. Constantin, B. Jung, S. de Lattre, J. Carr, M. Cissé, J.-Y. Lefrant, and S. Jaber. The measurement of pain in intensive care unit: comparison of 5 self-report intensity scales. *PAIN®*, 2010.

[22] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

[23] Y. Chen, P. Moreira, W. Liu, M. Michelle, N. T. Le Ha, and A. Wang. Is there a gap between artificial intelligence applications and priorities in health care and nursing management? *Journal of Nursing Management*, 2022. doi: 10.1111/jonm.13851.

[24] Z. Chen, R. Ansari, and D. Wilkie. Learning pain from action unit combinations: a weakly supervised approach via multiple instance learning. *IEEE transactions on affective computing*, 2019.

[25] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2017.

[26] E. Chou, M. Tan, C. Zou, M. Guo, A. Haque, A. Milstein, and L. Fei-Fei. Privacy-preserving action recognition for smart hospitals using low-resolution depth images. *ArXiv*, abs/1811.09950, 2018.

[27] S. Coomber, C. Todd, G. Park, P. Baxter, J. Firth-Cozens, and S. Shore. Stress in uk intensive care unit doctors. *British journal of anaesthesia*, 89(6):873–881, 2002.

[28] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[29] A. Davoudi, K. R. Malhotra, B. Shickel, S. Siegel, S. Williams, M. Ruppert, E. Bihorac, T. Ozrazgat-Baslanti, P. J. Tighe, A. Bihorac, et al. Intelligent icu for autonomous patient monitoring using pervasive sensing and deep learning. *Scientific reports*, 9(1):8020, 2019.

[30] A. Davoudi, T. Ozrazgat-Baslanti, P. J. Tighe, A. Bihorac, and P. Rashidi. Pain and physical activity association in critically ill patients. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020.

[31] A. Davoudi, P. J. Tighe, A. Bihorac, and P. Rashidi. Posture recognition in the critical care settings using wearable devices. *arXiv preprint arXiv:2110.02768*, 2021.

[32] A. De Jong, N. Molinari, S. De Lattre, C. Gniadek, J. Carr, M. Conseil, M.-P. Susbielles, B. Jung, S. Jaber, and G. Chanques. Decreasing severe pain and serious adverse events while moving intensive care unit patients: a prospective interventional study (the nurse-do project). *Critical care*, 2013.

[33] L. Delaney, E. Litton, K. Melehan, H. Huang, V. Lopez, and F. Van Haren. The feasibility and reliability of actigraphy to monitor sleep in intensive care patients: an observational study. *CRITICAL CARE*, 2021.

[34] L. J. Delaney, E. Litton, H. C. Huang, V. Lopez, and F. M. van Haren. The accuracy of simple, feasible alternatives to polysomnography for assessing sleep in intensive care: An observational study. *Australian Critical Care*, 36(3):361–369, 2023.

[35] F. Demrozi, G. Pravadelli, P. J. Tighe, A. Bihorac, and P. Rashidi. Joint distribution and transitions of pain and activity in critically ill patients. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020.

[36] S. V. Desai, T. J. Law, and D. M. Needham. Long-term complications of critical care. *Critical care medicine*, 39(2):371–379, 2011.

[37] Y. Dikkema, N. Mouton, K. Gerrits, T. Valk, M. van der Steen-Diepenrink, H. Eshuis, H. Houdijk, C. van der Schans, A. Niemeijer, and M. Nieuwenhuis. Identification and quantification of activities common to intensive care patients; development and validation of a dual-accelerometer-based algorithm. *Sensors*, 2023.

[38] I. Dirgová Luptáková, M. Kubovčík, and J. Pospíchal. Wearable sensor-based human activity recognition with transformer model. *Sensors*, 2022.

[39] A. V. Dorogush, V. Ershov, and A. Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[40] C. Duffield, D. Diers, L. O'Brien-Pallas, C. Aisbett, M. Roche, M. King, and K. Aisbett. Nursing staffing, nursing workload, the work environment and patient outcomes. *Applied Nursing Research*, 2011.

[41] S. Erden, N. Demir, G. A. Ugras, U. Arslan, and S. Arslan. Vital signs: Valid indicators to assess pain in intensive care unit patients? an observational, descriptive study. *Nursing & Health Sciences*, 2018.

[42] P. Esmaeilzadeh, T. Mirzaei, and S. Dharanikota. Patients' perceptions toward human–artificial intelligence interaction in health care: Experimental study. *Journal of Medical Internet Research*, 2021. doi: 10.2196/25856.

[43] G. Ferrante, A. Licari, S. Fasola, and S. Grutta. Artificial intelligence in the diagnosis of pediatric allergic diseases. *Pediatric Allergy and Immunology*, 32:405–413, 2020. doi: 10.1111/pai.13419.

[44] H. FR, van Haaren JHL, K. R, van Delden RW, V. PH, and G. JG. Objective quantification of in-hospital patient mobilization after cardiacsurgery using accelerometers: Selection, use, and analysis. *Sensors (Basel)*, 2021.

[45] S. Gandotra, D. Files, K. Shields, M. Berry, and R. Bakhru. Activity levels in survivors of the intensive care unit. *PHYSICAL THERAPY*, 2021.

[46] J. Gardner-Thorpe, N. Love, J. Wrightson, S. Walsh, and N. Keeling. The value of modified early warning score (mews) in surgical in-patients: a prospective observational study. *The Annals of The Royal College of Surgeons of England*, 88(6): 571–575, 2006.

[47] C. Gélinas and C. Johnston. Pain assessment in the critically ill ventilated adult: validation of the critical-care pain observation tool and physiologic indicators. *The Clinical journal of pain*, 2007.

[48] C. Gelinas, K. A. Puntillo, A. M. Joffe, and J. Barr. A validated approach to evaluating psychometric properties of pain assessment tools for use in nonverbal critically ill adults. In *Seminars in Respiratory and Critical Care Medicine*, 2013.

[49] E. Georgiou, M. Hadjibalassi, E. Lambrinou, P. Andreou, and E. D. Papathanassoglou. The impact of pain assessment on critically ill patients' outcomes: a systematic review. *BioMed research international*, 2015.

[50] E. Georgiou, M. Hadjibalassi, E. Lambrinou, P. Andreou, and E. D. E. Papathanassoglou. The impact of pain assessment on critically ill patients' outcomes: A systematic review. *BioMed Research International*, 2015.

[51] G. R. Gonçalves, J. Sena, W. R. Schwartz, and C. A. Caetano. Pixel-level class-agnostic object detection using texture quantization. In *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, volume 1, pages 31–36. IEEE, 2022.

[52] F. González-Seguel, A. Camus-Molina, M. Leiva-Corvalán, K. P. Mayer, and J. Leppe. Uninterrupted actigraphy recording to quantify physical activity and sedentary behaviors in mechanically ventilated adults: A feasibility prospective observational study. *Journal of Acute Care Physical Therapy*, 2022.

[53] Y. Guo, Z. Hao, S. Zhao, J. Gong, and F. Yang. Artificial intelligence in health care: Bibliometric analysis. *Journal of Medical Internet Research*, 2020. doi: 10.2196/18228.

[54] P. Gupta, J. L. Martin, A. Malhotra, J. Bergstrom, M. A. Grandner, and B. B. Kamdar. Circadian rest-activity misalignment in critically ill medical intensive care unit patients. *Journal of sleep research*, 2022.

[55] M. H and G. M. Sedentary time in older adults with acute cardiovascular disease. *CJC Open*, 2022.

[56] M. B. Happ. Communicating with mechanically ventilated patients: state of the science. *AACN Advanced Critical Care*, 2001.

[57] C. T. Hartrick, J. P. Kovan, and S. Shapiro. The numeric rating scale for clinical pain measurement: a ratio measure? *Pain Practice*, 3(4):310–316, 2003.

[58] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[59] B. Hirose, K. Ikeda, D. Yamamoto, E. Tsuda, R. Yamauchi, T. Hozuki, Y. Masuda, and T. Imai. Measurement of excitation-contraction coupling time in critical illness myopathy. *Clinical Neurophysiology*, 2022.

[60] A. L. Holder, S. P. Shashikumar, G. Wardi, T. G. Buchman, and S. Nemati. A locally optimized data-driven tool to predict sepsis-associated vasopressor use in the icu. *Critical Care Medicine*, 49(12):e1196–e1205, 2021.

[61] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[62] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[63] S. Jaiswal, S. Bagsic, E. Takata, B. Kamdar, S. Ancoli-Israel, and R. Owens. Actigraphy-based sleep and activity measurements in intensive care unit patients

randomized to ramelteon or placebo for delirium prevention. *SCIENTIFIC REPORTS*, 2023.

[64] A. Jalali, D. Bender, M. Rehman, V. Nadkanri, and C. Nataraj. Advanced analytics for outcome prediction in intensive care units. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016.

[65] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2017. doi: 10.1136/svn-2017-000101.

[66] A. Jordao, A. C. Nazare Jr, J. Sena, and W. R. Schwartz. Human activity recognition based on wearable sensor data: A standardization of the state-of-the-art. *arXiv preprint arXiv:1806.05226*, 2018.

[67] B. Kamdar, J. Fine, M. Pavini, S. Ardren, S. Burns, J. Bates, R. McGinnis, V. Pandian, B. Lin, D. Needham, and R. Stapleton. Phase i pilot safety and feasibility of a novel restraint device for critically ill patients requiring mechanical ventilation. *JOURNAL OF THE INTENSIVE CARE SOCIETY*, 2023.

[68] R. L. Kane, T. A. Shamliyan, C. Mueller, S. Duval, and T. J. Wilt. The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis. *Medical care*, pages 1195–1204, 2007.

[69] H. Kemp, C. Bantel, F. Gordon, S. Brett, PLAN, SEARCH, H. Laycock, S. Bampoe, C. Bantel, M. Gooneratne, et al. Pain assessment in int ensive care (paint): an observational study of physician-documented pain assessment in 45 intensive care units in the united kingdom. *Anaesthesia*, 2017.

[70] A. G. Khalf, k. Abdelhafez, and s. khalab. Health care providers' perception about artificial intelligence applications. *Assiut Scientific Nursing Journal*, 2022. doi: 10.21608/asnj.2022.144712.1397.

[71] S. Y. Kim, S. Kim, J. Cho, Y. S. Kim, I. S. Sol, Y. Sung, I. Cho, M. Park, H. Jang, Y. H. Kim, et al. A deep learning model for real-time mortality prediction in critically ill children. *Critical care*, 23(1):1–10, 2019.

[72] E. Kipnis, D. Ramsingh, M. Bhargava, E. Dincer, M. Cannesson, A. Broccard, B. Vallet, K. Bendjelid, R. Thibault, et al. Monitoring in the intensive care. *Critical care research and practice*, 2012, 2012.

[73] W. A. Knaus, J. E. Zimmerman, D. P. Wagner, E. A. Draper, and D. E. Lawrence. Apache—acute physiology and chronic health evaluation: a physiologically based classification system. *Critical care medicine*, 9(8):591–597, 1981.

[74] A. Korompeli, N. Kavrochorianou, and P. Myrianthefs. Rest-activity circadian rhythm and light exposure using wrist actigraphy in icu patients. *Archives in Neurology & Neuroscience*, 2022.

[75] H. Ksouri, P.-Y. Balanant, J.-M. Tadié, G. Heraud, I. Abboud, N. Lerolle, A. Novara, J.-Y. Fagon, and C. Faisy. Impact of morbidity and mortality conferences on analysis of mortality and critical events in intensive care practice. *American Journal of Critical Care*, 19(2):135–145, 2010.

[76] J.-R. Le Gall, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, and D. Villers. A simplified acute physiology score for icu patients. *Critical care medicine*, 12(11):975–977, 1984.

[77] L. Lehmkuhl, H. Olsen, J. Brond, M. Rothmann, P. Dreyer, and E. Jespersen. Daily variation in physical activity during mechanical ventilation and stay in the intensive care unit. *ACTA ANAESTHESIOLOGICA SCANDINAVICA*, 2023.

[78] C. Li, Z. Zhang, Y. Ren, H. Nie, Y. Lei, H. Qiu, Z. Xu, and X. Pu. Machine learning based early mortality prediction in the emergency department. *International Journal of Medical Informatics*, 155:104570, 2021.

[79] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable ai: a review of machine learning interpretability methods. *Entropy*, 23:18, 2020. doi: 10.3390/e23010018.

[80] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[81] R. Liu, A. A. Ramli, H. Zhang, E. Henricson, and X. Liu. An overview of human activity recognition using wearable sensors: Healthcare and artificial intelligence. In *International Conference on Internet of Things*. Springer, 2021.

[82] X. Liu, P. Hu, W. Yeung, Z. Zhang, V. Ho, C. Liu, C. Dumontier, P. J. Thoral, Z. Mao, D. Cao, et al. Illness severity assessment of older adults in critical illness using machine learning (elder-icu): an international multicentre study with subgroup bias evaluation. *The Lancet Digital Health*, 2023.

[83] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[84] I. D. Luptáková, M. Kubovčík, and J. Pospíchal. Wearable sensor-based human activity recognition with transformer model. *Sensors*, 22(5), jan 2022. doi: 10.3390/s22051911.

[85] L. Mamykina, D. K. Vawdrey, and G. Hripcsak. How do residents spend their shift time? a time and motion study with a particular focus on the use of computers. *Academic medicine: journal of the Association of American Medical Colleges*, 91 (6):827, 2016.

[86] S. Messe, S. Kasner, B. Cucchiara, M. McGarvey, S. Cummings, M. Acker, N. Desai, P. Atluri, G. Wang, B. Jackson, and J. Weimer. Derivation and validation of an algorithm to detect stroke using arm accelerometry data. *JOURNAL OF THE AMERICAN HEART ASSOCIATION*, 2023.

[87] C. Munro, Z. Liang, M. Elias, M. Ji, X. Chen, and K. Calero. Sleep and activity patterns are altered during early critical illness in mechanically ventilated adults. *DIMENSIONS OF CRITICAL CARE NURSING*, 2021.

[88] S. Nerella, J. Cupka, M. Ruppert, P. Tighe, A. Bihorac, and P. Rashidi. Pain action unit detection in critically ill patients. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021.

[89] S. Nerella, S. Bandyopadhyay, J. Zhang, M. Contreras, S. Siegel, A. Bumin, B. Silva, J. Sena, B. Shickel, A. Bihorac, K. Khezeli, and P. Rashidi. Transformers in healthcare: A survey, 2023.

[90] S. Nerella, Z. Guan, S. Siegel, J. Zhang, K. Khezeli, A. Bihorac, and P. Rashidi. Ai-enhanced intensive care unit: Revolutionizing patient care with pervasive sensing. *arXiv preprint arXiv:2303.06252*, 2023.

[91] S. Nerella, Z. Guan, S. Siegel, J. Zhang, K. Khezeli, A. Bihorac, and P. Rashidi. Ai-enhanced intensive care unit: Revolutionizing patient care with pervasive sensing. *arXiv preprint arXiv:2303.06252*, 2023.

[92] I. Newton and N. Chittenden. *Newton's Principia: the mathematical principles of natural philosophy*. Geo. P. Putnam, 1850.

[93] M. Odhner, D. Wegman, N. Freeland, A. Steinmetz, and G. L. Ingersoll. Assessing pain control in nonverbal critically ill adults. *Dimensions of Critical Care Nursing*, 2003.

[94] I. Ohu, P. K. Benny, S. Rodrigues, and J. N. Carlson. Applications of machine learning in acute care research. *Journal of the American College of Emergency Physicians Open*, 2020. doi: 10.1002/emp2.12156.

[95] A. Pagnamenta, G. Rabito, A. Arosio, A. Perren, R. Malacrida, F. Barazzoni, and G. Domenighetti. Adverse event reporting in adult intensive care units and the

impact of a multifaceted intervention on drug-related adverse events. *Annals of intensive Care*, 2(1):1–10, 2012.

[96] J. P. Parreco, A. E. Hidalgo, A. D. Badilla, O. Ilyas, and R. Rattan. Predicting central line-associated bloodstream infections and mortality using supervised machine learning. *Journal of critical care*, 45:156–162, 2018.

[97] S. M. Parry, C. L. Granger, S. Berney, J. Jones, L. Beach, D. El-Ansary, R. Koopman, and L. Denehy. Assessment of impairment and activity limitations in the critically ill: a systematic review of measurement instruments and their clinimetric properties. *Intensive care medicine*, 41:744–762, 2015.

[98] J.-F. Payen, J.-L. Bosson, G. Chanques, J. Mantz, J. Labarere, and for the DOLOREA Investigators. Pain assessment is associated with decreased duration of mechanical ventilation in the intensive care unit. *Anesthesiology*, 2009.

[99] S. Petersen, M. Abdulkareem, and T. Leiner. Artificial intelligence will transform cardiac imaging—opportunities and challenges. *Frontiers in Cardiovascular Medicine*, 6, 2019. doi: 10.3389/fcvm.2019.00133.

[100] R. C. Polomano, K. T. Galloway, M. L. Kent, H. Brandon-Edwards, K. N. Kwon, C. Morales, and C. T. Buckenmaier III. Psychometric testing of the defense and veterans pain rating scale (dvprs): a new pain scale for military population. *Pain Medicine*, 17(8):1505–1519, 2016.

[101] S. Prakash, J. Balaji, A. Joshi, and K. Surapaneni. Ethical conundrums in the application of artificial intelligence (ai) in healthcare—a scoping review of reviews. *Journal of Personalized Medicine*, 12:1914, 2022. doi: 10.3390/jpm12111914.

[102] R. Raj, K. Ussavarungsi, and K. Nugent. Accelerometer-based devices can be used to monitor sedation/agitation in the intensive care unit. *Journal of critical care*, 29 5:748–52, 2014. doi: 10.1016/j.jcrc.2014.05.014.

[103] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[104] E. Ramanujam, T. Perumal, and S. Padmavathi. Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors Journal*, 2021.

[105] S. Raziani and M. Azimbagirad. Deep cnn hyperparameter optimization algorithms for sensor-based human activity recognition. *Neuroscience Informatics*, 2022.

[106] S. Reddy, S. Allan, S. Coghlan, and P. Cooper. A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, 27:491–497, 2019. doi: 10.1093/jamia/ocz192.

[107] N. Ren, X. Zhao, and X. Zhang. Mortality prediction in icu using a stacked ensemble model. *Computational and Mathematical Methods in Medicine*, 2022:1–12, 2022. doi: 10.1155/2022/3938492.

[108] Y. Ren et al. Development of computable phenotype to identify and characterize transitions in acuity status in intensive care unit. *arXiv*, 2020.

[109] T. C. Rollinson, B. Connolly, D. J. Berlowitz, and S. Berney. Physical activity of patients with critical illness undergoing rehabilitation in intensive care and on the acute ward: An observational cohort study. *Australian Critical Care*, 2022.

[110] H. B. Rubins and M. A. Moskowitz. Complications of care in a medical intensive care unit. *Journal of General Internal Medicine*, 5:104–109, 1990.

[111] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. doi: 10.1038/s42256-019-0048-x.

[112] D. Saraswat, P. Bhattacharya, A. Verma, V. Prasad, S. Tanwar, G. Sharma, P. Bokoro, and R. Sharma. Explainable ai for healthcare 5.0: opportunities and challenges. *Ieee Access*, 10:84486–84517, 2022. doi: 10.1109/access.2022.3197671.

[113] D. S. Schujmann, T. T. Gomes, A. C. Lunardi, and C. Fu. Factors associated with functional decline in an intensive care unit: a prospective study on the level of physical activity and clinical factors. *Revista Brasileira de Terapia Intensiva*, 2022.

[114] J. Sena, J. Barreto, C. Caetano, G. Cramer, and W. R. Schwartz. Human Activity Recognition based on Smartphone and Wearable Sensors using Multiscale DCNN Ensemble. *Neurocomputing*, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.04.151. URL https://www.sciencedirect.com/science/article/pii/S0925231220317823.

[115] J. Sena, A. Jordão, and W. R. Schwartz. A Content-Based Late Fusion Approach applied to Pedestrian Detection. *Journal of Visual Communication and Image Representation*, page 103091, 2021. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2021.103091. URL https://www.sciencedirect.com/science/article/pii/S1047320321000559.

[116] J. Sena, S. Bandyopadhyay, A. Davidson, Z. Guan, J. Barreto, T. Ozrazgat-Baslanti, P. Tighe, W. R. Schwartz, A. Bihorac, and P. Rashidi. Diurnal pain classification in

critically ill patients using machine learning on accelerometry and analgesic data. In *IEEE BIBM IEEE, International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023.

[117] J. Sena, M. T. Mostafiz, J. Zhang, A. Davidson, S. Bandyopadhyay, R. Yuanfang, T. Ozrazgat-Baslanti, B. Shickel, T. Loftus, W. R. Schwartz, A. Bihorac, and P. Rashidi. The potential of wearable sensors for assessing patient acuity in intensive care unit (icu), 2023.

[118] J. Sena, M. T. Mostafiz, J. Zhang, A. E. Davidson, S. Bandyopadhyay, S. Nerella, Y. Ren, T. Ozrazgat-Baslanti, B. Shickel, T. Loftus, W. R. Schwartz, A. Bihorac, and P. Rashidi. Wearable sensors in patient acuity assessment in critical care. *Frontiers in Neurology*, 15, 2024. ISSN 1664-2295. doi: 10.3389/fneur.2024.1386728. URL https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2024.1386728.

[119] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[120] B. Shickel, T. J. Loftus, L. Adhikari, T. Ozrazgat-Baslanti, A. Bihorac, and P. Rashidi. Deepsofa: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific reports*, 9(1):1879, 2019.

[121] B. Shickel, A. Davoudi, T. Ozrazgat-Baslanti, M. Ruppert, A. Bihorac, and P. Rashidi. Deep multi-modal transfer learning for augmented patient acuity assessment in the intelligent icu. *Frontiers in digital health*, 2021.

[122] Shimmer. Shimmer3 EMG Unit. [Apparatus]. https://shimmersensing.com/product/shimmer3-emg-unit/.

[123] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[124] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[125] J. Smichenko, T. Shochat, and A. Zisberg. Assessment of sleep duration and number of awakenings based on ankle and wrist actigraphy in medical hospitalized older patients. *Biological Research For Nursing*, 2022.

[126] D. Soydaner. Attention mechanism in neural networks: Where it comes and where it goes. *Neural Computing & Applications*, 2022.

[127] A. Thrush, M. Rozek, and J. L. Dekerlegand. The clinical utility of the functional status score for the intensive care unit (fss-icu) at a long-term acute care hospital: a prospective cohort study. *Physical therapy*, 92(12):1536–1545, 2012.

[128] M. Turmell, A. Cooley, T. L. Yap, J. Alderden, V. K. Sabol, J.-R. A. Lin, and S. M. Kennerly. Improving pressure injury prevention by using wearable sensors to cue critical care patient repositioning. *American Journal of Critical Care*, 2022.

[129] A. Valentin, M. Schiffinger, J. Steyrer, C. Huber, and G. Strunk. Safety climate reduces medication and dislodgement errors in routine intensive care practice. *Intensive care medicine*, 39:391–398, 2013.

[130] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[131] A. C. Verceles and E. R. Hager. Use of accelerometry to monitor physical activity in critically ill subjects: a systematic review. *Respiratory care*, 2015.

[132] M. Verdon, P. Merlani, T. Perneger, and B. Ricou. Burnout in a surgical icu team. *Intensive care medicine*, 34:152–156, 2008.

[133] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. G. Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix), 1996.

[134] M. Watson, B. Hasan, and N. Moubayed. Agree to disagree: when deep learning models with identical architectures produce distinct explanations. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. doi: 10.1109/wacv51458.2022.00159.

[135] C. M. Wollschlager and A. R. Conrad. Common complications in critically ill patients. *Disease-a-month*, 34(5):225–293, 1988.

[136] D. H. Wong, Y. Gallegos, M. B. Weinger, S. Clack, J. Slagle, and C. T. Anderson. Changes in intensive care unit nurse task activity after installation of a third-generation intensive care unit information system. *Critical care medicine*, 31(10): 2488–2494, 2003.

[137] H. Wu, N.-K. Chan, C. J. P. Zhang, and W. Ming. The role of the sharing economy and artificial intelligence in health care: Opportunities and challenges. *Journal of Medical Internet Research*, 2019. doi: 10.2196/13469.

[138] S. Xiao et al. Two-stream transformer network for sensor-based human activity recognition. *Neurocomputing*, 2022.

[139] Q. Xu, W. Xie, B. Liao, C. Hu, L. Qin, Z. Yang, H. Xiong, Y. Lv, Y. Zhou, and A. Luo. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: a systematic review. *Journal of Healthcare Engineering*, 2023:1–13, 2023. doi: 10.1155/2023/9919269.

[140] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[141] M. Zaid, S. Ahmad, A. Suliman, M. Camazine, I. Weber, J. Sheppard, M. Popescu, J. Keller, L. Despins, M. Skubic, et al. Noninvasive cardiovascular monitoring based on electrocardiography and ballistocardiography: a feasibility study on patients in the surgical intensive care unit. In *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021.

Appendix A

Pain Classification Results

Table A.1: DVPRS 11 Classes - Day - Accelerometer

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
LogisticRegression	0.63	0.42	0.55	0.39	0.48	0.09
LinearDiscriminantAnalysis	0.55	0.43	0.50	0.43	0.48	0.10
MLPClassifier	0.63	0.43	0.47	0.39	0.47	0.71
BernoulliNB	0.58	0.42	0.47	0.38	0.45	0.06
CatBoostClassifier	0.60	0.39	0.44	0.36	0.43	8.23
CalibratedClassifierCV	0.62	0.38	0.54	0.29	0.42	0.20
AdaBoostClassifier	0.55	0.39	0.40	0.38	0.42	0.19
DummyClassifier	0.50	0.38	0.54	0.29	0.41	0.05
RandomForestClassifier	0.55	0.36	0.41	0.33	0.40	0.29
ExtraTreesClassifier	0.53	0.36	0.38	0.35	0.39	0.21
XGBClassifier	0.54	0.33	0.34	0.33	0.37	0.79
LGBMClassifier	0.55	0.32	0.34	0.31	0.36	0.73
DecisionTreeClassifier	0.51	0.33	0.32	0.34	0.36	0.06
BaggingClassifier	0.49	0.31	0.35	0.29	0.35	0.09
ExtraTreeClassifier	0.51	0.30	0.29	0.33	0.34	0.06
GaussianNB	0.53	0.05	0.06	0.55	0.10	0.06
QuadraticDiscriminantAnalysis	0.00	0.00	0.01	0.00	0.00	0.03

Table A.2: DVPRS 11 Classes - Day -Accelerometer and Clinical

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
DummyClassifier	0.5	0.38	0.54	0.29	0.40	0.05
CalibratedClassifierCV	0.48	0.38	0.54	0.29	0.40	0.27
LogisticRegression	0.48	0.38	0.53	0.29	0.40	0.11
LinearDiscriminantAnalysis	0.48	0.38	0.53	0.29	0.40	0.06
RandomForestClassifier	0.47	0.35	0.46	0.33	0.39	0.38
MLPClassifier	0.49	0.36	0.49	0.3	0.39	0.65
ExtraTreesClassifier	0.5	0.35	0.47	0.3	0.39	0.23
KNeighborsClassifier	0.47	0.36	0.44	0.31	0.38	0.08
AdaBoostClassifier	0.5	0.35	0.46	0.29	0.38	0.23
CatBoostClassifier	0.5	0.34	0.41	0.3	0.37	7.47
BernoulliNB	0.47	0.35	0.45	0.28	0.37	0.06
XGBClassifier	0.47	0.3	0.33	0.32	0.34	0.96
LGBMClassifier	0.48	0.28	0.28	0.29	0.32	0.82
BaggingClassifier	0.51	0.26	0.24	0.28	0.30	0.16
ExtraTreeClassifier	0.49	0.25	0.23	0.29	0.29	0.05
DecisionTreeClassifier	0.51	0.23	0.19	0.34	0.28	0.06
GaussianNB	0.52	0.05	0.07	0.35	0.10	0.06

Table A.3: DVPRS 11 Classes - Day -Accelerometer and Demographics

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
BaggingClassifier	0.57	0.42	0.54	0.39	0.47	0.14
LinearDiscriminantAnalysis	0.56	0.39	0.51	0.43	0.46	0.06
DecisionTreeClassifier	0.57	0.42	0.47	0.42	0.46	0.06
RandomForestClassifier	0.57	0.39	0.54	0.35	0.44	0.34
CatBoostClassifier	0.56	0.39	0.52	0.34	0.43	5.29
XGBClassifier	0.56	0.4	0.49	0.34	0.43	0.91
LogisticRegression	0.58	0.39	0.53	0.32	0.43	0.11
ExtraTreesClassifier	0.56	0.39	0.54	0.31	0.42	0.21
QuadraticDiscriminantAnalysis	0.53	0.39	0.54	0.31	0.42	0.1
MLPClassifier	0.5	0.39	0.5	0.33	0.42	0.69
CalibratedClassifierCV	0.56	0.38	0.54	0.29	0.41	0.34
DummyClassifier	0.5	0.38	0.54	0.29	0.40	0.05
LGBMClassifier	0.58	0.35	0.35	0.38	0.40	1.03
BernoulliNB	0.54	0.35	0.45	0.29	0.39	0.06
ExtraTreeClassifier	0.51	0.3	0.27	0.35	0.34	0.05
AdaBoostClassifier	0.5	0.3	0.3	0.31	0.34	0.24
GaussianNB	0.53	0.06	0.06	0.36	0.11	0.06

Table A.4: DVPRS 11 Classes - Day -Accelerometer and Diseases

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
DummyClassifier	0.5	0.38	0.54	0.29	0.402750518	0.05
BernoulliNB	0.46	0.38	0.49	0.32	0.40115097	0.06
RandomForestClassifier	0.47	0.38	0.54	0.29	0.397639365	0.39
ExtraTreeClassifier	0.5	0.37	0.49	0.3	0.39694935	0.05
CalibratedClassifierCV	0.46	0.38	0.54	0.29	0.395819368	0.29
MLPClassifier	0.44	0.36	0.45	0.35	0.394871795	0.74
LinearDiscriminantAnalysis	0.48	0.37	0.45	0.32	0.394739767	0.1
LogisticRegression	0.46	0.36	0.46	0.33	0.393859459	0.11
ExtraTreesClassifier	0.44	0.38	0.54	0.29	0.391986466	0.23
BaggingClassifier	0.5	0.36	0.48	0.29	0.387995912	0.17
CatBoostClassifier	0.47	0.37	0.49	0.29	0.387617389	12.81
XGBClassifier	0.45	0.35	0.47	0.29	0.375400017	0.95
LGBMClassifier	0.46	0.35	0.42	0.29	0.368314489	1.16
DecisionTreeClassifier	0.5	0.28	0.28	0.28	0.314606742	0.07
QuadraticDiscriminantAnalysis	0.46	0.24	0.24	0.26	0.278679612	0.11
AdaBoostClassifier	0.49	0.18	0.15	0.24	0.21704091	0.23
GaussianNB	0.46	0.06	0.06	0.43	0.105728285	0.06

Table A.5: DVPRS 11 Classes - Day -Accelerometer and Medications

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
MLPClassifier	0.59	0.42	0.48	0.41	0.47	0.81
LogisticRegression	0.62	0.41	0.51	0.38	0.46	0.12
LinearDiscriminantAnalysis	0.54	0.42	0.5	0.39	0.45	0.12
CalibratedClassifierCV	0.6	0.38	0.54	0.3	0.42	0.32
BernoulliNB	0.56	0.39	0.41	0.37	0.42	0.06
ExtraTreesClassifier	0.56	0.38	0.45	0.33	0.41	0.24
RandomForestClassifier	0.55	0.37	0.46	0.33	0.41	0.36
DummyClassifier	0.5	0.38	0.54	0.29	0.40	0.05
CatBoostClassifier	0.58	0.35	0.36	0.35	0.39	6
AdaBoostClassifier	0.52	0.36	0.38	0.34	0.39	0.25
XGBClassifier	0.53	0.33	0.33	0.35	0.37	1.03
LGBMClassifier	0.57	0.3	0.28	0.39	0.36	1.07
BaggingClassifier	0.53	0.31	0.32	0.32	0.35	0.18
DecisionTreeClassifier	0.51	0.25	0.21	0.33	0.29	0.07
ExtraTreeClassifier	0.5	0.24	0.2	0.32	0.28	0.05
QuadraticDiscriminantAnalysis	0.51	0.15	0.13	0.32	0.21	0.12
GaussianNB	0.52	0.03	0.05	0.42	0.07	0.06

Table A.6: DVPRS 11 Classes - Day -Accelerometer, Medications, Demographics and Clinical

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
LogisticRegression	0.57	0.45	0.5	0.41	0.48	0.2
MLPClassifier	0.58	0.42	0.44	0.4	0.45	2.61
LinearDiscriminantAnalysis	0.49	0.42	0.46	0.43	0.45	0.14
CatBoostClassifier	0.59	0.39	0.46	0.33	0.42	6.59
CalibratedClassifierCV	0.58	0.38	0.54	0.29	0.41	0.47
LGBMClassifier	0.56	0.37	0.43	0.35	0.41	1.13
XGBClassifier	0.57	0.37	0.44	0.33	0.41	1.1
QuadraticDiscriminantAnalysis	0.51	0.38	0.44	0.34	0.41	0.13
RandomForestClassifier	0.58	0.37	0.51	0.29	0.41	0.35
DummyClassifier	0.5	0.38	0.54	0.29	0.40	0.05
BaggingClassifier	0.51	0.37	0.45	0.31	0.40	0.17
BernoulliNB	0.56	0.35	0.33	0.37	0.39	0.13
ExtraTreesClassifier	0.53	0.35	0.45	0.28	0.38	0.23
ExtraTreeClassifier	0.51	0.33	0.32	0.35	0.36	0.05
DecisionTreeClassifier	0.48	0.31	0.34	0.3	0.35	0.07
AdaBoostClassifier	0.45	0.32	0.36	0.28	0.34	0.27
GaussianNB	0.5	0.02	0.05	0.55	0.05	0.06

Table A.7: DVPRS 11 Classes - Day -Accelerometer, Medications, Demographics, Clinical and Diseases

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
LogisticRegression	0.57	0.42	0.5	0.38	0.46	0.31
RandomForestClassifier	0.6	0.4	0.55	0.35	0.45	0.34
LinearDiscriminantAnalysis	0.53	0.41	0.43	0.43	0.45	0.19
CatBoostClassifier	0.57	0.41	0.48	0.37	0.45	7.56
MLPClassifier	0.57	0.41	0.45	0.38	0.44	2.81
BaggingClassifier	0.56	0.4	0.47	0.36	0.44	0.19
ExtraTreesClassifier	0.57	0.39	0.52	0.33	0.43	0.24
BernoulliNB	0.58	0.39	0.4	0.4	0.43	0.12
QuadraticDiscriminantAnalysis	0.54	0.4	0.52	0.32	0.43	0.14
CalibratedClassifierCV	0.57	0.38	0.54	0.29	0.41	0.47
LGBMClassifier	0.55	0.37	0.46	0.31	0.40	1.24
DummyClassifier	0.5	0.38	0.54	0.29	0.40	0.05
XGBClassifier	0.61	0.36	0.41	0.32	0.40	1.19
DecisionTreeClassifier	0.51	0.34	0.36	0.33	0.37	0.07
AdaBoostClassifier	0.43	0.35	0.42	0.3	0.37	0.29
ExtraTreeClassifier	0.49	0.15	0.11	0.31	0.19	0.05
GaussianNB	0.49	0.05	0.07	0.42	0.10	0.06

Table A.8: DVPRS 11 Classes - Day -Accelerometer, Medications, Demographics

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
MLPClassifier	0.61	0.43	0.49	0.4	0.47	3.01
CatBoostClassifier	0.61	0.41	0.49	0.38	0.46	6.67
LogisticRegression	0.59	0.42	0.5	0.37	0.46	0.12
ExtraTreesClassifier	0.55	0.41	0.52	0.34	0.44	0.23
LinearDiscriminantAnalysis	0.49	0.41	0.45	0.4	0.43	0.12
ExtraTreeClassifier	0.54	0.4	0.41	0.4	0.43	0.05
RandomForestClassifier	0.56	0.38	0.54	0.3	0.42	0.34
CalibratedClassifierCV	0.59	0.38	0.54	0.29	0.42	0.42
XGBClassifier	0.56	0.38	0.42	0.35	0.41	1.06
BaggingClassifier	0.56	0.37	0.42	0.33	0.40	0.18
DummyClassifier	0.5	0.38	0.54	0.29	0.40	0.05
LGBMClassifier	0.59	0.35	0.36	0.37	0.40	1.23
DecisionTreeClassifier	0.53	0.36	0.38	0.36	0.40	0.07
BernoulliNB	0.57	0.33	0.33	0.34	0.37	0.08
QuadraticDiscriminantAnalysis	0.53	0.3	0.27	0.35	0.34	0.12
AdaBoostClassifier	0.52	0.15	0.13	0.39	0.21	0.3
GaussianNB	0.51	0.03	0.05	0.43	0.07	0.06

Table A.9: DVPRS 11 Classes - Day -Clinical

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
ExtraTreesClassifier	0.48	0.38	0.44	0.35	0.41	0.17
DummyClassifier	0.50	0.38	0.54	0.29	0.41	0.05
LogisticRegression	0.48	0.38	0.54	0.29	0.40	0.10
LinearDiscriminantAnalysis	0.48	0.38	0.54	0.29	0.40	0.06
CalibratedClassifierCV	0.48	0.38	0.54	0.29	0.40	0.20
BernoulliNB	0.47	0.38	0.54	0.29	0.40	0.06
RandomForestClassifier	0.47	0.37	0.42	0.35	0.40	0.21
XGBClassifier	0.46	0.37	0.41	0.35	0.40	0.71
DecisionTreeClassifier	0.46	0.36	0.43	0.34	0.39	0.06
ExtraTreeClassifier	0.47	0.36	0.41	0.35	0.39	0.06
AdaBoostClassifier	0.49	0.36	0.43	0.31	0.39	0.19
KNeighborsClassifier	0.52	0.35	0.37	0.35	0.39	0.07
MLPClassifier	0.50	0.36	0.49	0.28	0.39	0.83
CatBoostClassifier	0.46	0.36	0.41	0.33	0.38	1.58
LGBMClassifier	0.46	0.35	0.42	0.33	0.38	0.59
BaggingClassifier	0.45	0.35	0.42	0.33	0.38	0.08
GaussianNB	0.51	0.02	0.06	0.41	0.06	0.06
QuadraticDiscriminantAnalysis	0.00	0.01	0.06	0.00	0.00	0.01

Table A.10: DVPRS 11 Classes - Day -Demographics

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
BaggingClassifier	0.61	0.42	0.54	0.34	0.45	0.08
CatBoostClassifier	0.60	0.39	0.52	0.32	0.43	1.86
MLPClassifier	0.59	0.40	0.50	0.33	0.43	0.84
XGBClassifier	0.55	0.39	0.50	0.33	0.42	0.65
LinearDiscriminantAnalysis	0.55	0.39	0.53	0.31	0.42	0.06
ExtraTreesClassifier	0.58	0.39	0.54	0.30	0.42	0.16
ExtraTreeClassifier	0.55	0.39	0.42	0.37	0.42	0.06
LGBMClassifier	0.56	0.39	0.52	0.31	0.42	0.60
RandomForestClassifier	0.59	0.38	0.54	0.29	0.42	0.20
BernoulliNB	0.55	0.39	0.54	0.30	0.42	0.06
LogisticRegression	0.57	0.38	0.54	0.29	0.42	0.12
DummyClassifier	0.50	0.38	0.54	0.29	0.41	0.05
CalibratedClassifierCV	0.46	0.38	0.54	0.29	0.40	0.23
AdaBoostClassifier	0.53	0.36	0.43	0.30	0.39	0.18
DecisionTreeClassifier	0.48	0.33	0.36	0.31	0.36	0.06
KNeighborsClassifier	0.50	0.30	0.30	0.31	0.34	0.08
QuadraticDiscriminantAnalysis	0.56	0.02	0.05	0.01	0.03	0.06
GaussianNB	0.51	0.01	0.04	0.00	0.01	0.06

Table A.11: DVPRS 11 Classes - Day -Diseases

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
DummyClassifier	0.50	0.38	0.54	0.29	0.41	0.05
BaggingClassifier	0.49	0.38	0.54	0.29	0.40	0.08
LGBMClassifier	0.45	0.39	0.54	0.30	0.40	0.42
BernoulliNB	0.47	0.38	0.53	0.30	0.40	0.06
CategoricalNB	0.47	0.38	0.53	0.30	0.40	0.06
CalibratedClassifierCV	0.48	0.38	0.54	0.29	0.40	0.21
ExtraTreesClassifier	0.47	0.38	0.54	0.29	0.40	0.16
RandomForestClassifier	0.47	0.38	0.54	0.29	0.40	0.18
DecisionTreeClassifier	0.46	0.38	0.54	0.29	0.40	0.05
XGBClassifier	0.44	0.38	0.54	0.29	0.39	0.62
CatBoostClassifier	0.44	0.38	0.54	0.29	0.39	13.16
LinearDiscriminantAnalysis	0.48	0.36	0.46	0.31	0.39	0.10
ExtraTreeClassifier	0.45	0.36	0.45	0.30	0.38	0.06
LogisticRegression	0.45	0.36	0.46	0.29	0.38	0.11
MLPClassifier	0.43	0.36	0.46	0.29	0.37	0.82
AdaBoostClassifier	0.43	0.36	0.46	0.29	0.37	0.19
GaussianNB	0.46	0.04	0.03	0.09	0.05	0.06
QuadraticDiscriminantAnalysis	0.00	0.38	0.54	0.29	0.00	0.01

Table A.12: DVPRS 11 Classes - Day -Medications

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
LogisticRegression	0.63	0.42	0.55	0.39	0.48	0.09
LinearDiscriminantAnalysis	0.55	0.43	0.50	0.43	0.48	0.10
MLPClassifier	0.63	0.43	0.47	0.39	0.47	0.71
BernoulliNB	0.58	0.42	0.47	0.38	0.45	0.06
CatBoostClassifier	0.60	0.39	0.44	0.36	0.43	8.23
CalibratedClassifierCV	0.62	0.38	0.54	0.29	0.42	0.20
AdaBoostClassifier	0.55	0.39	0.40	0.38	0.42	0.19
DummyClassifier	0.50	0.38	0.54	0.29	0.41	0.05
RandomForestClassifier	0.55	0.36	0.41	0.33	0.40	0.29
ExtraTreesClassifier	0.53	0.36	0.38	0.35	0.39	0.21
XGBClassifier	0.54	0.33	0.34	0.33	0.37	0.79
LGBMClassifier	0.55	0.32	0.34	0.31	0.36	0.73
DecisionTreeClassifier	0.51	0.33	0.32	0.34	0.36	0.06
BaggingClassifier	0.49	0.31	0.35	0.29	0.35	0.09
ExtraTreeClassifier	0.51	0.30	0.29	0.33	0.34	0.06
GaussianNB	0.53	0.05	0.06	0.55	0.10	0.06
QuadraticDiscriminantAnalysis	0.00	0.00	0.01	0.00	0.00	0.03

Table A.13: DVPRS 11 Classes - Night -Accelerometer

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
MLPClassifier	0.49	0.36	0.52	0.32	0.41	0.44
LogisticRegression	0.49	0.35	0.51	0.3	0.39	0.09
LinearDiscriminantAnalysis	0.5	0.36	0.51	0.28	0.39	0.06
BernoulliNB	0.52	0.35	0.49	0.28	0.38	0.06
CalibratedClassifierCV	0.5	0.35	0.51	0.26	0.38	0.2
DummyClassifier	0.5	0.35	0.51	0.26	0.38	0.05
AdaBoostClassifier	0.49	0.35	0.51	0.26	0.37	0.19
RandomForestClassifier	0.49	0.34	0.5	0.26	0.37	0.3
ExtraTreesClassifier	0.48	0.33	0.47	0.26	0.36	0.21
XGBClassifier	0.47	0.32	0.43	0.26	0.35	0.76
LGBMClassifier	0.49	0.32	0.41	0.26	0.35	1.09
KNeighborsClassifier	0.49	0.32	0.43	0.25	0.35	0.06
BaggingClassifier	0.46	0.32	0.4	0.27	0.35	0.12
CatBoostClassifier	0.48	0.31	0.38	0.26	0.34	8.8
ExtraTreeClassifier	0.49	0.18	0.16	0.24	0.22	0.05
DecisionTreeClassifier	0.5	0.17	0.15	0.25	0.22	0.06
QuadraticDiscriminantAnalysis	0.48	0.09	0.08	0.18	0.13	0.06
GaussianNB	0.48	0.05	0.04	0.21	0.08	0.06

Table A.14: DVPRS 11 Classes - Night -Accelerometer and Clinical

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
KNeighborsClassifier	0.53	0.4	0.51	0.36	0.44	0.07
ExtraTreesClassifier	0.51	0.36	0.48	0.29	0.39	0.19
AdaBoostClassifier	0.58	0.35	0.49	0.27	0.39	0.19
BernoulliNB	0.54	0.35	0.48	0.28	0.39	0.06
RandomForestClassifier	0.49	0.35	0.48	0.29	0.38	0.29
CatBoostClassifier	0.5	0.35	0.46	0.28	0.38	8.61
LinearDiscriminantAnalysis	0.51	0.35	0.51	0.26	0.38	0.06
DummyClassifier	0.5	0.35	0.51	0.26	0.38	0.05
LogisticRegression	0.5	0.35	0.51	0.26	0.38	0.09
MLPClassifier	0.49	0.35	0.51	0.26	0.37	0.47
CalibratedClassifierCV	0.49	0.35	0.51	0.26	0.37	0.2
XGBClassifier	0.52	0.33	0.4	0.29	0.37	0.65
BaggingClassifier	0.52	0.28	0.29	0.27	0.32	0.11
DecisionTreeClassifier	0.47	0.27	0.3	0.26	0.31	0.06
LGBMClassifier	0.52	0.26	0.27	0.28	0.31	0.73
ExtraTreeClassifier	0.47	0.23	0.22	0.25	0.27	0.05
GaussianNB	0.51	0.03	0.05	0.1	0.06	0.06

Table A.15: DVPRS 11 Classes - Night -Accelerometer and Demographics

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
LinearDiscriminantAnalysis	0.45	0.38	0.53	0.37	0.42	0.06
ExtraTreesClassifier	0.53	0.39	0.5	0.33	0.42	0.19
BaggingClassifier	0.48	0.38	0.46	0.35	0.41	0.12
CatBoostClassifier	0.5	0.38	0.47	0.33	0.41	7.83
BernoulliNB	0.51	0.37	0.49	0.3	0.40	0.06
XGBClassifier	0.51	0.36	0.48	0.29	0.39	0.68
RandomForestClassifier	0.51	0.35	0.51	0.26	0.38	0.27
DummyClassifier	0.5	0.35	0.51	0.26	0.38	0.05
AdaBoostClassifier	0.56	0.32	0.28	0.42	0.37	0.2
CalibratedClassifierCV	0.43	0.35	0.51	0.26	0.36	0.23
LogisticRegression	0.41	0.35	0.51	0.26	0.36	0.1
DecisionTreeClassifier	0.52	0.29	0.29	0.3	0.33	0.06
LGBMClassifier	0.54	0.28	0.29	0.29	0.32	0.87
ExtraTreeClassifier	0.48	0.25	0.25	0.26	0.29	0.05
MLPClassifier	0.37	0.22	0.23	0.2	0.24	0.51
GaussianNB	0.5	0.01	0.06	0.01	0.02	0.06
QuadraticDiscriminantAnalysis		0.01	0.07	0.01	0.01	0.01

Table A.16: DVPRS 11 Classes - Night -Accelerometer and Diseases

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
BernoulliNB	0.61	0.38	0.4	0.37	0.42	0.06
LinearDiscriminantAnalysis	0.45	0.39	0.51	0.31	0.40	0.06
CatBoostClassifier	0.51	0.36	0.51	0.28	0.39	12.74
XGBClassifier	0.54	0.34	0.43	0.31	0.39	0.7
MLPClassifier	0.43	0.37	0.5	0.29	0.38	0.48
DummyClassifier	0.5	0.35	0.51	0.26	0.38	0.05
RandomForestClassifier	0.49	0.35	0.51	0.26	0.37	0.28
AdaBoostClassifier	0.45	0.35	0.51	0.27	0.37	0.19
CalibratedClassifierCV	0.42	0.35	0.51	0.26	0.36	0.21
LogisticRegression	0.37	0.35	0.51	0.28	0.36	0.1
ExtraTreesClassifier	0.46	0.33	0.42	0.27	0.35	0.2
DecisionTreeClassifier	0.49	0.31	0.37	0.27	0.34	0.06
LGBMClassifier	0.47	0.26	0.3	0.23	0.29	0.87
BaggingClassifier	0.47	0.25	0.27	0.25	0.29	0.12
ExtraTreeClassifier	0.5	0.17	0.14	0.28	0.22	0.05
QuadraticDiscriminantAnalysis	0.41	0.15	0.12	0.23	0.18	0.06
GaussianNB	0.38	0.15	0.12	0.2	0.18	0.06

Table A.17: DVPRS 11 Classes - Night -Accelerometer and Medications

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
MLPClassifier	0.69	0.41	0.5	0.39	0.47	0.51
XGBClassifier	0.7	0.42	0.48	0.39	0.47	0.7
LGBMClassifier	0.71	0.42	0.48	0.38	0.47	0.95
LinearDiscriminantAnalysis	0.66	0.39	0.44	0.41	0.46	0.12
CatBoostClassifier	0.69	0.39	0.46	0.35	0.44	5.21
BaggingClassifier	0.61	0.4	0.42	0.39	0.44	0.12
DecisionTreeClassifier	0.56	0.39	0.38	0.43	0.43	0.06
ExtraTreeClassifier	0.56	0.39	0.42	0.37	0.42	0.05
RandomForestClassifier	0.66	0.36	0.46	0.33	0.42	0.27
LogisticRegression	0.71	0.36	0.5	0.3	0.42	0.1
CalibratedClassifierCV	0.71	0.35	0.51	0.27	0.40	0.22
AdaBoostClassifier	0.62	0.36	0.52	0.27	0.40	0.2
ExtraTreesClassifier	0.67	0.35	0.44	0.29	0.40	0.2
BernoulliNB	0.64	0.34	0.41	0.3	0.39	0.06
DummyClassifier	0.5	0.35	0.51	0.26	0.38	0.05
QuadraticDiscriminantAnalysis	0.48	0.04	0.08	0.22	0.09	0.13
GaussianNB	0.49	0.03	0.04	0.03	0.04	0.06

Table A.18: DVPRS 11 Classes - Night -Accelerometer, Medications, Demographics and Clinical

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
XGBClassifier	0.62	0.4	0.46	0.37	0.44	0.76
LinearDiscriminantAnalysis	0.64	0.39	0.45	0.37	0.44	0.12
BaggingClassifier	0.61	0.4	0.49	0.34	0.44	0.13
LogisticRegression	0.63	0.4	0.52	0.32	0.44	0.11
MLPClassifier	0.53	0.39	0.44	0.36	0.42	1.54
DecisionTreeClassifier	0.56	0.38	0.39	0.39	0.42	0.06
BernoulliNB	0.65	0.37	0.43	0.33	0.42	0.06
LGBMClassifier	0.65	0.36	0.46	0.32	0.42	0.93
ExtraTreesClassifier	0.6	0.36	0.51	0.31	0.42	0.19
CatBoostClassifier	0.68	0.36	0.47	0.29	0.41	5.39
AdaBoostClassifier	0.63	0.36	0.52	0.28	0.41	0.22
RandomForestClassifier	0.61	0.35	0.49	0.28	0.40	0.26
CalibratedClassifierCV	0.64	0.35	0.51	0.26	0.39	0.28
DummyClassifier	0.5	0.35	0.51	0.26	0.38	0.05
QuadraticDiscriminantAnalysis	0.51	0.34	0.42	0.29	0.37	0.12
ExtraTreeClassifier	0.5	0.32	0.38	0.28	0.35	0.05
GaussianNB	0.47	0.02	0.04	0.02	0.03	0.06

Table A.19: DVPRS 11 Classes - Night -Accelerometer, Medications, Demographics, Clinical and Diseases

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
DecisionTreeClassifier	0.6	0.44	0.4	0.51	0.48	0.06
BaggingClassifier	0.65	0.41	0.49	0.37	0.46	0.14
BernoulliNB	0.67	0.41	0.44	0.39	0.46	0.06
LogisticRegression	0.56	0.4	0.53	0.32	0.43	0.11
QuadraticDiscriminantAnalysis	0.58	0.39	0.42	0.38	0.43	0.12
XGBClassifier	0.61	0.37	0.45	0.34	0.42	0.79
CatBoostClassifier	0.67	0.36	0.48	0.31	0.42	6.58
AdaBoostClassifier	0.63	0.36	0.52	0.28	0.41	0.22
CalibratedClassifierCV	0.57	0.36	0.52	0.29	0.40	0.28
LGBMClassifier	0.65	0.35	0.45	0.29	0.40	0.97
ExtraTreesClassifier	0.6	0.35	0.5	0.27	0.39	0.2
RandomForestClassifier	0.63	0.35	0.51	0.26	0.39	0.26
LinearDiscriminantAnalysis	0.53	0.34	0.36	0.35	0.38	0.14
DummyClassifier	0.5	0.35	0.51	0.26	0.38	0.05
ExtraTreeClassifier	0.55	0.33	0.28	0.44	0.37	0.05
MLPClassifier	0.43	0.26	0.31	0.23	0.29	1.51
GaussianNB	0.39	0.16	0.13	0.21	0.19	0.06

Table A.20: DVPRS 11 Classes - Night -Accelerometer, Medications and Demographics

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
DecisionTreeClassifier	0.6	0.44	0.45	0.44	0.47	0.06
LinearDiscriminantAnalysis	0.65	0.4	0.45	0.43	0.47	0.11
MLPClassifier	0.56	0.4	0.45	0.37	0.43	1.49
XGBClassifier	0.64	0.4	0.46	0.37	0.45	0.71
LogisticRegression	0.65	0.4	0.53	0.32	0.44	0.1
ExtraTreesClassifier	0.62	0.39	0.53	0.31	0.43	0.19
ExtraTreeClassifier	0.54	0.39	0.43	0.36	0.42	0.05
BaggingClassifier	0.67	0.38	0.46	0.35	0.44	0.13
CatBoostClassifier	0.67	0.38	0.49	0.33	0.44	5.22
LGBMClassifier	0.66	0.37	0.44	0.32	0.42	0.89
CalibratedClassifierCV	0.65	0.37	0.52	0.3	0.42	0.26
AdaBoostClassifier	0.62	0.36	0.52	0.28	0.40	0.21
BernoulliNB	0.64	0.36	0.42	0.31	0.40	0.06
RandomForestClassifier	0.62	0.35	0.5	0.27	0.39	0.26
DummyClassifier	0.5	0.35	0.51	0.26	0.38	0.05
QuadraticDiscriminantAnalysis	0.49	0.34	0.42	0.28	0.37	0.12
GaussianNB	0.49	0.01	0.03	0.01	0.02	0.06

Table A.21: DVPRS 11 Classes - Night -Clinical

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
ExtraTreesClassifier	0.52	0.37	0.47	0.39	0.43	0.16
XGBClassifier	0.54	0.36	0.47	0.37	0.42	0.61
RandomForestClassifier	0.51	0.36	0.45	0.37	0.41	0.19
CatBoostClassifier	0.54	0.37	0.44	0.34	0.41	1.46
LGBMClassifier	0.55	0.36	0.44	0.34	0.40	0.58
DecisionTreeClassifier	0.52	0.35	0.46	0.31	0.39	0.05
MLPClassifier	0.54	0.35	0.51	0.26	0.38	0.59
BaggingClassifier	0.52	0.34	0.42	0.30	0.38	0.08
LinearDiscriminantAnalysis	0.51	0.35	0.51	0.26	0.38	0.05
LogisticRegression	0.51	0.35	0.51	0.26	0.38	0.08
BernoulliNB	0.51	0.35	0.51	0.26	0.38	0.05
DummyClassifier	0.50	0.35	0.51	0.26	0.38	0.05
ExtraTreeClassifier	0.49	0.34	0.45	0.29	0.38	0.05
CalibratedClassifierCV	0.43	0.35	0.51	0.26	0.37	0.19
KNeighborsClassifier	0.52	0.33	0.38	0.29	0.36	0.06
AdaBoostClassifier	0.51	0.26	0.25	0.38	0.32	0.16
GaussianNB	0.49	0.08	0.10	0.17	0.14	0.06
QuadraticDiscriminantAnalysis	0.00	0.01	0.08	0.01	0.00	0.01

Table A.22: DVPRS 11 Classes - Night -Demographics

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
ExtraTreesClassifier	0.58	0.37	0.52	0.29	0.41	0.15
LGBMClassifier	0.50	0.36	0.52	0.31	0.40	0.55
LinearDiscriminantAnalysis	0.49	0.37	0.52	0.29	0.40	0.06
BernoulliNB	0.46	0.37	0.52	0.29	0.39	0.05
BaggingClassifier	0.42	0.36	0.52	0.31	0.39	0.07
RandomForestClassifier	0.53	0.35	0.51	0.26	0.38	0.18
DummyClassifier	0.50	0.35	0.51	0.26	0.38	0.05
CalibratedClassifierCV	0.47	0.35	0.51	0.26	0.37	0.21
AdaBoostClassifier	0.46	0.35	0.45	0.29	0.37	0.16
LogisticRegression	0.44	0.35	0.51	0.26	0.37	0.10
XGBClassifier	0.47	0.33	0.45	0.26	0.35	0.55
CatBoostClassifier	0.48	0.32	0.35	0.30	0.35	1.63
MLPClassifier	0.41	0.33	0.46	0.25	0.35	0.60
DecisionTreeClassifier	0.41	0.30	0.40	0.24	0.32	0.05
KNeighborsClassifier	0.42	0.21	0.23	0.19	0.24	0.06
ExtraTreeClassifier	0.47	0.20	0.17	0.22	0.23	0.05
GaussianNB	0.46	0.01	0.07	0.01	0.01	0.05
QuadraticDiscriminantAnalysis	0.00	0.01	0.07	0.01	0.00	0.01

Table A.23: DVPRS 11 Classes - Night -Diseases

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
CatBoostClassifier	0.51	0.37	0.45	0.31	0.40	12.16
BernoulliNB	0.56	0.35	0.39	0.33	0.39	0.05
MLPClassifier	0.43	0.38	0.50	0.31	0.39	0.59
AdaBoostClassifier	0.42	0.38	0.45	0.34	0.39	0.16
CalibratedClassifierCV	0.54	0.35	0.51	0.26	0.38	0.18
DummyClassifier	0.50	0.35	0.51	0.26	0.38	0.05
LogisticRegression	0.38	0.37	0.52	0.29	0.37	0.10
LinearDiscriminantAnalysis	0.37	0.37	0.52	0.29	0.37	0.06
XGBClassifier	0.46	0.33	0.36	0.31	0.36	0.51
DecisionTreeClassifier	0.41	0.34	0.44	0.28	0.36	0.05
BaggingClassifier	0.41	0.34	0.44	0.28	0.36	0.08
RandomForestClassifier	0.42	0.33	0.44	0.27	0.35	0.17
ExtraTreesClassifier	0.41	0.33	0.44	0.27	0.35	0.14
ExtraTreeClassifier	0.38	0.33	0.44	0.27	0.34	0.05
LGBMClassifier	0.35	0.32	0.45	0.25	0.33	0.33
GaussianNB	0.39	0.13	0.11	0.18	0.16	0.06
QuadraticDiscriminantAnalysis	0.00	0.35	0.51	0.26	0.00	0.01

Table A.24: DVPRS 11 Classes - Night -Diseases

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
XGBClassifier	0.66	0.44	0.48	0.41	0.48	0.67
RandomForestClassifier	0.64	0.44	0.47	0.42	0.48	0.25
BaggingClassifier	0.64	0.43	0.45	0.44	0.48	0.08
LGBMClassifier	0.61	0.44	0.48	0.41	0.47	0.68
DecisionTreeClassifier	0.57	0.43	0.46	0.42	0.46	0.05
CatBoostClassifier	0.66	0.40	0.46	0.37	0.45	6.34
ExtraTreesClassifier	0.68	0.40	0.46	0.36	0.45	0.18
MLPClassifier	0.68	0.39	0.50	0.33	0.44	0.50
LinearDiscriminantAnalysis	0.64	0.37	0.44	0.34	0.42	0.06
LogisticRegression	0.71	0.36	0.51	0.28	0.41	0.09
CalibratedClassifierCV	0.70	0.35	0.51	0.27	0.40	0.18
AdaBoostClassifier	0.59	0.35	0.51	0.26	0.39	0.16
BernoulliNB	0.64	0.34	0.42	0.28	0.38	0.06
DummyClassifier	0.50	0.35	0.51	0.26	0.38	0.05
CategoricalNB	0.50	0.35	0.51	0.26	0.37	0.06
ExtraTreeClassifier	0.50	0.34	0.35	0.33	0.37	0.06
GaussianNB	0.48	0.02	0.04	0.02	0.03	0.06
QuadraticDiscriminantAnalysis	0.00	0.00	0.01	0.00	0.00	0.03

Table A.25: DVPRS 11 Classes - Night -Medication

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
XGBClassifier	0.66	0.44	0.48	0.41	0.48	0.67
RandomForestClassifier	0.64	0.44	0.47	0.42	0.48	0.25
BaggingClassifier	0.64	0.43	0.45	0.44	0.48	0.08
LGBMClassifier	0.61	0.44	0.48	0.41	0.47	0.68
DecisionTreeClassifier	0.57	0.43	0.46	0.42	0.46	0.05
CatBoostClassifier	0.66	0.40	0.46	0.37	0.45	6.34
ExtraTreesClassifier	0.68	0.40	0.46	0.36	0.45	0.18
MLPClassifier	0.68	0.39	0.50	0.33	0.44	0.50
LinearDiscriminantAnalysis	0.64	0.37	0.44	0.34	0.42	0.06
LogisticRegression	0.71	0.36	0.51	0.28	0.41	0.09
CalibratedClassifierCV	0.70	0.35	0.51	0.27	0.40	0.18
AdaBoostClassifier	0.59	0.35	0.51	0.26	0.39	0.16
BernoulliNB	0.64	0.34	0.42	0.28	0.38	0.06
DummyClassifier	0.50	0.35	0.51	0.26	0.38	0.05
CategoricalNB	0.50	0.35	0.51	0.26	0.37	0.06
ExtraTreeClassifier	0.50	0.34	0.35	0.33	0.37	0.06
GaussianNB	0.48	0.02	0.04	0.02	0.03	0.06
QuadraticDiscriminantAnalysis	0.00	0.00	0.01	0.00	0.00	0.03

Table A.26: Mild vs Moderate - Day -Accelerometer

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
MLPClassifier	0.52	0.56	0.68	0.67	0.60	0.62
BaggingClassifier	0.52	0.58	0.66	0.59	0.58	0.05
GaussianNB	0.49	0.55	0.64	0.53	0.55	0.01
BernoulliNB	0.49	0.55	0.62	0.53	0.54	0.01
DecisionTreeClassifier	0.48	0.55	0.61	0.53	0.54	0.01
KNeighborsClassifier	0.49	0.53	0.64	0.50	0.53	0.01
ExtraTreeClassifier	0.48	0.55	0.60	0.53	0.53	0.01
CalibratedClassifierCV	0.50	0.54	0.67	0.45	0.53	0.04
DummyClassifier	0.50	0.54	0.67	0.45	0.53	0.01
CatBoostClassifier	0.50	0.53	0.66	0.45	0.52	1.02
LogisticRegression	0.49	0.53	0.66	0.44	0.52	0.01
AdaBoostClassifier	0.54	0.48	0.48	0.60	0.52	0.11
LinearDiscriminantAnalysis	0.49	0.53	0.65	0.44	0.52	0.01
RandomForestClassifier	0.52	0.49	0.48	0.57	0.52	0.23
ExtraTreesClassifier	0.48	0.52	0.64	0.44	0.51	0.12
LGBMClassifier	0.46	0.52	0.61	0.46	0.51	0.50
XGBClassifier	0.46	0.51	0.60	0.46	0.50	0.12
QuadraticDiscriminantAnalysis	0.52	0.33	0.40	0.62	0.44	0.03

Table A.27: Mild vs Moderate - Day -Accelerometer and Clinical

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
CatBoostClassifier	0.58	0.64	0.66	0.63	0.62	0.98
LGBMClassifier	0.60	0.62	0.62	0.64	0.62	0.05
AdaBoostClassifier	0.64	0.57	0.57	0.72	0.62	0.10
CalibratedClassifierCV	0.51	0.55	0.68	0.78	0.61	0.04
RandomForestClassifier	0.56	0.61	0.62	0.61	0.60	0.20
ExtraTreesClassifier	0.55	0.61	0.64	0.61	0.60	0.11
KNeighborsClassifier	0.52	0.59	0.64	0.58	0.58	0.02
BernoulliNB	0.52	0.58	0.65	0.58	0.58	0.01
DecisionTreeClassifier	0.53	0.58	0.59	0.58	0.57	0.01
GaussianNB	0.49	0.55	0.64	0.53	0.55	0.01
ExtraTreeClassifier	0.50	0.56	0.57	0.56	0.55	0.01
QuadraticDiscriminantAnalysis	0.55	0.52	0.51	0.60	0.54	0.03
DummyClassifier	0.50	0.54	0.67	0.45	0.53	0.01
XGBClassifier	0.47	0.54	0.58	0.53	0.53	0.09
LogisticRegression	0.50	0.53	0.66	0.45	0.52	0.01
LinearDiscriminantAnalysis	0.49	0.53	0.66	0.44	0.52	0.01
MLPClassifier	0.49	0.53	0.66	0.44	0.52	0.69
BaggingClassifier	0.47	0.52	0.52	0.53	0.51	0.05

Table A.28: Mild vs Moderate - Day -Accelerometer and Demographics

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
RandomForestClassifier	0.62	0.59	0.58	0.67	0.61	0.18
LinearDiscriminantAnalysis	0.54	0.60	0.67	0.62	0.60	0.01
LogisticRegression	0.54	0.60	0.67	0.62	0.60	0.03
XGBClassifier	0.57	0.59	0.58	0.61	0.59	0.08
LGBMClassifier	0.53	0.59	0.62	0.59	0.58	0.05
ExtraTreesClassifier	0.52	0.58	0.65	0.58	0.58	0.11
GaussianNB	0.50	0.55	0.64	0.55	0.55	0.01
MLPClassifier	0.51	0.56	0.55	0.57	0.55	0.98
CatBoostClassifier	0.48	0.54	0.62	0.52	0.54	1.06
DummyClassifier	0.50	0.54	0.67	0.45	0.53	0.01
CalibratedClassifierCV	0.50	0.54	0.67	0.45	0.53	0.04
BaggingClassifier	0.55	0.48	0.48	0.61	0.53	0.05
BernoulliNB	0.47	0.52	0.52	0.53	0.51	0.01
ExtraTreeClassifier	0.45	0.52	0.52	0.51	0.50	0.01
AdaBoostClassifier	0.52	0.42	0.44	0.58	0.48	0.11
DecisionTreeClassifier	0.51	0.43	0.44	0.57	0.48	0.01
QuadraticDiscriminantAnalysis	0.42	0.44	0.42	0.49	0.44	0.04

Table A.29: Mild vs Moderate - Day -Accelerometer and Diseases

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
XGBClassifier	0.52	0.54	0.53	0.57	0.54	0.08
CatBoostClassifier	0.48	0.54	0.62	0.52	0.54	1.08
BaggingClassifier	0.52	0.53	0.52	0.57	0.53	0.05
CalibratedClassifierCV	0.50	0.54	0.67	0.45	0.53	0.04
DummyClassifier	0.50	0.54	0.67	0.45	0.53	0.01
DecisionTreeClassifier	0.48	0.54	0.53	0.54	0.52	0.01
AdaBoostClassifier	0.53	0.46	0.46	0.60	0.51	0.11
ExtraTreesClassifier	0.44	0.51	0.56	0.48	0.50	0.11
RandomForestClassifier	0.44	0.51	0.54	0.49	0.49	0.21
MLPClassifier	0.44	0.50	0.56	0.47	0.49	0.80
GaussianNB	0.44	0.50	0.56	0.47	0.49	0.01
LogisticRegression	0.45	0.50	0.60	0.43	0.48	0.07
LinearDiscriminantAnalysis	0.44	0.50	0.59	0.43	0.48	0.01
LGBMClassifier	0.49	0.45	0.44	0.54	0.48	0.06
ExtraTreeClassifier	0.43	0.49	0.48	0.50	0.47	0.01
QuadraticDiscriminantAnalysis	0.38	0.45	0.44	0.45	0.43	0.03
BernoulliNB	0.38	0.45	0.44	0.45	0.43	0.01

Table A.30: Mild vs Moderate - Day -Accelerometer and Medications

Model	ROC AUC	F1 Score	Recall	Precision	HAR MEAN	Time Taken
RandomForestClassifier	0.64	0.66	0.66	0.68	0.66	0.20
AdaBoostClassifier	0.62	0.64	0.64	0.66	0.64	0.12
XGBClassifier	0.61	0.64	0.64	0.65	0.64	0.10
LGBMClassifier	0.60	0.64	0.64	0.65	0.63	0.06
CatBoostClassifier	0.58	0.64	0.65	0.63	0.62	1.15
BaggingClassifier	0.61	0.60	0.59	0.66	0.61	0.06
LinearDiscriminantAnalysis	0.51	0.58	0.64	0.58	0.57	0.02
BernoulliNB	0.51	0.57	0.65	0.57	0.57	0.01
ExtraTreeClassifier	0.53	0.55	0.54	0.59	0.55	0.01
ExtraTreesClassifier	0.49	0.56	0.60	0.55	0.55	0.11
DecisionTreeClassifier	0.53	0.52	0.51	0.58	0.53	0.01
DummyClassifier	0.50	0.54	0.67	0.45	0.53	0.01
CalibratedClassifierCV	0.50	0.54	0.67	0.45	0.53	0.04
MLPClassifier	0.48	0.53	0.63	0.48	0.52	0.79
LogisticRegression	0.49	0.53	0.65	0.44	0.52	0.07
QuadraticDiscriminantAnalysis	0.52	0.27	0.38	0.66	0.41	0.02
GaussianNB	0.42	0.26	0.32	0.40	0.34	0.01