CHRONIC PAIN AS A CONTINUUM: UNSUPERVISED LEARNING FOR IDENTIFICATION OF CO-EXISTING CHRONIC PAIN MECHANISMS

CHRONIC PAIN AS A CONTINUUM: AUTOENCODER AND UNSUPERVISED LEARNING METHODS FOR ARCHETYPE CLUSTERING AND IDENTIFYING CO-EXISTING CHRONIC PAIN MECHANISMS

By MD ASIF KHAN, BS

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Applied Science

McMaster University
MASTER OF APPLIED SCIENCE (2022)
Hamilton, Ontario, Canada (Electrical and Computer Engineering)

TITLE:	Chronic Pain as a Continuum: Autoencoder and Unsupervised Learning Methods for Archetype Clustering and Identifying Co-existing Chronic Pain Mechanisms
AUTHOR:	Md Asif Khan BS (Computer Science and Engineering), North South University, Dhaka, Bangladesh
SUPERVISOR:	Dr. Thomas E. Doyle
NUMBER OF PAGES:	xxiv, 226

Lay Abstract

Chronic pain (CP) is a global burden and the primary cause for patients to seek medical attention. Despite continuous efforts in this area, CP remains clinically challenging to manage. The most effective method of treating CP is identifying the underlying cause or mechanism, which is often unattainable. This thesis attempted to identify the CP mechanisms existing in a patient while quantifying them from patient-reported history and questionnaire data. Unsupervised Learning was used to identify clinically meaningful clusters that revealed the three main CP mechanisms, i.e., Nociceptive, Neuropathic, and Nociplastic, achieving acceptable hamming loss (0.43) and average precision (0.5). The results exhibited that the CP mechanisms co-exist and CP should be regarded as a continuum rather than distinct entities. The algorithm successfully indicated the dominant CP mechanism, a goal for optimal CP management and treatment. The results were also validated by a comparative analysis with data from another cohort that demonstrated a similar trend.

Abstract

Chronic pain (CP) is a personal and economic burden that affects more than 30% of the world's population. While being the leading cause of disability, it is complicated to diagnose and manage. The optimal way to treat CP is to identify the pain mechanism or the underlying cause. The substantial overlap of the pain mechanisms (i.e., Nociceptive, Neuropathic, and Nociplastic) usually makes identification unreachable in a clinical setting where finding the dominant mechanism is complicated. Additionally, many specialists regard CP classification as a spectrum or continuum.

Despite the importance, a data-driven way to identify co-existing CP mechanisms and quantification is still absent. This work successfully identified the co-existing CP mechanisms within a patient using Unsupervised Learning while quantifying them without the help of diagnosis established by the clinicians. Two different datasets from different cohorts comprised of patient-reported history and questionnaires were used in this work. Unsupervised Learning (k-prototypes) revealed notable overlaps in the data. It was further emphasized by the outcomes of the Semi-supervised Learning algorithms when the same trend was observed with some diagnosis or class information. It became evident that the CP mechanisms overlap and cannot be classified as distinct conditions. Additionally, mixed pain mechanisms do not make an individual cluster or class, and CP should be considered as a continuum.

To reduce data dimension and extract hidden features, Autoencoder was used. Using an overlapping clustering technique, the pain mechanisms were identified. The pain mechanisms were also quantified while elucidating overlaps, and the dominant CP mechanism was successfully pointed out with explainable element. The hamming loss of 0.43 and average precision of 0.5 were achieved when considered as a multi-label classification problem.

This work is a data-driven validation that there are significant overlaps in CP conditions, and CP should be considered a continuum where all CP mechanisms may co-exist.

Acknowledgements

First, I would like to sincerely thank and express my deepest gratitude to my supervisor Dr. Thomas E. Doyle, for his confidence and trust in me and for supporting me from the inception of this journey. I am grateful for his guidance and excellent supervision, from which I learned many avenues of research, developed reporting and communication skills. Thank you for allowing me to have the honor of being a part of Biomedic.AI lab at McMaster University.

I also want to express my heartfelt thanks to Dr. Dinesh Kumbhare for his guidance, support, and suggestions whenever I needed help in the clinical domain. Thank you for all the insights and utmost cooperation which have made this work possible! It has been a privilege to have the opportunity to learn from you.

I want to extend my appreciation to Dr. Ryan G. L. Koh and Dr. Samah Hassan at UHN. Thank you, Dr. Hassan, for your domain expertise that helped me understand the data better and for all the suggestions that allowed me to organize my thoughts. Ryan, I cannot thank you enough! I got you during the weekends too! Thank you for all the help and advice that helped me present this thesis better. Will miss our silly talks while working at ϟ speed! It was awesome working with you, and I'm sure you'll become the students' favorite professor soon! And thank you both for helping me with STAR-ML and the literature review.

Additionally, I would like to thank my colleagues at Biomedic.AI. Dr. Mohammadreza Heydarian, Dr. Omar Boursalie, Sajjad Rashidiani, Calvin Zhu, Mohammad Siddiqui, Ama Simons, Ryan Clark, Victoria Tucci, and Theodore Liu. Thank you for being so cooperative and fun! You all have augmented my graduate experience. Theo and Victoria, you both helped me in the development of STAR-ML and literature review when I needed speed! Thank you for being so responsive 🚀 ! Especial thanks to Sajjad, my friend! Your help throughout this time was monumental. I will miss the late-night work at ITB A202 and ☕ at Main Streets Tim Hortons!

Thanks to Ms. Laura Banfield, Health Sciences Librarian at McMaster University, for her guidance and help in the literature review. It was crucial for preparing a comprehensive but definitive search.

Outside of McMaster University and UHN, I could constantly rely on the support of my family, especially, you, Prianka ❤ ! Without their never-ending support, I could not have completed my thesis.

I would also like to thank my seniors and juniors from Bangladesh in Hamilton! I had one of the best times in my life with you all. Thank you all for your support!

Last but not least, I must acknowledge the support of the Canadian Department of National Defence IDEaS and the Department of Electrical and Computer Engineering at McMaster University, Hamilton, Ontario, Canada for funding me and providing me with the opportunity to pursue graduate studies.

Dedicated to my family

and

everyone who is suffering from chronic pain.

Table of Contents

Lay Abstract **iii**

Abstract **iv**

Acknowledgements **vi**

Notations and Abbreviations **xxii**

Declaration of Academic Achievement **xxv**

1 Introduction **1**

1.1 Motivation 2

1.1.1 Chronic Pain 2

1.1.2 Impacts of Chronic Pain 3

1.2 Machine Learning 5

1.2.1 Unsupervised Learning 6

1.2.2 Semi-supervised Learning 7

1.3 Overview of the CP Field 8

1.3.1 Challenges in Understanding CP and Treatment 9

1.3.2 Pain as a Continuum 10

1.3.3 Best Practices in CP Management 11
1.4 The Gaps 12
1.5 Objectives and Research Question 12
1.6 Evaluation Overview 14
1.7 Organization and Scope 15

2 Literature Review **17**
2.1 Theoretical Background 17
2.1.1 Data Preprocessing and Analysis 18
2.1.1.1 Data Preprocessing 18
2.1.1.2 Descriptive Statistics 19
2.1.1.3 Histogram and Barchart 19
2.1.1.4 Correlation Heatmap 20
2.1.1.5 Outlier Detection and Removal 20
2.1.1.6 Missing Data Handling 21
2.1.2 Unsupervised Learning 22
2.1.2.1 k-prototypes 22
2.1.2.2 Fuzzy C-Means Clustering 25
2.1.3 Semi-supervised Learning 27
2.1.3.1 Self-training 27
2.1.3.2 Label Propagation 28
2.1.4 Dimension Reduction and Feature Extraction 30
2.1.4.1 Uniform Manifold Approximation and Projection . . 30
2.1.4.2 Artificial Neural Network 31
2.1.5 Performance Evaluation 35

2.1.5.1 Performance Evaluation of Unsupervised Learning . 35
2.1.5.2 Performance Evaluation of Semi-supervised Learning 39
2.1.5.3 Performance Evaluation of Multi-label Classification 42
2.1.6 Explainable AI and Interpretable AI 44
2.2 Literature Review: Focus on Approach 45
2.2.1 Clustering in CP 45
2.2.2 Towards a Mechanism-based Approach in CP 52
2.3 Scoping Review: "AI in Chronic Pain" 57
2.3.1 Review type selection 58
2.4 Search Terms and Research Databases 58
2.4.1 Search Strings 59
2.4.2 Inclusion and Exclusion Criteria 60
2.4.3 Study Selection 61
2.5 Conducting the Review 63
2.6 Reporting the Review 65
2.6.1 Applications of AI/ML in CP Research 66
2.6.1.1 Data Used 67
2.6.1.2 CP Conditions 68
2.6.1.3 AI/ML Method Used 75
2.6.2 Gaps and Findings 83
2.6.2.1 State of AI in CP Research 83
2.6.2.2 Rationale of Choice 83
2.6.2.3 Future of AI in CP Research 85
2.6.2.4 Call for Open-sourced Pain Data 86

2.6.3 Limitations of the Scoping Review 87

3 Domain Data **89**

3.1 Data Description 89

3.1.1 Datasets and Study Population 89

3.1.2 Definition of Variables 90

3.1.2.1 Questionnaires in the Datasets 91

3.1.3 Study Population and Subject Selection Criteria 91

3.1.3.1 Population 91

3.1.4 Datatype 92

3.2 Data Preprocessing 92

3.2.1 Data Cleaning 93

3.2.1.1 Duplicate Removal 94

3.2.1.2 Irrelevant Data Fixing 95

3.2.2 Encoding 96

3.2.3 Recoding 96

3.2.4 Unstandardized Column Removal 96

3.2.4.1 Missing Data Handling 97

3.3 Data Analysis 100

3.3.1 Exploratory Data Analysis 102

3.3.1.1 Descriptive Statistics 102

3.3.1.2 Visualization and Analysis 107

3.3.1.3 Outlier Detection and Removal 112

3.3.1.4 Data Imputation 114

3.3.1.5 Feature Scaling 117

3.3.2 Data Visualization 117
3.4 Secondary Use of Data 119
3.5 Availability of the Datasets 119

4 Machine Learning Model **120**
4.1 k-prototypes 120
4.2 Semi-supervised Learning (SVM) 121
4.3 FCM 123
4.3.1 Feature Extraction 123
4.3.2 FCM Implementation 125

5 Methodology **126**
5.1 Identifying Distinguishable Clusters 126
5.1.1 Finding the Optimal Number of Clusters 126
5.1.2 Cluster Validation 128
5.2 Semi-supervised Learning in Identifying CP Mechanisms 128
5.3 Unsupervised Learning in Identifying CP Mechanisms 131
5.3.1 DADOSD Dataset 133
5.3.2 TRICD Dataset 134
5.3.3 Performance Evaluation 134

6 Results **136**
6.1 Unsupervised Learning 136
6.1.1 k-prototypes 138
6.2 Semi-supervised Learning 139
6.2.1 Semi-supervised SVM 139

6.2.2 Label Propagation 141
6.2.3 Label Spreading 142
6.3 Overlapping Clustering 143
6.3.1 FCM 144
6.3.1.1 Regular Clustering 144
6.3.1.2 Soft Clustering 145

7 Discussion **148**
7.1 Interpreting the Results 148
7.1.1 The Trend in Clustering CP Data 149
7.1.2 Should Mixed-Pain be Considered Separately? 150
7.1.3 Identifying Co-existing CPs 151
7.2 Why Were the Clusters Not Distinguishable? 152
7.3 Chronic Pain is a Continuum 153
7.4 Generalizability 153
7.5 Explainability 154
7.6 Impact and Novelty 155
7.7 Possible Implications 156
7.8 Limitations 156

8 Conclusion **159**
8.1 Summary of Contributions 159
8.2 Future Directions 161

A Literature Review **164**
A.1 Search Strings 164

A.2 Screening the Articles 165

B Datasets **167**

B.1 Ethical Approval and Data Retention 167

B.1.1 Research Ethics Board Approval 167

B.1.2 Storage and Use 167

B.2 Comparison Between the Datasets 168

B.3 Reasons for Feature and Instance Removal 170

B.4 Data Variables 171

C Data and Model **177**

C.1 Data Analysis 178

C.2 Correlation Heatmap 178

C.3 Model Hyper-parameters and Libraries 179

C.3.1 Visualization 180

C.3.1.1 UMAP 180

C.3.2 Unsupervised Learning (Regular Clustering) 180

C.3.2.1 k-prototypes 180

C.3.3 Semi-supervised Learning 181

C.3.3.1 SVC 181

C.3.3.2 Label Propagation 183

C.3.3.3 Label Spreading 183

C.3.4 Unsupervised Learning (AE and FCM) 183

C.3.4.1 Autoencoder 183

C.3.4.2 FCM 185

C.4 Finding Alignment with FCM Clusters 185

List of Figures

2.1 Illustration of an AE network architecture 35
2.2 MEDLINE search 59
2.3 PRISMA diagram for the scoping review 64
2.4 Focus of AI/ML applications in CP research 66
2.5 Column chart of the data sources used in CP research 67
2.6 Column chart of CP condition in focus 68
2.7 Column chart of AI/ML algorithm used in the publications 75
2.8 Line chart of the number of publications by years 83
2.9 Rationale of selecting the AI/ML algorithm 84
2.10 Spider plot of different ML-related aspects 85
2.11 Dataset availability status in the articles 87
3.1 Data cleaning steps 94
3.2 Missing values in raw DADOSD dataset 97
3.3 Missing values in raw TRICD dataset 98
3.4 Missing values in DADOSD dataset before imputation 99
3.5 Missing values in TRICD dataset before imputation 99
3.6 Cleaned data to Model-Ready data 101
3.7 Histogram and barchart before outlier removal for DADOSD dataset 109

3.8 Histogram and barchart after outlier removal for DADOSD dataset . 110
3.9 Histogram and barchart before outlier removal for TRICD dataset . . 111
3.10 Histogram and barchart after outlier removal for TRICD dataset . . . 112
3.11 Boxplot before outlier removal for DADOSD dataset 113
3.12 Boxplot before outlier removal for TRICD dataset 114
3.13 Boxplot after outlier removal and imputation in DADOSD dataset . . 115
3.14 Boxplot after outlier removal and imputation in TRICD dataset . . . 116
3.15 The datapoints in DADOSD dataset are visualized as a scatterplot using UMAP. 118
3.16 The datapoints in TRICD dataset are visualized as a scatterplot using UMAP. 118
5.1 Applying k-prototypes to the Model-Ready data 126
5.2 Optimal number of clusters in DADOSD dataset 127
5.3 Optimal number of clusters in TRICD dataset 127
5.4 Applying Semi-supervised Learning to the Model-Ready data 129
5.5 Class distribution in training and test set (DADOSD dataset) 130
5.6 AE model training history 132
5.7 Applying AE and FCM to the Model-Ready data 132
5.8 AE architecture for the DADOSD dataset 133
5.9 AE architecture for the TRICD dataset 134
6.1 Confusion matrix of Semi-supervised SVC (DADOSD Dataset) 140
6.2 Confusion matrix of Label Propagation (DADOSD Dataset) 141
6.3 Confusion matrix of Label Spreading (DADOSD Dataset) 143

6.4 The AE encoded features from DADOSD dataset visualized as a scatterplot with FCM cluster labels (regular clustering) 145
6.5 Multi-label confusion matrix of FCM using AE features 147
A.1 Scoping review: Full-text screening 166
C.1 Correlation heatmap of numerical variables of DADOSD dataset . . . 178
C.2 Correlation heatmap of numerical variables of TRICD dataset 179

List of Tables

2.1 Summary of the related works 51

2.2 Information of included studies (focused on data used) 69

2.3 Information of included studies (focused on ML/AI algorithm used) . 76

3.1 Summary of age and sex information of the datasets 92

3.2 Number of instances and features 93

3.3 Descriptive statistics for numerical variables in DADOSD dataset (after outlier removal and imputation) 102

3.4 Descriptive statistics for categorical variables in DADOSD dataset (after outlier removal and imputation) 102

3.5 Descriptive statistics for mechanistic classes in DADOSD dataset . . 105

3.6 Descriptive statistics for numerical variables in TRICD dataset (after outlier removal and imputation) 105

3.7 Descriptive statistics for categorical variables in TRICD dataset (after outlier removal and imputation) 106

4.1 Model training and test split (DADOSD dataset) 122

4.2 AE training and validation split 124

6.1 k-prototypes on DADOSD Dataset ($k = 4$) 138

6.2 k-prototypes on DADOSD Dataset ($k = 3$) 138

6.3 k-prototypes on TRICD Dataset ($k = 4$) 138
6.4 k-prototypes on TRICD Dataset ($k = 3$) 139
6.5 Report showing the classification metrics for Semi-supervised SVC . . 140
6.6 Report showing the classification metrics for Label Propagation . . . 141
6.7 Report showing the classification metrics for Label Spreading 142
6.8 FCM on AE extracted features DADOSD dataset ($k = 3$) 144
6.9 FCM on AE extracted features TRICD dataset ($k = 3$) 144
6.10 Example of FCM outputs 145
6.11 Overall multi-label classification performance 147
B.1 Comparison between TRICD and DADOSD datasets 168
B.2 DADOSD dataset variables and their description 171
B.3 TRICD dataset variables and their description 174
C.1 List of orders of the CP categories 185

Notations and Abbreviations

AE	Autoencoder
AI	Artificial Intelligence
AMI	Adjusted Mutual Information
ANN	Artificial Neural Network
ARI	Adjusted Rand Index
CBP	Chronic Back Pain
cLBP	Chronic Low Back Pain
CM	Confusion Matrix
CNS	Central Nervous System
CS	Central Sensitization
CP	Chronic Pain
CPTF	Canadian Pain Task Force
DADOSD	DAta Driven Outcome System Dataset

DBI	Davies-Bouldin Index
DL	Deep Learning
FCM	Fuzzy C-Means
FMI	Fowlkes-Mallows Index
FS	Feature Selection
ICD	International Classification of Diseases
IASP	International Association for the Study of Pain
LBP	Low Back Pain
MeSH	Medical Subject Headings
ML	Machine Learning
MLCM	Multi-Label Confusion Matrix
MSE	Mean Squared Error
NLP	Natural Language Processing
NPL	No Predicted Label
NTL	No True Label
PCA	Principal Component Analysis
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
QST	Quantitative Sensory Testing

ReLU	Rectified Linear Unit
RI	Rand Index
RNN	Recurrent Neural Network
STAR-ML	Screening Tool for Assessing Reporting of Machine Learning
SME	Subject-Matter Expert
SVC	Support Vector Classification
SVM	Support Vector Machine
TMD	Temporomandibular Disorder
TRICD	Toronto Rehabilitation Institute's Clinic Dataset
UHN	University Health Network
UMAP	Uniform Manifold Approximation and Projection
XAI	Explainable Artificial Intelligence

Declaration of Academic Achievement

The following is a declaration that the research represented in this thesis was completed by Mr. Md Asif Khan and acknowledges the contributions of Dr. Thomas E. Doyle and Dr. Dinesh Kumbhare. Md Asif Khan contributed to the inception of the study, study design, and was responsible for the data preprocessing, conducting analysis, experiments, and the writing of the manuscript. Dr. Thomas E. Doyle contributed to the inception of the study, study design, computation power needed to run the experiments, review of the manuscript, and has provided guidance and support at all stages of this thesis. Dr. Dinesh Kumbhare supported with data used in this work, study design, review of the manuscript, and has guided with clinical knowledge needed for this thesis.

This thesis contains no material that has been submitted or published previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is Md Asif Khan's work.

Chapter 1

Introduction

The abundance of healthcare data and rapid progress in analysis techniques has caused the healthcare sector to experience a paradigm shift [1]. However, this large amount of available data is often difficult for clinicians to process and help guide clinical decisions, thus, often left unutilized. Advances in techniques (e.g., Machine Learning) allow efficient information extraction from an enormous amount of data and can aid clinical decision-making. Different Machine Learning (ML) and Deep Learning (DL) techniques are currently being applied in clinical diagnoses, prognosis, epidemic outbreak prediction, drug discovery, and development [1–3].

This thesis explores the application of Unsupervised Learning and Semi-supervised Learning on chronic pain (CP) data to identify clinically meaningful clusters of patients identifying CP mechanisms. The following sections introduce the motivations for the exploration, field and method overview, including the hypothesis, evaluation overview, and thesis outline.

1.1 Motivation

This section discusses the motivation behind this work from the perspectives of CP and ML.

1.1.1 Chronic Pain

Pain is considered a major global healthcare problem and one of the most prevalent causes for patients to seek medical attention [4–6]. Global estimates suggest that one in five adults suffer from pain, and another one in 10 adults are diagnosed with CP every year [5]. More than 1.5 billion people worldwide live with CP [7]. According to the United States Pain Foundation, 50 million American adults live with CP, and it is the leading cause of long-term disability [8]. It is also the main cause of accessing the healthcare system affecting one in three Americans, and in the US alone, US$560 to US$635 billion is spent each year on CP treatments, disability payments, and loss of productivity, excluding the cost of care for military personnel, institutionalized individuals (e.g., prisoners or nursing home patients), and children [8, 9]. The Canadian Pain Task Force (CPTF) report of 2020 estimated that 7.63 million or one in four Canadians aged 15 or older live with CP, which contributed to the total direct and indirect cost of $38.3 to $40.4 billion in 2019 [10]. Although CP and its associated diseases are not immediately life-threatening, it is a leading cause of disability, and suffering [4, 5, 11].

Any pain that lasts for 3 to 6 months or more is addressed as CPs by the physicians [12, 13]. CP can be induced by injuries, surgeries, nerve damage, infections, migraines, bad posture or improper sitting and working position, sleeping on a poor mattress, ordinary aging, or it may not be associated with a physical cause [14].

CP can be categorized into three major categories [15–19]:

1. *Nociceptive* (from tissue damage)
2. *Nociplastic* (from a sensitized nervous system without evidence of tissue or nerve damage)
3. *Neuropathic* (from nerve disease or injury)

However, many specialists regard pain classification as a spectrum or continuum as there is significant overlap in the types of pain mechanisms within and between patients [15, 18, 19].

1.1.2 Impacts of Chronic Pain

CP has an immense personal and economic burden affecting 30% of the world's population. CP prevalence rates vary from 11% to more than 40% [15]. In 2016, the United States Centers for Disease Control and Prevention (CDC) estimated prevalence at 20.4% (50 million adults) [20]. A systematic review of studies done in the United Kingdom estimated the CP prevalence rate to be 43.5% (pooled). It also indicated that CP affects between one-third to half of the population (approximately 28 million adults) in the United Kingdom. This prevalence is expected to increase further with an aging population [21]. Research suggests that women report greater frequency, intensity, and duration of pain while men and women are suffering from the same painful condition. A study among the Brazilian population showed that the prevalence of CP in women was close to 50%, nearly double what was observed in men (28.36%) [22]. According to CDC, higher CP prevalence rates are found in women, individuals from lower socioeconomic backgrounds, rural areas, and military veterans [21]. The

prevalence and CP-associated disability in low-income countries are higher than in high-income countries [23].

Additionally, CP can have severe effects on a person's daily life, including his/her mental health. It becomes very difficult to lead day-to-day life for an individual with CP, and it may also lead to temporary or permanent disability [11]. Additionally, it can have critical effects on a person's daily life, including some severe consequences, including depression, anxiety, inability to work, disruption in social relationships, and suicidal thoughts [5, 10]. According to the CPTF report of 2019 [24], many Canadians lack access to appropriate pain management services, leading to poor early treatment and aggravating problems with time.

The Global Burden of Diseases, Injuries, and Risk Factors Study (GBD) estimates disability-adjusted life-years (DALYs: Years lived with disability or YLDs + Years of life lost or YLLs) due to 369 diseases and injuries for 204 countries and territories in 2019 [25]. Chronic low back pain (cLBP) is one of the 10 most important drivers for increasing the causes that had the largest absolute increases in the number of DALYs between 1990 and 2019 and is common from teenage to old age [25]. The GBD in 2013 revealed cLBP as the single most significant contributor to the YLDs globally [13].

While CP can cause depression, anxiety, inadequate sleep, and unfavorable social conditions, these issues can also induce CP conditions. Psychological factors affecting the development of CP include depression, anxiety, emotional distress, and other negative emotions. Psychological and physical trauma, catastrophizing, pain-coping, lack of family and social support, level of education, race, age, sex, and genetics are associated factors with CP that can also cause increased pain intensity, and psychological distress [26, 22]. However, psychological distress and sleep problems are

associated with pain and have a bi-directional association [15]. Moreover, CP also impacts relationships and self-esteem, which is associated with higher divorce and suicide rates [15, 27]. Furthermore, CP is associated with increased mortality [28].

However, it is debatable if CP is a pathologic entity or not [29]. Considering the importance of CP management, it was recently recognized by the World Health Organization (WHO) as a disease in its own right and implemented in revisions to the current version of the International Classification of Diseases (ICD-11) that will lead to improved classification and diagnostic coding [13, 24]. The International Pain Society and Global Health Community addressed the failure to treat pain as an abrogation of fundamental human rights, and pain management has been considered a basic human right in international law since 2004 [30].

Although pain clinics focus on diagnosing and managing CP, there is only one board-certified pain specialist for every 10,000 patients with severe CP. Patients receive only 30% pain reduction of various available treatments [8]. Therefore, introducing new efficient techniques to optimize CP treatments is crucial to offer the best possible treatments with limited resources.

1.2 Machine Learning

Artificial Intelligence (AI) or ML is gradually transforming research in the healthcare and biomedical domain and has led to man-machine collaboration in the healthcare sector [3]. AI can enhance the precision of diagnosis, enable early prognosis, and aid clinical decision-making while reducing healthcare costs [31, 2, 32, 33]. ML [34, 35] and DL [36–38] techniques have been successfully applied in clinical predictive modeling. However, DL-based methods require large and high-quality datasets to be generalizable

and learn the underlying semantics of the inputs [39, 40]. DL models can provide comparable or even better performance than medical experts in diagnosing particular diseases if quality training samples are available [41–43, 3].

ML is a technique that employs data to imitate the human learning process by automatically identifying patterns in the data. The uncovered patterns are then utilized to derive, extract, classify, cluster, or predict information for unseen or future data unveiling meaningful insights. Based on the amount of information given and the type of learning, the algorithms can be subdivided into four main groups:

1. *Supervised Learning* (data labels provided),
2. *Semi-supervised Learning* (data labels are provided for a small subset of the data, and the rest of the data are unknown or not labeled),
3. *Unsupervised Learning* (no data labels provided),
4. *Reinforcement Learning* (develops patterns based on positive and negative rewards) [44].

Another approach is DL which is a part of a broader family of ML algorithms based on Artificial Neural Networks (ANN) that is inspired by the brain. Information is processed like interconnected brain cells where the learning can be supervised, semi-supervised or unsupervised [45]. A suitable algorithm is selected or developed for application depending on the available data and the objective.

1.2.1 Unsupervised Learning

Unsupervised Learning technique, commonly known as clustering, is one of the most significant and primitive human activities that is used to inspect unrevealed insights

by unscrambling a finite dataset having little to no ground truth, into a finite and discrete set of naturally hidden data patterns or structures [46, 47]. In clustering, data samples are grouped together on the basis of some innate similarities [46]. They are often divided into two main categories based on their inherent techniques; hierarchical and partitioning [46].

Nowadays, large amounts of data are being generated, and clustering is becoming increasingly popular as it is a good way to deal with this amount of unlabeled data [47]. Clustering techniques have been used in different areas of study, such as computer science, engineering, astronomy, biology, geology, sociology, and economics [47]. Clustering techniques are evident in literature for biomedical applications [48]. In biomedical research, clustering algorithms are used ubiquitously [49], such as biomedical document mining, gene expression and genome sequence analysis, and magnetic resonance imaging (MRI) data analysis. As clustering works well to extract knowledge and insights from a large volume of data, it can be utilized in disease prevention, early diagnosis, prognosis, and treatment [47].

1.2.2 Semi-supervised Learning

In Semi-supervised Learning, a small amount of labeled data and a relatively larger amount of unlabelled data are used to perform specific learning tasks. It allows leveraging the large amount of unlabelled data available, combining the smaller set of labeled information. It combines Supervised and Unsupervised Learning in an attempt to improve performance. For example, in a classification problem, additional instances for which the label is unknown might be utilized, unlike Supervised Learning to aid in the classification process. On the other hand, for clustering approaches, the learning

technique might benefit from the true knowledge that certain instances belong to a specific group [50].

1.3 Overview of the CP Field

Usually, CP conditions are divided into subtypes based on criteria defined by clinical examinations, which consider anatomical locations and relevant symptoms overlooking etiology [51, 52]. Additionally, these empirical classification approaches are limited in scope and often disregard known pathophysiological mechanisms, leading to sub-optimal treatment outcomes [51]. These can also add cognitive bias [53] in the diagnosis affecting patient outcomes. Pain categorization can influence prognosis, diagnosis, and treatment with the provision of services implications [15]. Thus, there is a need for an unbiased and strategic approach to identify the exact mechanisms driving the pain phenotype for improving CP treatment and management [54].

The CP categories stated in Subsection 1.1.1 are elaborated here.

Nociceptive pain Nociceptive pain is the most common form of CP. It is caused by activity in neural pathways and is secondary to stimuli that might cause tissue damage [55].

Neuropathic pain According to the International Association for the Study of Pain (IASP), Neuropathic pain is a term used for a group of conditions caused by a lesion or disease affecting the somatosensory nervous system [56]. About 15-25% of people with CP are estimated to have Neuropathic pain [57].

Nociplastic pain Nociplastic pain can be described as pain arising from the altered function of pain-related sensory pathways in the central and peripheral nervous system without any evidence of tissue damage or discrete pathology concerning the somatosensory nervous system. The mechanisms causing Nociplastic pain are not fully understood, but it is considered that augmented central nervous system (CNS) pain and sensory processing, and abnormal pain modulation play major roles. It can emerge in isolation or as a comorbidity with CP conditions that are primarily Neuropathic or Nociceptive.

There are overlapping conditions characterized by Nociplastic pain with other pain conditions [58, 59]. Usual descriptors for Nociceptive pain include terms such as aching and throbbing, whereas Neuropathic pain is commonly described as lancinating and shooting [15].

1.3.1 Challenges in Understanding CP and Treatment

Identifying the underlying cause is considered the most effective mode of management and treatment for CP [53]. CP management is challenging to manage clinically, even with the increase in potential treatments and individualized pain therapy along with mechanism-based approaches [54]. Mechanism-based pain treatment is optimal, but identifying the mechanisms behind the pain in clinical practice can be very difficult or unattainable [15, 17]. Therefore, treatments are generally symptom-based or disease-based [15]. The anatomic or radiographic diagnoses are not sufficient to guide rehabilitative care without considering the underlying pain mechanism(s). Evaluation of pain mechanisms can help personalize care and is a step forward in providing

precision medicine to pain patients. Pain mechanism evaluation is guided by patient-reported history, questionnaires, and potentially quantitative sensory testing (QST), which is a psychophysical test. Though QST is being used to infer the presence of pain mechanisms, it lacks norms to aid in interpreting findings and established test metric standards [60, 61, 17].

A patient's description of pain should be accepted as true in the absence of other contradictory evidence, as pain is mostly subjective [15]. This is especially true when pain is complicated to convey, as seen in patients with Nociplastic pain. Sometimes there are associated subjective symptoms, and missing biomarkers make it even more difficult to diagnose. As a result, physicians are often unable to reach high certainty about the diagnosis, and consequently, patients are aggrieved that their symptoms are doubted [58, 62]. Physicians might rely on other means like facial expressions or imaging to assess pain and to identify causes [15]. However, pain diagnosis heavily relies on or is based on clinical judgment. Therefore, it relies heavily on the experience, skills, and available resources for assessment [56, 18].

1.3.2 Pain as a Continuum

The concept of mixed pain is increasingly being acknowledged suggesting that many pain conditions have a blended pain phenotype and do not fall into one particular category, i.e., clinically substantial overlap of Nociceptive and Neuropathic symptoms. For patients with CP in primary care and orthopedic settings, the prevalence of pain with mixed pathophysiology was estimated to be 59.3% [63, 64, 17–19, 65].

There is evidence of overlap of Nociceptive, Neuropathic, and Nociplastic pain or mixed pain, implying these pains are part of a CP continuum. For example, a patient's

chronic low back pain may be a mixed pain consisting of Nociceptive, Neuropathic, and Nociplastic components [58, 17–19, 65].

1.3.3 Best Practices in CP Management

Proper CP management and treatment can reverse functional and structural brain abnormalities [66]. Tailored therapy for the patients to improve their quality of life should be the goal of the treatment. Pain treatment guidelines have recommended a personalized approach that uses a shared-decision model, as pain is a dynamic consequence of biological, psychological, and social factors [15, 67, 68]. The US Department of Health and Human Services 2019 report on Pain Management Best Practices suggested a multidisciplinary approach should be taken for CP across different disciplines using one or more treatment modalities. The report emphasized an individualized, patient-centered approach, and the care should be based on the biopsychosocial model [69]. A multimodal approach should be employed, which includes self-care, a healthy lifestyle including exercise, proper nutrition, maintained sleep hygiene, smoking cessation, and ergonomic changes if needed. This may also include other treatments like psychological therapies, opioid and non-opioid pharmacological therapies, and other complementary treatments [15, 67, 69].

In theory, mechanism-based pain treatment is considered superior to symptom or disease-based treatments as it focuses on the underlying mechanism of the pain. Personalized multimodal and interdisciplinary treatment approaches that might include psychotherapy, pharmacotherapy, integrative treatments, and invasive procedures are often advised by clinical trials and guidelines. With adequate pain management, neuroplastic changes might also be reversible. Additionally, emotional support systems

and well-being can boost healing and diminish pain chronification [15].

1.4 The Gaps

Several studies have been administered in this field where only a few studies tried to identify pain archetypes using clustering techniques [70–73, 51, 52, 54, 74]. Additionally, most of the works only focus on a particular pain disorder or a small set of pain disorders. Although the concept of mixed pain is recognized by clinicians and researchers, and the prevalence of co-existing CP conditions is very high [15, 18, 19], there is no study that leverages ML to identify mixed pain in patients. Additionally, there is no study that tried to employ clustering techniques to identify clinically meaningful clusters to reveal CP mechanisms and indicate the dominant mechanism. Moreover, "CP as a continuum" rather than being distinct clinical conditions is yet to be validated in a data-driven way.

1.5 Objectives and Research Question

As previously stated, biological pain mechanisms can be categorized into three classes or categories, i.e., Nociceptive, Nociplastic, and Neuropathic [15, 60, 17–19]. However, the identification of the mechanisms is not directly measurable but instead is inferred from indirect assessments, which makes it more challenging to discern [60, 17].

Identification of pain mechanisms can improve personalized treatment by helping both clinicians and physical therapists. The efficiency of an intervention can be maximized if multiple pain mechanisms can be addressed simultaneously [19]. This allows for prioritizing specific treatments based on the mechanisms reasoned to be

involved rather than a general diagnosis. After identifying the pain mechanisms, the next step is to provide treatment(s) targeting the present mechanisms. Unfortunately, there is currently no tool to discern the relative roles of the pain mechanisms [60, 75, 19]. Additionally, it is feasible to use questionnaires to distinguish the pain mechanisms [17]. Therefore, there is a need for an unbiased and strategic approach that can utilize patient-reported history, and questionnaire data [10] to identify existing pain mechanisms while quantifying them.

ML is yet to be explored to identify CP mechanisms. This thesis tries to answer the following question:

> "Can ML (i.e., unsupervised or semi-supervised) identify existing pain mechanisms (i.e., Nociceptive, Neuropathic, and Nociplastic) in a CP patient without the help of diagnosis and treatment information?"

To answer this question, Unsupervised Learning was applied to the patient-reported questionnaire data to identify pain mechanisms without any diagnosis or treatment information. Therefore, the objective of this thesis will be to develop a technique that can extract clinically meaningful information from CP patient data in order to provide clinicians with better information about the pain mechanisms to allow for better pain management and treatment.

This thesis also tries to validate the notion of pain as continuum in a data-driven way. The data should make separable clusters if there is no overlap in the data. If the clustering cannot reveal separable clusters, then Semi-supervised Learning will be employed to see if a small amount of class information (mechanistic classes) can help the algorithm to identify mechanisms. The failure of Semi-supervised Learning will suggest that the data has overlaps indicating co-existing CP mechanisms are

significant. In this case, if an overlapping clustering technique exhibits overlapping clusters, then it indicates CPs are not distinct conditions but a spectrum or continuum.

If pain can be clustered in different groups or categories (mechanistic classification) other than the classical anatomical approach, then that could lead to optimized treatment and management of CP, including the development of personalized/tailored rehabilitation programs [19, 65]. As a result, the expected outcomes can be maximized with the available resources. In addition to that, it has the potential to increase effectiveness and satisfaction among patients, which is an essential factor in pain management. This may also improve AI algorithms across other pain groups/cohorts resulting in better resource allocation for hospitals, pain clinics, and rehabilitation centres.

1.6 Evaluation Overview

AI has emerged as a great asset in medical applications, and its potential in medical sectors is rising. Its abilities are now far beyond merely assisting doctors in providing simple diagnoses [1].

This work investigates CP patients' data using Unsupervised Learning to reveal clinically meaningful groupings and tries to explore the clusters in the light of mechanistic CP classes. Then semi-supervised approaches were examined where the algorithm benefits from true class information. Finally, a DL approach was tried to extract features from data, and overlapping clustering was applied to identify CP mechanisms and validate the notion of pain as a continuum. These methods are tested and validated using suitable metrics.

As only patient-reported history and questionnaire data are taken into account,

it is free from cognitive bias that might come from doctors' diagnoses. If similar groups can be found, it could be beneficial to understanding the medical condition better or from a new viewpoint. If the clusters are clinically meaningful, i.e., represent mechanistic classes of CP, it would aid doctors in faster decision making leading to faster treatment and improved pain management. The dependency on a diagnostic test gets minimized as only the questionnaire is involved [10, 68]. It can also help better manage or treat the medical condition/s, which is necessary given the minimum resources. In addition, it would be interesting to observe the generalizability and consistency of the model by applying it to another set of similar data. Domain experts' help was sought to validate the result of the algorithms in the clinical context.

1.7 Organization and Scope

The thesis has 8 chapters and is structured as follows: Chapter 2 presents the literature review in view of theoretical background, relevant works, and a scoping review with gap analysis on the relevant research.

Chapter 3 gives a description of the data used in this thesis with analysis with visualizations.

Chapter 4 gives an overview of the use of the ML algorithms used in this work.

Chapter 5 describes the model implementations and evaluations, whereas Chapter 6 presents the results.

Chapter 7 discusses the experimental validation, interprets the results, and expands upon the potential of the interpreted results along with the limitations.

Finally, Chapter 8 presents conclusions and discusses directions for future work.

The author brings to the reader's attention that part of this thesis was published

in 2022 [76]. This publication is made by the author of this thesis, as the lead author, in collaboration with his supervisor at McMaster University and co-authors at UHN. The scoping review of the related literature and identifying the gaps in the areas of CP with ML application in Chapter 2 and the following chapters are the contributions that have only been published in this thesis.

Chapter 2

Literature Review

This chapter is divided into three main parts. The first part presents the theoretical backgrounds relevant to this thesis (Section 2.1). The second part presents a literature review focused on the approach taken for this study (Section 2.2), and the last part is a scoping review on the current state of AI in the field of CP following the guidelines for scoping review [77]. The scoping literature review titled "AI in Chronic Pain" tries to summarize the current trends in the use of AI in CP (diagnosis, prognosis, clinical decision support, self-management, and rehabilitation) during the last 10 years (2012-2022) while identifying gaps. The scoping review is summarized in four stages: planning, screening protocol, conducting, and reporting the review. The scoping review is presented in Section 2.3.

2.1 Theoretical Background

This section gives a theoretical overview of the used steps and techniques in this thesis.

2.1.1 Data Preprocessing and Analysis

2.1.1.1 Data Preprocessing

Data preprocessing is a vital step in improving data quality and preparing data for analysis. Unfortunately, real-world data are influenced by negative factors such the inconsistent and superfluous data, the presence of noise or outliers, errors, discrepancies in codes or names, and duplicates. Low-quality data leads to low-quality ML/DL models. Data preprocessing includes data cleaning such as handling the removal of noise and inconsistent data), missing data handling, data transformation into forms that are appropriate for the ML models, feature selection and extraction, etc. [78].

Feature Scaling Independent features present in the data are standardized or normalized in a range using feature scaling techniques. With few exceptions, ML algorithms don't perform well if numerical features have different scales. Feature scaling is especially important for classifiers that calculate distances between data, such as Support Vector Machines, k-nearest neighbors (k-NN). Without scaling, features with larger numerical values have a more significant effect on the distance and dominate other features when calculating distances.

There are two common ways to get all numerical features to have the same scale: min-max normalization and standardization.

1. Min-max normalization: This technique re-scales feature values ranging from 0 to 1.

2. Standardization: It re-scales the feature values so that the resulting distribution has a unit variance distribution with 0 mean value.

Standardization is less affected by outliers than normalization. However, unlike min-max normalization, standardization does not confine values in 0 to 1 range which may create issues for some algorithms (e.g., ANNs often expect input values ranging from 0 to 1) [44].

2.1.1.2 Descriptive Statistics

Descriptive statistics acts as an initial descriptor of the data and provides simple summaries about the sample and the observations made. It is a summary statistic that summarizes and describes features quantitatively from a dataset. It consists of methods for organizing, visualizing, and describing data using tables, graphs, and summary measures [79]. Here the central tendency (mean, median, or mode) of datasets is reported. Three measures of dispersion or spread, i.e., range, standard deviation, and interquartile range, have been reported as well [80].

2.1.1.3 Histogram and Barchart

A histogram gives an approximate representation of the frequency distribution of numerical features or data. It is constructed using 'bin' with a range and putting the values inside the range or intervals to make columns. The bins are adjacent without any gap and are mostly of equal size [81]. In order to indicate the original variable is continuous, the rectangles of a histogram touch each other [82].

While a histogram is used for continuous data, a bar chart plots a graphical comparison of categorical variables. It is recommended that bar charts should have gaps between the rectangles to clarify the distinction [82, 83].

2.1.1.4 Correlation Heatmap

Correlation is the measure of how two features are correlated with each other. The standard correlation coefficient or Pearson correlation coefficient is calculated and plotted as a heatmap [84]. Pearson correlation coefficient is the ratio between the covariance of two variables and the product of their standard deviations. For this, it always has a value between -1 and 1 where the negative value indicates a negative correlation and the positive value indicates positive correlation [85].

2.1.1.5 Outlier Detection and Removal

The data points that differ significantly from others within a given dataset are considered outliers [86, 87]. Outlier detection is a fundamental issue in data mining and ML [88]. Identifying outliers and eliminating them is vital for building stable ML models. It is a standard practice in ML problems as it helps to make better assumptions about the data to uncover the underlying pattern of the machine learning algorithms [89, 88].

The anomalies or outliers may result from typos, i.e., misplaced decimal points, transmission errors, or during exceptional cases or circumstances, e.g., health data from a patient with a rare disease that can add irregularities in the measured data. Sometimes, only a few outliers can distort the group results by altering the mean and variance or by increasing the standard deviation of data. Studies have shown that data analysis considerably depends on how outliers or missing values are handled [89].

This work uses a standard and widely used descriptive statistical method called interquartile range (IQR) to identify any outliers from the density distribution of the features. It is based on the mean and variance of each group of data and is defined as

the difference between the 25th and 75th percentiles of the data. To calculate IQR, the feature is divided into quartiles via linear interpolation. Plotting boxplots are particularly helpful in visualizing the distribution of normal and outlier data points [90, 89, 88].

Boxplot A boxplot or box-and-whisker plot shows quantitative data distribution to facilitate comparisons between variables or across levels of a categorical variable. It shows the quartiles of the variable while the whiskers show the rest of the distribution while pointing out the 'outliers' using an inter-quartile range method [91].

2.1.1.6 Missing Data Handling

Missing data not only leads to significantly inconsistent results by the ML algorithms but also to the scientific soundness of the study being compromised. It is also important to know why the values are missing. In many cases, all the samples with missing values are deleted, which causes information loss. Adding reasonable estimates of missing data is better than removing the sample or leaving it untreated [92, 78].

To resolve this issue, imputation techniques are used. The goal of imputation is to fill the missing points of the variables with intuitive data. Imputing missing values with central tendency measures of features such as mean, median, or mode are simple and widely used [92]. For a feature or column, the mean is the average of all values in the column, where the median is the middle number (sorted by size), and mode is the most frequent numerical value. However, imputing missing data with mean or median values can only be done with numerical features, whereas mode imputation can be done with numerical and categorical features. One issue with these techniques, they ignore relationships with other variables [92]. However, it is not recommended to use

the mean imputation if the distribution of the feature is skewed or contains outliers [93].

2.1.2 Unsupervised Learning

Clustering approaches are usually divided into a few types, including hierarchical, centroid-based, distribution-based, density-based, and self-organizing maps. However, clustering techniques are highly dependent on the types of data in use [94]. The clustering algorithms in the scope of this thesis are discussed below:

2.1.2.1 k-prototypes

k-prototypes is a partitioning-based clustering method. It is an improvement from the k-means and k-modes clustering techniques to handle clustering with mixed data type [95]. Like k-means, it measures the distance between numerical features using Euclidean distance, but it measures the distance between categorical features using the number of matching categories as well. The k-prototypes algorithm is more useful than k-means and k-modes as real-world data mostly hold mixed-type objects [95]. The features with categorical variables need to be filtered carefully for the implementation of the algorithm. Besides that, the quality of the input data may affect cluster initialization.

The k-prototypes algorithm defines k virtual points or prototypes as the centers of the groups or clusters. For the numerical variables, these prototypes are represented by mean values, and for categorical variables, the prototypes are represented by mode values. The dissimilarity between two mixed-type objects A and B can be measured by d using equation 2.1.1. A and B are expressed by n number of

features $X_1^r, X_2^r, ..., X_p^r, X_{p+1}^c, ..., X_n^c$ where first p features are numeric and the rest are categorical $(n-p)$.

$$d(A,B) = \sum_{j=1}^{p}(a_i - b_i)^2 + \gamma \sum_{i=p+1}^{n} \delta(a_i, b_i) \tag{2.1.1}$$

Here, the first term denotes the squared Euclidean distance (is useful for comparing distances) measurement for the continuous variables, and the second term calculates the simple matching dissimilarity measure on categorical features [96]. This measure is defined by

$$d_s(A,B) = \sum_{i=1}^{m} \delta(a_i, b_i) \tag{2.1.2}$$

where

$$\delta(a_i, b_i) = \begin{cases} 0 & (a_i = b_i) \\ 1 & (a_i \neq b_i) \end{cases} \tag{2.1.3}$$

To avoid favoring any type of feature, i.e., numerical and categorical, the weight γ is used. This can either be user-specified or estimated by a combined variance of the data describing the influence of the categorical versus numerical features. Hence, the distance measure is a linear combination of the Euclidean measure and the simple matching coefficient.

The goal is to minimize the objective function, which is the total sum of distances d_T between the instances and the prototype of the belonging class u_l:

$$d_T = \sum_{l=1}^{k} \sum_{a \in C_l} \left(\sum_{i=1}^{p}(a_i - u_{\mathrm{l,i}})^2 + \gamma \sum_{i=p+1}^{n} \delta(a_i, u_{\mathrm{l,i}}) \right) \tag{2.1.4}$$

The initial k prototypes are selected as temporary centers of the clusters, then each instance is assigned to the closest prototypes. After allocating all the instances, the prototypes are updated to represent their optimal clusters. After that, the instances are reallocated to the updated prototypes if necessary, and the process is repeated till the partitions become stable [97].

The workflow of the algorithm can be described in a few steps [98]:

1. Randomly initializes k cluster prototypes (one for each cluster).
2. For each instance/data point in the dataset:
    2.1 Allocate each instance to the cluster whose prototype is closest according to equation 2.1.1.
    2.2 Update cluster prototypes of the corresponding cluster after each allocation to be the new center of the data points in the cluster.
3. If no data points are left to be assigned to a cluster:
    3.1 Recalculate the similarity of all instances against the current prototypes.
    3.2 If a data point is closer to another prototype than the clusters it belongs to, reassign the data point to that cluster and update the prototypes of both clusters.
4. Repeat steps 2-3 until no data points have changed clusters in step 3.2 or the maximum number of iterations set by the user has been reached.

2.1.2.2 Fuzzy C-Means Clustering

Lotfi A. Zadeh introduced fuzzy sets as an extension of the classical notion of set where the set elements have degrees of membership [99]. Fuzzy C-Means is an extension of the k-means algorithm [100], with the concept of fuzzy sets. Unlike k-means, clusters are not considered mutually exclusive partitions, but flexible sets that can overlap some of the other clusters. All the instances are assigned to all the clusters, but a weight vector determines the membership level of belonging to the clusters. Partially overlapped properties can be described by adjacent clusters where a given instance can have a non-zero weight for multiple clusters. The degree of membership is determined by the magnitude of the weight.

In clustering techniques like k-prototypes where a data point can only belong or assigned to one cluster are referred to as hard or crisp clustering. This restriction is relaxed for fuzzy clustering, and an instance can belong to all of the clusters with a certain degree of membership [101, 94]. If the clusters are overlapping and ambiguous, this is very useful. Additionally, the memberships may help uncover more sophisticated relations between a given data point and the disclosed clusters [94].

The fuzzy c-means or FCM [102] is one of the most used fuzzy clustering algorithms [103]. FCM attempts to find c fuzzy clusters for a set of data points $a_j \in \Re^d, j = 1, ..., N$ while minimizing the cost function

$$J(U, M) = \sum_{i=1}^{c} \sum_{j=1}^{N} (u_{i,j})^m D_{ij} \tag{2.1.5}$$

where $U = [u_{i,j}]_{c \times N}$ is the fuzzy partition matrix and $u_{i,j} \in [0, 1]$ is the membership coefficient of the jth object in the ith cluster; $M = [m_1, m_2, ..., m_c]$ is the cluster prototype (mean or center) matrix; $m \in [1, \infty)$ is the fuzzification parameter and

usually is set to 2 [129]; $D_{ij} = D(x_j, m_i)$ is the distance measure between x_j and m_i.

We summarize the standard FCM as follows, in which the Euclidean or L_2 norm distance function is used.

1. Select appropriate values for m, c, and a small positive number ε. Initialize the prototype matrix M randomly. Set step variable $t = 0$.

2. Calculate (at $t = 0$) or update (at $t \geq 1$) the membership matrix U by

$$u_{ij}^{t+1} = \left(\sum_{l=1}^{c} \left(\frac{D_{lj}}{D_{ij}} \right)^{(1-m)^{-1}} \right)^{-1} \tag{2.1.6}$$

   for $i = 1, 2, ..., c$ and $j = 1, 2, ..., N$.

3. Update the prototype matrix by

$$m_i^{t+1} = \left(\sum_{j=1}^{N} \left(u_{ij}^{t+1} \right)^m x_j \right) \Big/ \left(\sum_{j=1}^{N} \left(u_{ij}^{t+1} \right)^m \right) \tag{2.1.7}$$

   for for $i = 1, 2, ..., c$.

4. Repeat steps 2–3 until $\|M^{t+1} - M^t\| < \varepsilon$.

FCM suffers to produce good results in presence of outliers. Additionally, if the initial cluster centers are not optimal it may also result in sub-optimal clusters. Algorithm k-means++ can use for optimized center initialization [94, 104]. It is developed by augmenting k-means with a randomized seeding technique to obtain a faster algorithm ($O(\log k)$) compared to k-means along with the optimal clustering. In k-means++, the first center is randomly chosen from input data with uniform

distribution. Then, the probability P of being center is calculated for each data point:

$$P_i = \frac{D(x_i)}{\sum_{j=0}^{N} D(x_j)} \tag{2.1.8}$$

where $D(x_i)$ is distance from point i to the closest center. Based on these probabilities, the next center is chosen. This step is repeated till the required amount of centers is initialized. k-means++ can be used for optimal center initialization for the FCM algorithm [105].

2.1.3 Semi-supervised Learning

2.1.3.1 Self-training

Semi-supervised Learning has some popular models and one of them is self-training [106]. Self-training performs Semi-supervised Learning using the model itself as a pseudo labeler. It reinforces the understanding of the model by iterating through the samples.

Self-training inspired by Yarowsky [107, 106], can help a supervised classifier to function as a semi-supervised classifier and allow it to learn from unlabeled data [108]. One of the main advantages of self-training is it is simple and not ambiguous. Additionally, it is a wrapper method as the choice of learner method is left open for the user [106].

Support Vector Machines Support Vector Machines (SVMs) are a set of popular ML methods for classification, regression, and other learning tasks. SVM classifier or Support Vector Classification (SVC) [109] is a much-prevailing classification algorithm defined by a separating hyperplane. It produces support vectors and tries to maximize

the Euclidean distance (margin) between the data points and the decision boundary. It is a non-probabilistic classifier and it performs well in case of a small number of samples [110, 109, 111]. SVMs are effective in high dimensional spaces, the number of samples is smaller than the number of dimensions, and different kernel functions can be used for the decision function, e.g, polynomial kernel. However, if the number of samples is significantly smaller than the number of features, it is important to avoid over-fitting by choosing proper kernel functions and regularization term [112].

As SVMs can only solve binary classification problems, in the case of multi-class classification, it utilizes the one-versus-one scheme by fitting all binary sub-classifiers. Then it assigns a class to samples by a voting mechanism [113].

Support Vector Machine (SVM) is a well-established classification algorithm, defined by a separating hyperplane. It produces support vectors and tries to maximize the Euclidean distance (margin) between the data points and the decision boundary. It is a non-probabilistic classifier and has the hinge loss function which measures the number of misclassified data examples. In this work, Support Vector Classification (SVC) with polynomial kernel was used. It belongs to the SVM family and used for two-class and multi-class problems. For a given training vector x, the polynomial kernel is given by $(\gamma \langle x, x' \rangle + r)^d$ where γ defines the influence of a single training example, r is the correlation coefficient, and d denotes the degree [109, 112].

2.1.3.2 Label Propagation

Label Propagation is a semi-supervised graph inference algorithm that is used for classification tasks. It works by constructing a similarity graph over all instances in the input dataset.

Zhu and Ghahramani [114] proposed a simple iterative label propagation algorithm to learn from both labeled and unlabeled data. Labels are propagated with a combination of random walk and clamping. Similar to other similar Semi-supervised Learning algorithms, label propagation works as intended if labeled data reveals the structure of the data distribution to fit the classification goal. It is known by Label Propagation [115].

Zhou et al. [116] proposed another label propagation-based algorithm to learn from labeled and unlabeled data. Often referred to as Label Spreading, is a Semi-supervised Learning approach to design a classifying function, sufficiently smooth with respect to the inherent structure revealed by labeled and unlabeled data points from the entire dataset [117].

Both of these two algorithms work by constructing a similarity graph over all items in the input dataset. But, they differ in modifications to the similarity matrix that graph and the clamping effect on the label distributions. Clamping allows the algorithm to change the weight of the true labeled data. Label Propagation algorithm performs hard clamping (clamping factor $\alpha = 0$) of input labels. This can be relaxed, e.g., $\alpha = 0.3$, 70% of the original label distribution will be retained where the algorithm changes its confidence of the distribution within 30%.

Label Propagation employs the raw similarity matrix constructed from the input data, whereas, Label Spreading minimizes a loss function with regularization properties making it more robust to noise. Label Spreading utilizes affinity matrix based on the normalized graph Laplacian and soft clamping across the labels [118].

Label Propagation computes a similarity matrix between samples with a k-nearest neighbor (k-NN) kernel to propagate samples and produce a sparse matrix with

significantly reduced running times. It can be expressed by $1[x' \in kNN(x))]$ where k is number of neighbours.

2.1.4 Dimension Reduction and Feature Extraction

Dimension reduction is a fundamental technique for visualizing high dimensional data and pre-processing it for ML algorithms [119]. In high-dimension, data can have highly correlated and redundant features, and all dimensions do not necessarily contain an equal amount of useful information. The goal of dimension reduction is to retain the most possible information about the data while representing the data with fewer features in lower dimensional latent space or embedding space.

Feature selection (FS) and feature extraction (FE) are two major aspects of ML. Feature selection keeps only the relevant or distinguishing features, while feature extraction uses some transformations to generate novel and useful features from the original ones. Both play significant roles to the effectiveness of clustering algorithms [120, 94].

The right FS method or a combination of FS method/s and ML algorithm/s including clustering can significantly improve the efficacy, reduce overfitting, and enhance the computational efficiency of the overall system [121, 122]. FE can not only help in dimensionality reduction but also help in data visualization. Overall, these methods can decrease the computation time and simplify the ML process [94].

2.1.4.1 Uniform Manifold Approximation and Projection

Uniform Manifold Approximation and Projection (UMAP) is a manifold learning technique for data visualization and non-linear dimension reduction for ML. It can

preserve more of the global structure and faster than commonly used t-Distributed Stochastic Neighbor Embedding (t-SNE). UMAP also ensures that the local structure in the data is preserved in balance with global structure [119].

However, like t-SNE, the relative size of clusters and distances between the clusters in lower dimension might not be meaningful due to using local distances when constructing its high-dimensional graph representation [123].

2.1.4.2 Artificial Neural Network

There are several different types of neural networks, including convolutional neural networks (CNN), and RNN [124]. A generalized workflow of an ANN is described here. A typical ANN consists of different layers [124]:

1. input layer
2. one or more hidden layers
3. output layer

The nodes or neurons of each layer usually represent the number of features. Similar to the human brain, the nodes are mapped through links called "synapses" to the nodes of the hidden layers and finally to the output layer. These links are associated with weights representing feature strength that help to decide which feature should be passed to the subsequent layers. An ANN can adjust the weights by learning through an optimization process, i.e., Adam is a computationally efficient optimization algorithm which is a stochastic gradient descent method that is based on adaptive estimation of moments [125].

Activation functions such as sigmoid, tangent hyperbolic (tanh), Rectified Linear Unit (ReLU), and leaky version of a ReLU (LeakyReLU) which allows a small gradient when the unit is not active [85]. The nodes of the hidden layers utilize activation functions on the weighted sum of inputs and then map them to the outputs holding the predicted values. When the weights get adjusted, the output layer constructs a vector of probabilities for the outputs, and the one with the minimum error rate is chosen.

The training of an ANN is an iterative task where it tries to minimize the error, e.g., mean squared error (MSE) which computes the mean of squares of errors between labels and predictions to adjust the weights. In order to learn sparse features or internal representations while reducing overfitting, regularizers are used. It also improves the model's generalizability to new observations. For example, there are layer weight regularizers that allow to apply penalties on layer activity or parameters during optimization. These penalties are summed into the loss function e.g., $L2$ regularizer (the sum of the squared values) [126]. The errors are backpropagated into the network from the output layer in order to find the most optimal values for errors, and the weights are adjusted accordingly. This is repeated several times while the weights are re-adjusted until there is an improvement in the predicted values or the cost. This can be done efficiently by 'Early Stopping', which stops training when a monitored metric has stopped improving [127]. When the cost function is minimized, the model is trained.

Dimension Reduction for Clustering Most clustering algorithms are susceptible to the data dimensions due to unreliable similarity metrics and suffer from the 'curse of dimensionality' [94, 128, 129]. The term 'curse of dimensionality' denotes the

complexity growth for multivariate function estimation under a high dimensional space [130].

Especially, distance-based clustering algorithms may not be effective in a high dimensional space as there is no difference between the distance of the nearest points from other points when the dimension is high enough [94, 131]. Transforming the data to a lower dimension is vital for clustering to make the high-dimensional data manageable for the algorithms while being computationally less expensive. Unfortunately, dimension reduction incurs information loss, may distort real clusters and impairs interpretability [94, 128].

However, high-dimensional data usually have an inherent dimension considerably lower than the original [94]. Considering the data have some low-dimensional latent representation, autoencoder can be used to get such representations consisting of fewer features compared to the high-dimensional feature space [129].

Autoencoder An autoencoder (AE) is a type of ANN that is designed to encode the input into a compressed and meaningful representation and then reconstruct it back by decoding in a way that the reconstructed input is as similar as possible to the original input [129]. AE can be used as a generative model, data denoising, and dimensionality reduction. While Principal Component Analysis (PCA) is a linear projection of data points into a lower dimensional space, non-linear methods, such as AE, can often achieve superior results. The encoded latent representation can be used as features for classification, and clustering techniques as well [124, 132, 128, 129].

The goal of AE is to capture the most important features present in the data. AE is composed of two parts:

Encoder The encoder part of the network transforms the inputs into a latent-space

representation while keeping the most important features. It can be defined as $h = f(x)$ where f is a function that takes x as input and transforms or maps x into the latent space representation h.

Decoder The decoder part uses the latent representation of the inputs with the aim of reconstructing the inputs. It can be presented as $x' = g(h)$ where g is a function that takes h as input to construct x' with the objective of making x' as similar as possible to x.

The AE (encoder and decoder) can therefore be described by the function $g(f(x)) = x'$. AE's objective is not just copying the input into the output but generating the h such that it holds the important properties of the dataset that can be utilized for further analysis. To be able to extract only important features from the given data, a set of constraints can be set on the function that generates h in order to achieve the resulting form with smaller dimensions than x. As a result, the quality of representing the most important features relies on the constraints defined on h. When h allows a good reconstruction of x then it has retained most of the information present in the input [124].

There are several variations of AEs. An illustration of an AE network architecture is shown in Figure 2.1. However, a multi-layer under-complete AE is often used for feature extraction to be utilized by classification or clustering algorithms [128, 129]. This type of AE has more than one hidden layer and has a smaller dimensional h compared to the input layer (x). It is particularly helpful as additional internal hidden layers can extract the hidden features better and smaller dimensional h force the AE to capture the most salient features of the input data [128].

To use the low-dimensional latent feature representation as features for clustering

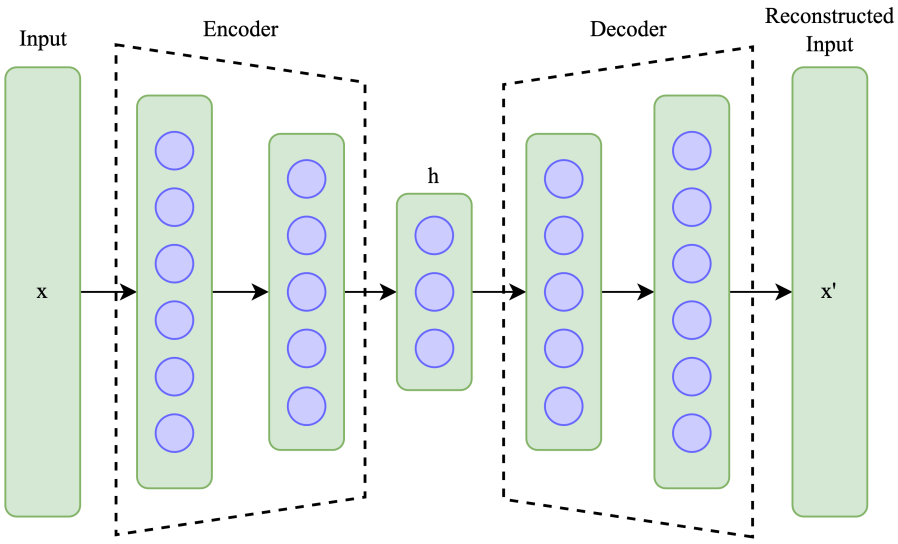


Figure 2.1: Illustration of an AE network architecture. It has an encoder on the left and a decoder on the right. Here h is the latent representation or the bottleneck.

or classification algorithms, the AE is trained similarly to ANN. Then, the decoder is put aside, and the output from the encoder is used as the features for the clustering or classification algorithms [124, 132, 128, 129].

2.1.5 Performance Evaluation

2.1.5.1 Performance Evaluation of Unsupervised Learning

Performance evaluation of a clustering algorithm is more complex than counting the number of errors or the precision and recall of a supervised classification algorithm. In particular, the absolute values of the cluster labels should not be considered by an evaluation metric. The goal should be to check if the clustering defines separations of similar data points to the ground truth set of classes (if available) or satisfies assumptions according to the defined similarity metric/s, such as members of a class being more similar than members of different classes. The evaluation metric used for

clustering is presented here, separated into two types based on the availability of the true class information.

Rand Index and Adjusted Rand Index Rand index (RI) is a similarity measure between two clustering's assignments [133]. A variation of RI is adjusted Rand index (ARI) [134] corrects the RI for agreements due to chance. However, the RI remained a popular clustering validity index as it has a simple and natural interpretation [135].

The knowledge of ground truth or true labels is necessary to compute these indexes. Similar clusterings have a high RI and ARI, where 1.0 is the perfect score. The score ranges are $[0, 1]$ for the RI and $[-1, 1]$ for the ARI. For RI, 0 indicates that the two data clusterings do not agree. For ARI, a lower score means poor agreements like RI, but a negative score the agreement is less than expected from a random result [136, 137].

Adjusted Mutual Information Adjusted Mutual Information (AMI) originated from information theory which measures the agreement of the two clusterings. It is normalized against chance, similar to ARI [137, 138].

Ground truth or true labels are needed to compute AMI. The AMI is in $[-1, 1]$ where bad or independent labeling results in negative scores, random label assignments score close to 0, and 1 indicates equal or perfect assignments [136]. Unlike ARI, when the ground truth clustering has unbalanced clusters, including small clusters, AMI provides a better comparison with a high AMI score representing pure clusters [139].

Homogeneity and Completeness Rosenberg [140] defined two objectives, i.e., homogeneity and completeness, which are two objectives of any cluster assignment.

Homogeneity is satisfied when all clusters contain instances that are members of a single class, and completeness is satisfied when all the instances that are members of a given class are members of the same cluster.

Homogeneity and completeness also require knowledge of the ground truth classes. Both can range from 0.0 to 1.0. A higher value indicates a better clustering result. Increasing the homogeneity of a clustering assignment often results in decreasing its completeness [140, 136].

Fowlkes-Mallows Index The Fowlkes-Mallows index (FMI) is expressed as the geometric mean of the pairwise precision and recall. Though it was introduced as a measure for comparing hierarchical clustering, it can be used for other clustering methods.

To be able to calculate FMI, the true class assignments of the samples should be known. FMI score can range from 0 to 1, where values close to 0 indicate largely independent label assignments and values close to 1 indicate significant agreement. However, this measure has an issue in the case of small numbers of clusters where the value is very high, even where the clusterings are independent [140, 137, 136].

Silhouette Coefficient Score Silhouette coefficient score helps to interpret and validate the consistency within clusters [141]. It is a measure of how similar an instance is to its own cluster in comparison to other clusters. The silhouette coefficient is defined for every sample in the data based on two scores:

1. Mean distance between a sample and all other samples in the same cluster.
2. Mean distance between a sample and all other samples in the next nearest cluster.

When the ground truths are not known, this measure helps to evaluate the clustering model's performance by applying it to the results of a cluster analysis. For a set of samples, the silhouette coefficient is given as the average of the silhouette coefficients for each sample [136, 141].

It ranges from $[-1, 1]$ where a higher score indicates better-defined and dense clusters. A negative score refers to incorrect clustering, and scores close to 0 indicate overlapping clusters [136].

Davies-Bouldin Index Davies–Bouldin index (DBI) is a metric for evaluating clustering algorithms defined as the average similarity measure between clusters. Here the similarity measure is the ratio of within-cluster distances compared to between-cluster distances. The validation of how well the clustering performed is made using quantities and features of the dataset by computing only point-wise distances [142].

True class information is not needed to calculate DBI. It is applied to the results of a cluster model where 0 is the lowest possible score. Scores closer to 0 indicate better partitioning referring clusters are further apart [136].

Finding Optimal Number of Clusters Finding the optimal number of clusters is a necessary step for an Unsupervised Learning or clustering algorithm. Two methods in the scope of this thesis are discussed below.

Elbow Method The elbow method considers the percentage of explained variation or dispersion as a function of the number of clusters. If plotted against the number of clusters, the variation changes rapidly at the beginning for a small number of clusters. At some point, the change slows down (less variance), leading to an angle

or elbow formation in the graph. The optimal number of clusters is selected at this point [143, 144].

This method usually uses the sum of squared errors (SSE) or within the sum of squared errors (WSSE) as a performance metric, traverses the number of clusters (k values), and finds the elbow point [143]. The inflection point should be evident for the k value to be determined. For k-prototypes to find the optimal number of clusters, the cost function (sum distance of all points to their respective cluster centroids) that combines the calculation for numerical and categorical variables is used instead of SSE or WSSE [145, 146].

Average Silhouette Score Another criterion for estimating the natural number of clusters is the average silhouette score of the data [141]. To determine the optimal number of clusters, the average silhouette coefficient is calculated for all the samples for different numbers of clusters. As the silhouette score close to 1 indicates appropriate clustering, the number of clusters that yields the largest average silhouette score is considered as the optimal number of clusters [147].

2.1.5.2 Performance Evaluation of Semi-supervised Learning

Parameter Tuning The parameters which are not directly learned during training by the classifiers are called hyper-parameters. It is often passed as arguments to the constructor of the estimator or classifier classes, e.g., C, *kernel*, and *gamma* for SVC [148].

It is recommended to search for the best set of hyper-parameters that gives the best cross-validation score. Grid search is a commonly used way that exhaustively considers all parameter combinations until a good combination of hyper-parameter

values is found. Given a set of values for the hyper-parameters, it evaluates all the possible combinations of hyper-parameter values using cross-validation [44, 149].

Accuracy and Balanced Accuracy Accuracy is the percentage of labels that a classifier successfully predicted. It is defined by:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{2.1.9}$$

Here, TP represents the number of true positives, FP is the number of false positives, FN is the number of false negatives, and TN is the number of true negatives. Therefore, accuracy is the ratio of the sum of TPs and TNs out of all the predictions [150].

Accuracy is not a good measure if the dataset is imbalanced. Balanced accuracy is used to deal with imbalanced datasets, both binary and multi-class classifications. It is the average of recall obtained on each class [151].

Precision Precision is a measure of the classifier of not labeling a negative sample as a positive prediction. It is defined by the following equation:

$$Precision = \frac{TP}{TP + FP} \tag{2.1.10}$$

where TP represents the number of true positives and FP is the number of false positives.

Recall Recall or Sensitivity measures the classifier's ability to find all the positive samples. It is given by:

$$Recall = \frac{TP}{TP + FN} \tag{2.1.11}$$

where TP represents the number of true positives and FN is the number of false negatives.

F1-Score F1-Score combines precision and recall and is described as the harmonic mean of them. It is a single metric that weights the precision and recall in a balanced way such that a higher value is required for both metrics to increase the value of the F1-Score. It is defined by:

$$F1\text{-}Score = \frac{Precision \times Recall}{Precision + Recall} \tag{2.1.12}$$

For multi-class classification, F1-Score is calculated for each class in a one-vs-rest approach as opposed to an overall F1-Score (binary classification). Additionally, unlike binary classification, there are no positive or negative classes for multi-class classification problems. The TP, TN, FP, and FN are considered for each class separately to calculate the measures mentioned above.

Support Support depicts the number of occurrences of each class in ground truth or true labels.

Weight Weight considers both TP and FN counts for each class which is ignored in the regular one-vs-rest confusion matrix [152, 153].

The following aggregated metrics are calculated for the aforementioned metrics (Precision, Recall, F1-Score) in the case of multiclass/multilabel classification tasks:

Micro Metrics are calculated globally by counting the total TPs, FNs, and FPs.

Macro Metrics are calculated for each label along with their unweighted mean.

Weighted Metrics are calculated for each label that returns the average weighted by support. Label imbalances are taken into account in this case which may result in an F1-Score that is not between precision and recall.

Samples Metrics are calculated for each instance and return their average. It is only meaningful for multi-label classification where this differs from the accuracy score.

Confusion Matrix A confusion matrix (CM) is a two-dimensional matrix that allows visualization of a classifier's performance by representing the true labels in the rows and predicted labels by the classifier in the columns [154].

For multi-class classification tasks where each instance can belong to a single class, the confusion matrix is an essential tool for performance evaluation [152].

2.1.5.3 Performance Evaluation of Multi-label Classification

Hamming Loss Traditional accuracy measures cannot evaluate the performance of a multi-class multi-label classifier as a misclassification is no longer entirely wrong or right. If a subset of the actual classes is predicted correctly, then it should be considered better than a prediction that contains no actual class [155].

Hamming loss is one of the most used metrics to evaluate performance for multi-label classifiers [156]. It is the fraction of incorrectly predicted labels. Hamming loss considers the prediction error and the missing error (a relevant label not predicted by the classifier) and normalizes over the total number of classes and the total number of instances [157]. In other words, hamming loss only penalizes individual incorrect predictions.

Hamming loss value ranges from 0 to 1 where a lesser value indicates a better classifier [157].

Multi-Label Confusion Matrix In a multi-label classification task, each instance can have multiple labels belonging to multiple classes. Multi-Label Confusion Matrix (MLCM) is a variation of regular CM that precisely calculates performance measures such as precision, recall, and F1-score for multi-label problems by extracting the accurate TP, TN, FP, and FN information of each class without ignoring the combination of true and predicted labels together. It also handles the possibility of not predicting any labels for some or all of the true labels and adds an extra column to the CM as 'No Predicted Label (NPL)' case. Additionally, it also tackles the possibility of not assigning any label by the classifier for some instances, which is added as an additional row to the CM as 'No True Label (NTL)' [152].

The summation of each row is used for calculating the weighted average of the performance measures. Unlike 'Support', 'Weight' in MLCM considers both TP and FN counts for each class which is ignored in the regular one-vs-rest confusion matrix [152, 153].

2.1.6 Explainable AI and Interpretable AI

Often the results of the ML or AI-based solution cannot be understood or communicated properly. Especially if the application is related to medical systems, it is a serious issue. Therefore, Explainable AI (XAI) [158] for transparency in the decision-making process by the AI model will be introduced. It is also to inject trust into the system with proper reasoning while making it communicable and easier to understand. There are methods that might be introduced to enhance model explainability by addressing why a subject has been assigned to a particular cluster or a certain pain archetype. The relevance between the hypothesized method and the currently used approach may also be assessed.

SHAP (Shapley Additive Explanations) was introduced by Lundberg and Lee [159] and is a famous method for explaining individual predictions based on the optimal Shapley values in game theory. The predictor values of a data sample act as players in a coalition where the Shapley value is the average of marginal contribution of the predictor values across all possible coalitions. Submodular Pick-Local Interpretable Model-Agnostic Explanations or SP-LIME is another method that provides a global view of the model to users by selecting a set of representative instances with explanations to address the trusting issue of the models [160].

The explanations by XAI are necessary for algorithmic fairness and identifying potential bias in the training data and address why the algorithm made the decision it took, but it does not describe how it arrived at that decision [160]. However, interpretability aims to describe the internals of a system in a way human understands. A system is considered interpretable if it produces simple descriptions using meaningful vocabulary to make the user understood [161]. Interpretable AI clarifies how it arrived

at the decision whereas the answer of 'why' comes from XAI.

Local Interpretable Model-Agnostic Explanations (LIME) is an explanation technique that helps to explain the predictions of any classifier or regressor in an interpretable way by approximating the predictions locally with an interpretable model [160].

2.2 Literature Review: Focus on Approach

2.2.1 Clustering in CP

Clustering is common in pain research [71]. In [72], E. Bäckryd et al. indicate that clustering clinically important subgroups and the comparison of their responses to any interventions is a significant area of research. In terms of pain-related data, it is difficult to know the true cluster structure, and this is necessary to evaluate the correctness of the clustering methods [71]. Lötsch, J. and Malkusch, S. proposed an approach using Explainable AI (XAI) methods to interpret cluster structures related to pain [71]. A few supervised ML methods were used to interpret the clusters by considering the meaning of cluster identification as a classification problem.

Pain can be stated as a subjective experience controlled by psychosocial and contextual factors [72]. Anxiety disorder was found to be associated with 17-35.1% of the cohorts [162, 163], and almost 35% of the CP patients have reported depression [164]. However, it was observed in [70], not depression or anxiety, sleep is the most important factor in CP.

Bruehl et al. [165] discussed approaches for evaluating validity and reliability related to diagnostic criteria of CP and spotted the light on the challenges regarding the

validation criteria when pathophysiologic mechanisms cannot be identified. However, [51] suggests that the multidimensional diagnostic criteria of CP addressed in [166] are encouraging to classify patients based on psychosocial factors, comorbidities, and functional consequences along with core diagnostic features. It also suggests that the phenotypic cluster profile might become more beneficial than the anatomical diagnoses by considering fundamental pain processing mechanisms and psychological distress [51, 54].

T. Miettinen et al. [70] used a data-driven approach in 320 patients with persistent pain to find the most contributing factors related to different pain phenotypes. The dataset used in this work contains 59 predictors that were grouped into 7 different categories, including pain phenotype-related factors, demographic factors, pain etiology, comorbidities, lifestyle factors, psychological variables, and treatment-related factors. They considered comorbidities and lifestyle factors for their analyses that were missing in the previous studies. The data was analyzed using a few simple unsupervised and supervised machine learning methods. Surprisingly, not depression but they found sleep problems to be the most prevalent factor associated with the extreme phenotypes of pain. Though this work recognizes an important factor for the assignment of pain phenotypes but their proposed algorithm did not perform very well in terms of the non-pain phenotype variables, and the dataset is also considerably small. Additionally, clustering analysis is vulnerable to outliers in the data, which was not considered.

In [72], the authors analyzed a dataset of 4665 patients using principal component analysis, hierarchical clustering analysis, and partial least squares-discriminant analysis to subgroup CP patients for helping the development of tailored rehabilitation programs. They proposed four subgroups using psychometric data and also identified

the attributes most responsible for the subgroup discernment to better understand the reason behind the discrimination. This work also addresses if there was any association between the subgroups and ICD-10 diagnoses for consistency. Unlike [70], outliers were identified and discarded. However, the data used in this study used inquiry form as a replacement for systematic clinical examination of anxiety and depression, which could be misleading. Additionally, this study design lacks the directions of causality.

The authors of [73] clustered low back pain patients into four clusters based on their individual course of low back pain (LBP) over time with distinctly different clinical courses and validated their results against clinical variables and outcomes. This observational study considered 6-month clinical course of 176 patients with non-specific LBP, along with measurements of bothersomeness collected from weekly text messages from them. They used four derived parameters from the actual set of 26 parameters in the dataset for cluster analysis using hierarchical clustering. K-means was used along with Ward's method to optimize the cluster allocation and the resultant clusters were described according to the initial level of bothersomeness, rate of early improvement, and the point of change. It is interesting how the data were collected using text messages. However, they also proposed a 7-cluster solution, but they considered the 4-cluster solution to make the solution more describable and generalized.

Vardeh et al. in [54] also used chronic low back pain (cLBP) to indicate the current knowledge of pathophysiological pain mechanisms in clinical practice and used this as a driver of potential treatment choice. This work is an extension of the 5 dimensions of pain mechanism that need to be considered during pain diagnostic classification suggested by the Analgesic, Anesthetic, and Addiction Clinical Trial Translations, Innovations, Opportunities, and Networks (ACTTION)- American Pain Society (APS)

Pain Taxonomy (AAPT). A 3-stage pain diagnostic ladder is proposed to specify the treatment choice by narrowing down the pain mechanisms which is a complex approach to put into clinical practice due to the absence of precise diagnostic tools for mechanisms and lack of specific analgesic treatments for particular mechanisms.

In [74], the authors used UKBioBank data of 4,156 chronic back pain (CBP) and 14,927 pain-free controls and employed FCM clustering to derive CBP sub-groups. The variables consisted of psychosocial, brain, and physical factors. From 1502 variables, 100 variables were selected after t-test. From the 100, according to feature weighting tests, only 10 variables were considered for subsequent analysis. However, only two dimensions (loneliness and depressive symptoms) were used for the clustering analysis, which indicated five optimal clusters based on cluster validity measures, i.e., silhouette value, Calinski-Harabasz, and Davies-Bouldin index. The cluster labels were then used as class labels by SVM, Naïve Bayes, k-NN, and Random Forest classifiers to determine classification accuracy. The best classification accuracy achieved was about 95% when only CBP sub-groups were assessed, but when the healthy controls were added, the misclassification in CBP sub-groups increased to 35-53% across the classifiers. The authors indicated that there were overlaps in CP patients' data, and the sub-grouping accuracy might improve with the inclusion of pain processing mechanisms. Though the authors claimed that this was the first study to develop and classify sub-groups of CBP based on psychosocial, physical, and nervous system measures, they did not use any physical or nervous system features for clustering.

The authors of [51, 52] developed an algorithm, namely Rapid OPPERA Algorithm, to identify groups of individuals based on biopsychosocial risk factors for a limited set of CP conditions. Study participants responded about bothersomeness by particular

symptoms and the feature set of 4 variables was derived to address the generalizability and reliability of the cluster model. The feature set to assign an individual to a cluster were muscle pain sensitivity, somatic symptoms, anxiety, and depression. They extended their work in [51] by using the clustering algorithm on two additional cohorts to check the generalizability and stability of the clusters. This work considered cohorts with temporomandibular disorder (TMD) and other CP conditions and is agnostic to the anatomical location of the pain. But it lacks validation in terms of diverse type of CPs and the cohort mainly consist of female and white patients. Additionally, it did not address comorbidity or disability associated with the pain disorders. It indicates the need to consider further tactics like mixed models to use large datasets where the missingness in cluster features can be observed. In [51, 167], the authors found pain amplification and psychosocial distress as contributing factors to CP. In [52], the authors identified three subgroups by supervised cluster analysis clusters using a comprehensive array of biopsychosocial measures. The cluster membership was found to be highly associated with chronic TMD. The study of [51], suggests that the distinguishing feature of the three clusters is consistent among cohorts.

Though the number of potential treatment targets has increased and mechanism-based along with individualized pain therapy approach is introduced, CP management is still very challenging to manage clinically [54]. Usually, CP conditions are classified or divided into subtypes by criteria-defined clinical examinations based on their anatomical location and relevant symptoms overlooking etiology, which might impede optimal treatments [51, 52]. Additionally, these empirical classification approaches are limited in scope and often disregard known pathophysiological mechanisms which consequently leads to sub-optimal treatment outcomes [51]. Thus, there is a need for

a strategic approach to identify the exact mechanisms driving the pain phenotype to improve CP management [54].

A summary of relevant works, along with their approaches, is listed in Table 2.1. Several studies [70–73, 51, 52, 54, 74] have been administered in this field, where only a few studies tried to identify pain archetypes using clustering techniques. Additionally, most of the works only focused on a particular pain disorder or a small set of pain disorders. There is a lack of addressing all the pain mechanisms despite being the most effective way to treat CP. A crucial detail of associating the explainability and interpretability of the used models in those studies is still rare. Moreover, the use of Principal Component Analysis (PCA) is very common when it comes to clustering pain [70, 72, 71], which hampers the explainability and the interpretability of the model. Therefore, there is a necessity for studies considering explainability while identifying pain mechanisms using clustering techniques to minimize bias.

Table 2.1: Summary of the related works

Reference	**Work Done**	**Approaches**
[70]	Identification and Interpretation of the pain phenotype cluster structure, most contributing factors related to different pain phenotypes	PCA, Ward's k-means
[71]	Cluster structures' interpretation in pain-related phenotype	PCA, k-means, ABC Analysis, Tree-based methods (XAI)
[72]	Sub-grouped CP patients into 4 groups	PCA, hierarchical clustering analysis, partial least squares-discriminant analysis
[73]	Clustered in 4 groups based on low back pain	Ward's Method, k-means
[74]	Found 5 clusters in CBP patients with loneliness and depressive symptoms then checked classification performance using cluster labels	FCM, SVM, Naïve Bayes, k-NN, Random Forest
[51, 52]	Identified groups of individuals with pain conditions and used the clustering algorithm on two additional cohorts to check the generalizability	Rapid OPPERA Algorithm
[54]	Introduced a 3-stage pain diagnostic ladder to specify the treatment choice of chronic low back pain	Mechanism-based approach to diagnose cLBP

2.2.2 Towards a Mechanism-based Approach in CP

There is a lack of gold standard in differentiating the underlying pain mechanisms, i.e., Nociceptive, Neuropathic, and Nociplastic. Due to the absence of a gold standard, the verification of discriminating features for the mechanisms in the clinical setting depends on expert consensus [19].

In 2010, K.M. Smart and co-authors derived expert consensus lists of clinical indicators of nociceptive, peripheral neuropathic, and central sensitization (CS) mechanisms of musculoskeletal pain for clinicians' use for mechanistic classification [168]. Through a web-based 3-round Delphi survey method, 103 clinical experts were surveyed to set the criteria for mechanistic inferences. Later, in a series of three articles [169–171], the authors tried binary classification of the mechanisms for LBP patients using the set criteria. In their work, they tried to identify feature sets contributing to each pain mechanism, but the mixed pain or indeterminate pain state participants were excluded from the studies.

In a study conducted in 2015, J. Nijs et al. [172], reviewed original research articles and conducted a Delphi study to set classification criteria for peripheral Neuropathic, Nociceptive, and CS in the LBP population. This also requires validation as the conclusion is not based on LBP patients.

M. C. Kolski [173] aimed to validate the clinical application of a pain mechanism classification system (PMCS) in clinical practice. Documented signs and symptoms for musculoskeletal pain data from medical records were analyzed to find the agreement between the PMCS determined by physical therapists and the category assigned by statistical model using patients' signs and symptoms. Unlike the other studies, five pain mechanism categories were considered, such as inflammatory, ischemia, peripheral

neurogenic, central, and other (multiple pain mechanisms). With unweighted pair-group method with arithmetic mean (UPGMA) and k-means cluster analysis, the assumed five groups were classified with sensitivity ranging from 15.7% to 83.1%. The sensitivity was calculated as the number of patients assigned to a category by the statistical model divided by the number of patients assigned to the category by physical therapists. While the study provided empirical support that a PMCS could be implemented in an outpatient pain clinical practice, it underscored the need for further research on the mixed pain mechanisms (75.3% of patients were labeled as other by the physical therapists).

In 2021, E. Kosek et al. tried to set clinical grading criteria and presented an algorithm for diagnosing possible or probable Nociplastic pain [18]. As Nociplastic pain can also co-occur with Neuropathic and especially Nociceptive pain, there remains an open question of when a patient with nociceptive pain should be classified as also having Nociplastic pain. Though Quantitative sensory testing, offset analgesia, and functional neuroimaging are helpful in Nociplastic pain, they are not readily accessible. So, the IASP recognized the need for clinical criteria for Nociplastic pain. The authors used a consensus on 3 questions answered by 55 experts and leaders to define a set of clinical and research-applicable criteria for Nociplastic pain.

Four criteria were set to be applied to reach a clinical diagnosis of possible or probable Nociplastic pain where a possible diagnosis requires the presence of two criteria (1-2), and all four should be present for a probable diagnosis. The criteria are as follows:

1. Pain is present for 3 months, regional, and cannot be explained entirely by Neuropathic or Nociceptive pain.

2. Clinical elicited findings of evoked pain hypersensitivity in the region of pain.

3. History of pain sensitivity (i.e., to touch, pressure, movement, or temperature) in the region.

4. Presence of comorbidities, i.e., hypersensitivity to sound/light/odor or all, sleep disturbance, fatigue, or cognitive problems.

These criteria were set by consensus, which associated bias, and the authors highlighted that these proposed criteria require further testing for validity. It is also to be noted that, to identify Nociplastic pain, differentiation from Neuropathic and Nociceptive mechanisms is also required. As these also depend on clinical judgments, this becomes subjective.

M. A. Shraim et al. conducted a review of methods to discriminate the mechanism-based categories of pain in the musculoskeletal system [17]. Using a framework developed in [65] to cluster literature in terms of 3 CP mechanisms, 188 articles were clustered into five themes that contributed to discrimination between mechanisms: clinical examination, QST, pain-type questionnaires, diagnostic and laboratory testing, and imaging [17]. These themes were discussed in 93%, 44%, 42%, 35%, and 31% of the articles, respectively. However, these numbers do not indicate that a highly used method is superior to others. The authors presented a summary of commonly used methods to discriminate between pain mechanism categories. A combination of features and methods was recommended to discriminate the CP mechanisms. Several impedances were identified in the application of methods to discriminate between CP mechanisms, such as lack of consensus on clear definitions and criteria for Nociceptive and Nociplastic pain, claim to discriminate between specific pairs of pain mechanisms, not all three or a mixed representation, poor validity, and reliability of the used

methods. A need for an expert consensus method to guide the recommendation of the combination of methods to aid the identification of CP mechanisms in clinical and research grounds was raised until further advances in the identification of the mechanisms.

As identified in [17], M. A. Shraim et al. ran a Delphi expert consensus study and generated features to discriminate between mechanism-based categories of musculoskeletal pain [19]. From the 196 features that reached consensus, 120 features were shared between pairs of pain mechanism categories; from the 76 features, 17, 37, and 22 were unique to Nociceptive, Neuropathic, and Nociplastic pain mechanisms, respectively.

Unique features achieving the most significant consensus for the mechanisms are as follows:

1. Nociceptive: Responsiveness to nonsteroidal anti-inflammatory drugs (NSAIDs), inflammation signs, and predictable pain recovery established on expected time of tissue recovery.

2. Neuropathic: Nerve damage-related features, e.g., diagnostic tests documenting nerve damage.

3. Nociplastic: Diffuse, widespread, or poorly localized pain, generalized hypersensitivity, and multiple somatic symptoms.

The pain mechanisms have significant overlaps in terms of features. The co-existence of multiple mechanisms within an individual is likely to explain the overlaps. Thus, discriminating between the mechanisms depends on a combination of features. The authors underscore the need for the development of a tool or tools to identify

pain mechanism categories in pain patients, where the aim should be to identify which mechanism contributes most to an individual's current presentation rather than focusing on identifying only one.

The studies involving Delphi had conflicting views from the experts due to various reasons such as experience, understanding of pain, personal bias, and language barrier. These need further validation with computational techniques. Moreover, the number of features or criteria to be considered to diagnose pain mechanisms are also challenging to assess for practitioners, if attainable.

2.3 Scoping Review: "AI in Chronic Pain"

Choosing the right type of review plays a key role in knowledge synthesis and accelerating the research with evidence. In a study [174], 14 most common types of literature reviews, especially in the health information domain, are listed. As the target is to identify knowledge gaps and to assess the scope of use of AI in the context of CP, scoping review is the suitable one in its own right [175].

The Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement is a widely used tool by authors to improve the reporting of systematic reviews while maintaining transparency in every step of conducting the review [176]. The PRISMA 2009 statement gets replaced by the PRISMA 2020 statement and includes a new reporting guideline that reflects advances in methods to identify, select, appraise, and synthesize articles [177].

In 2018, an extension of PRISMA was published for scoping reviews. The checklist consists of 20 essential reporting items and two optional items to be included while conducting and reporting a scoping review. Critical appraisal to include individual sources of evidence was also done in this review accordingly [77].

While there are many tools to assess the methodological quality of studies [178–181], there was a lack of a tool to assess the quality of ML algorithms' reporting in studies. To consistently assess the reporting quality and only include well-reported articles in the review paper, a new tool (Screening Tool for Assessing Reporting of Machine Learning; STAR-ML) was developed by the author that can be used to screen articles for a systematic or scoping review focusing on the reporting of the ML algorithm [76].

PRISMA 2020 flow diagram [177] template has been followed for the review, while STAR-ML was used as a component of the "Reports assessed for eligibility" step to

include only well-reported research in the review.

2.3.1 Review type selection

The research question was "What is known from the existing literature about the use and focus of ML/AI in the domain of Chronic Pain?". Having some degree of ambiguity in the question could gain benefits by capturing varying degrees of work, the current state of ML/AI's application in CP research, and its limitations. An excellent balance of keywords should be maintained in the search to reduce the chances of missing relevant articles while generating an efficient and manageable number of references. Therefore, a comprehensive approach in order to generate relevant but wider coverage was taken, which is recommended by Arksey et al. [182].

According to the goal, scoping review was chosen, and the guideline established by Arksey et al. was followed, which is highly adapted for scoping reviews [182].

2.4 Search Terms and Research Databases

The search strategy was developed in Ovid MEDLINE and then adapted to the other databases. The comprehensive search string was developed and finalized after multiple reviews and iterations with the librarian (Ms. Leeanne Romane), Subject-matter experts or SMEs (Dr. Samah Hassan and Dr. Kumbhare), and Dr. Doyle. Searches were carried out across four electronic databases: MEDLINE, Web of Science Core Collection, ACM Digital Library, and IEEE Xplore from 2012 to 2022. These four databases were chosen to have a good combination of medical and engineering databases. The final search was conducted on February 28, 2022. Keywords and

Medical Subject Headings (MeSH) regarding CP and AI were used and are given in section 2.4.1.

2.4.1 Search Strings

The search string used to search in MEDLINE is given below. The search strings for the other databases were adapted from the MEDLINE search.

MEDLINE Figure 2.2 shows the search developed in Ovid MEDLINE.

1. artificial intelligence/ or exp machine learning/ or natural language processing/ or neural networks, computer/ or cluster analysis/
2. artificial* intelligen*.ti,ab,kf,kw.
3. machine learning.ti,ab,kf,kw.
4. (deep learning or convolutional neural network or artificial neural network).ti,ab,kf,kw.
5. (cluster analysis or (unsupervised adj2 learning)).ti,ab,kf,kw.
6. natural language processing.ti,ab,kf,kw.
7. computer neural network*.ti,ab,kf,kw.
8. or/1-7
9. Chronic Pain/
10. ((Chronic* or Recurrent or Persistent*) adj3 pain*).ti,ab,kf,kw.
11. or/9-10
12. 8 and 11
13. limit 12 to english language
14. 13 not (animals/ not (humans/ and animals/))
15. limit 14 to journal article
16. limit 15 to ("review articles" or meta analysis or "systematic review" or comment or editorial)
17. 15 not 16
18. exp Neoplasms/
19. 17 not 18
20. remove duplicates from 19
21. limit 20 to yr="2012 -Current"

Figure 2.2: MEDLINE search

Search strings for the other databases can be found in Appendix A.1.

2.4.2 Inclusion and Exclusion Criteria

This review included all studies that used AI techniques in the CP domain. Separate inclusion and exclusion criteria for first and second level screening were set before conducting the review [183]. Articles were included if they satisfied the following criteria:

1. Studies published in English
2. Studies involved only human participants/data
3. Original research article, i.e., not a review article or letter
4. Peer-reviewed
5. Studies focused on CP
6. Used AI or ML methods
7. Studies focused on physically adults (17+)
8. Studies excluding only healthy participants or synthetic data
9. Studies scored 6 or more in the STAR-ML
10. Studies cannot be duplicate

The detailed 'Title and Abstract screening' and 'Full-text screening' guidelines can be found in Appendix A.

2.4.3 Study Selection

The entire process, from duplicate removal to the final study selection, was organized and conducted using Covidence systematic review software (Veritas Health Innovation, Australia). Covidence is a web application that helps to collaborate and streamlines the production of systematic and other literature reviews [184].

Duplicate Removal There were overlaps among the studies across the selected databases. Among 691 total studies from the four databases, 207 were found to be duplicates and excluded by Covidence.

Title and Abstract Screening During the first phase, all the articles were screened independently by two researchers based on the title and abstract and were voted as Yes, No, or Maybe. The discrepancies (voted as Maybe) were resolved by group discussions and consensus. Articles meeting the inclusion criteria underwent a full-text review by two researchers to determine final eligibility and confirm the accuracy of the extracted data. Articles that passed the first screening were transferred to the full-text review section in Covidence for full-text review.

Full-text Screening Based on the exclusion criteria, four independent reviewers screened the full texts and excluded the articles.

Assessment of Reporting Quality With the other exclusion criteria, the STAR-ML tool (version 2) was utilized to consistently score and include only the articles which were well-reported in terms of data and ML/AI algorithm used. It was used as a criterion where only the well-reported articles would be included. Four

independent reviewers scored the articles using STAR-ML version 2 [76].

Data Charting Process Following the identification of the eligible publications, all relevant data were gathered in a structured way using a Microsoft Excel file.

The extracted data contain pain context, focus on the mechanism, the number of participants, if only female participants were present in the data, the focus of the application, data source, type of AI/ML algorithm, AI/ML algorithm used, data availability, and pain-related questionnaires or scales used.

The extracted relevant data also contains the rationale for the use of the AI/ML algorithm, variables used, imbalanced data, imputation, feature selection, feature scaling, training, hyperparameter tuning, and performance metrics. Altogether, these data formed the basis of the analysis. A uniform approach was sought by two reviewers on included 60 studies in the review. However, in practice, it was often challenging (sometimes impossible) to extract relevant information for the listed fields either due to the nature of the work (i.e., data, algorithm) or where research reports failed to include relevant information. Though we used STAR-ML to consider only high-quality reports, not all of them have perfect scores as data are not always presented in the most accessible of formats in the reports [185].

The extracted data were grouped into five categories based on the focus of the application, i.e., diagnosis, prognosis, clinical decision support, self-management, and rehabilitation. They were reported accordingly in two separate tables from the perspectives of the data and algorithm used.

Data Extraction The articles which passed the full-text screening step were passed to the 'Extraction' step into Covidence. From a total of 691 articles, 60 met the

inclusion criteria for data extraction for this scoping review.

2.5 Conducting the Review

Figure 2.3 shows the PRISMA diagram (based on PRISMA 2020) where all the data of each step are recorded.

From the four databases, in total, 691 studies (MEDLINE n = 320, Web of Science Core Collection n = 309, IEEE Xplore n = 55, ACM Digital Library n = 7) were identified that were published in the past decade (2012 to February 2022). There were overlaps in the search results, mostly between MEDLINE and Web of Science (197 duplicates). In total, 207 duplicates were identified by Covidence and discarded.

Before the full-text screening, 484 remaining articles after the duplicate removal were screened based on the abstracts by two independent reviewers. A total of 195 studies were excluded based on the set exclusion criteria (language, topic, i.e., CP, method, i.e., AI/ML, publication type, i.e., original articles, and population, i.e., human). Though these criteria existed for MEDLINE and Web of Science, they were not perfect, and in the case of other databases, all these criteria were not present. The conflicts between the two reviewers were resolved by consensus.

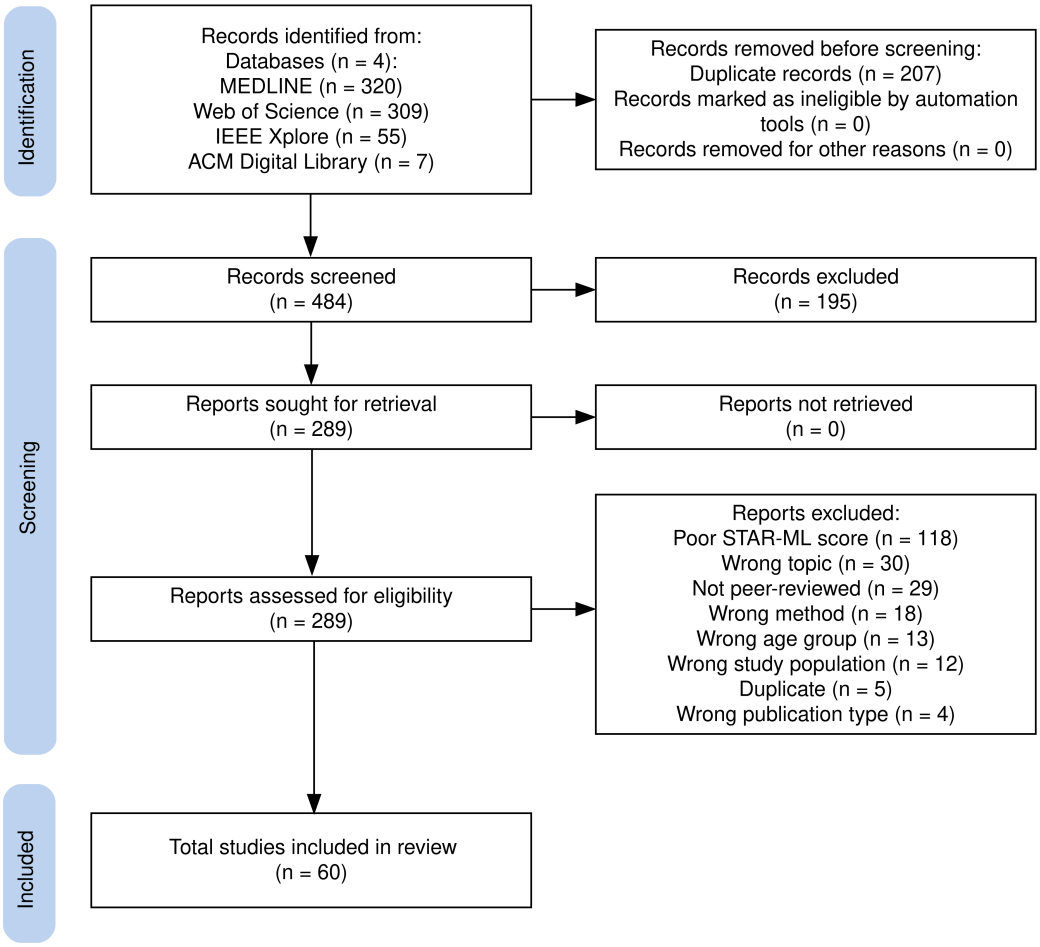


Figure 2.3: PRISMA flow diagram depicting the data in every step taken for this review. Out of 691 identified articles, 60 were included in the review.

The remaining 289 papers went into the second level or full-text screening, and we were able to retrieve all of them. These 289 papers were assessed based on the 10 eligibility criteria that we set. During the screening process, the conflicts were resolved by consensus and 60 papers were included in the final review. At this stage, 229 articles were excluded by the following criteria:

- Poor STAR-ML score (≤ 5): 118 papers

- Wrong topic: 30 papers
- Not peer-reviewed: 29 papers
- Wrong method: 18 papers
- Wrong age group: 13 papers
- Wrong study population: 12 papers
- Duplicates (missed by Covidence or extended publications): 5
- Wrong publication type: 4 papers

Some of the articles had more than one reason to exclude, but the reason was selected based on the sequence of the exclusion criteria (Appendix A).

2.6 Reporting the Review

This section provides a summary of the contributions from the selected papers according to areas of interest, i.e., data and algorithm. Five focus of application was chosen considering the overall CP field and application of AI, and they are listed below:

1. Diagnosis
2. Prognosis
3. Clinical Decision Support
4. Self-Management
5. Rehabilitation

First, a few charts and figures are presented to give an overview of the field. Then a summary table is presented containing relevant information focused on data and categorized by the focus of the application (Table 2.2). Then, another table (Table 2.3) is presented, focused on the ML/AI algorithms and their implementation, categorized by the focus of the application.

2.6.1 Applications of AI/ML in CP Research

The most prevalent diagnosis of AI applications in CP research is for diagnosis. The other major fields of application (i.e., clinical decision support, prognosis, self-management, and rehabilitation) have not got enough attention. No publication focused on Rehabilitation where ML/AI was used as a method. Figure 2.4 shows the number of publications in each of the categories of focus.

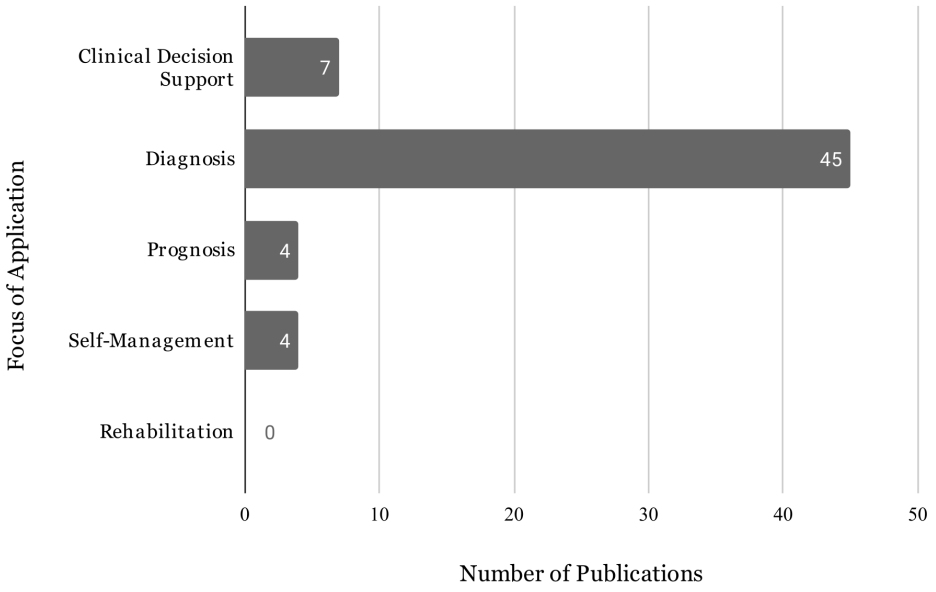


Figure 2.4: Focus of AI/ML applications in CP research. Table 2.2 contains detailed information categorized by the focus of applications.

2.6.1.1 Data Used

Data from different sources were utilized in the articles. However, several articles utilized multiple sources of data (Table 2.2). Though MRI and fMRI was the most common choice, clinical and questionnaire data were also used by a significant number of articles (Figure 2.5).

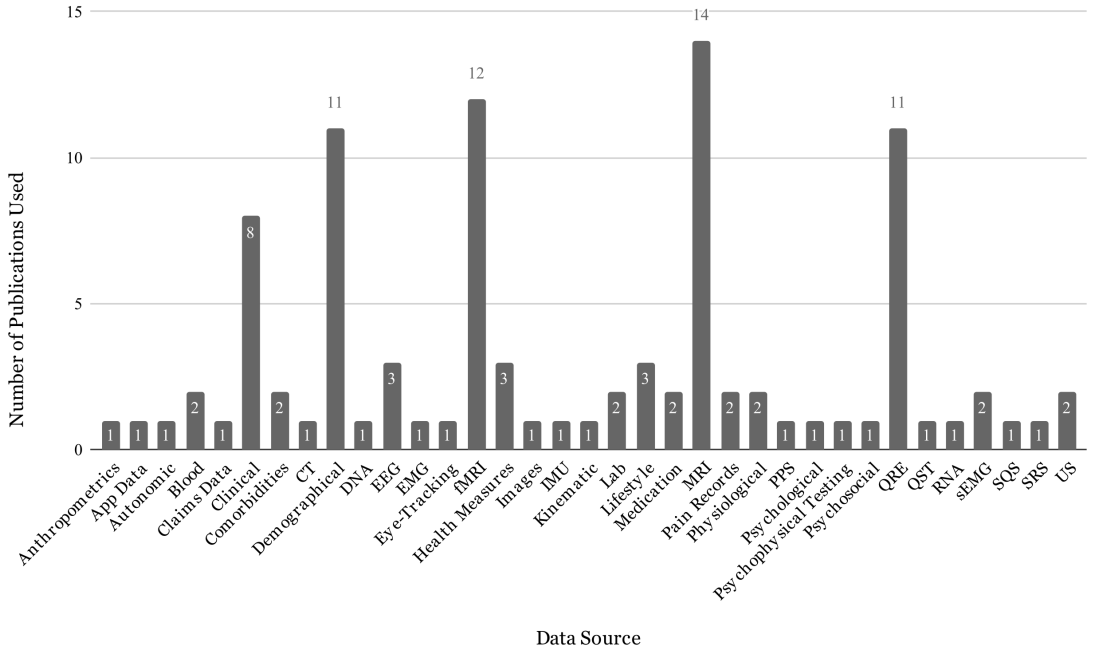


Figure 2.5: Column chart of the data sources used in CP research. The abbreviations of the acronyms are provided as a footnote under Table 2.2.

Questionnaire has the potential as it considers important factors like perception, fear, and anxiety, which are crucial in pain treatment. Sometimes left unused, 29 out of 60 articles mentioned having pain-related questionnaire data in their datasets.

2.6.1.2 CP Conditions

A column chart is given in Figure 2.6. Fibromyalgia (FM) and cLBP were considered in most research. The number of research mentioned CP was 10, where they discussed about CP in general and tried to distinguish CP from healthy controls (HCs). However, the focus on the mechanism is rare.

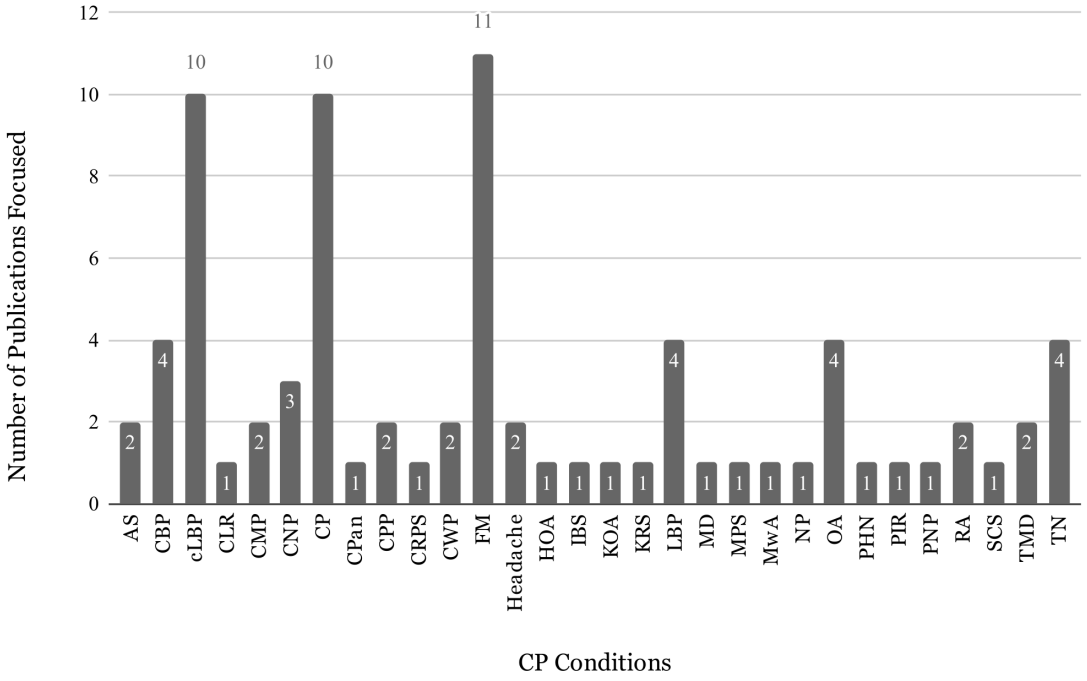


Figure 2.6: Column chart of CP condition in focus in the publications. The abbreviations of the acronyms are provided as a footnote under Table 2.2.

All data-relevant information, along with the type of ML used (i.e., supervised, semi-supervised, unsupervised, and reinforcement), has been presented in Table 2.2. The supervised algorithms were more dominant than Unsupervised Learning (60 supervised, 13 unsupervised). However, no application of Semi-supervised or Reinforcement Learning was found.

Table 2.2: Information of included studies (focused on data used)

Authors	**Year**	**Pain Context**	**Mech. Focus**	**Participant Number**	**Female Only**	**Data Source**	**Type of ML**	**Data Availability**
Clinical Decision Support								
Almeida et al. [186]	2019	CMP	No	80	Yes	PPS, Psychosocial	Unsupervised	No
Fodeh et al. [187]	2017	PIR, MD	No	184	No	Clinical	Supervised	No
Kang et al. [188]	2012	CNP	No	121	No	Demographical, Health Measures, QRE	Unsupervised	No
Pancino et al. [189]	2021	CP	No	376	-	Eye-Tracking Images	Supervised, Unsupervised	No
Rogachov et al. [190]	2018	CP, AS	No	133	No	fMRI, Psychophysical Testing	Supervised	No
Santana et al. [191]	2020	CP	No	997	No	QRE, QST	Supervised	No
Santana et al. [192]	2019	FM, CBP	No	158	No	fMRI	Supervised	AURR
Diagnosis								
Alge et al. [193]	2020	FM	No	26	-	SQS	Supervised	No
Antonucci et al. [194]	2020	CP	No	204	-	QRE	Supervised	AURR
Bagarinao et al. [195]	2014	CPP	No	66	Yes	MRI	Supervised	No
Bair et al. [196]	2016	TMD	No	4678	No	QRE, Health Measures, Clinical	Supervised	No

Table 2.2 continued from the previous page

Authors	Year	Pain Context	Mech. Focus	Participant Number	Female Only	Data Source	Type of ML	Data Availability
Barroso et al. [197]	2020	KOA, HOA	Neuropathic	151	No	MRI, QRE	Supervised	No
Behr et al. [198]	2019	MPS	No	69	No	US	Supervised	No
Behr et al. [199]	2020	FM	No	128	-	US	Supervised	No
Bianchi et al. [200]	2020	OA	No	92	No	Blood, CT	Supervised	AURR
Callan et al. [201]	2014	CBP	No	26	No	fMRI	Supervised	No
Caza-Szoka et al. [202]	2016	cLBP	No	24	-	sEMG	Supervised	No
Chen et al. [203]	2022	LBP	No	10	-	EEG	Supervised	No
Cheng et al. [204]	2018	AS	Neuropathic	133	No	MRI	Supervised	No
Gaynor et al. [205]	2021	TMD, FM, TN, Headache	No	5585	No	Demographical, Clinical	Unsupervised	No
Gudin et al. [206]	2020	CP	No	631	-	Demographical, Clinical, Medication	Supervised, Unsupervised	AURR
Harte et al. [207]	2016	FM	No	62	Yes	fMRI	Supervised	No
Holton et al. [208]	2020	CWP	No	40	No	QRE	Unsupervised	No

Table 2.2 continued from the previous page

Authors	**Year**	**Pain Context**	**Mech. Focus**	**Participant Number**	**Female Only**	**Data Source**	**Type of ML**	**Data Availability**
Ichesco et al. [209]	2021	FM	No	28	Yes	fMRI	Supervised	AURR
Jimenez-Grande et al. [210]	2021	CNP	No	40	-	Kinematic	Supervised	No
Jimenez-Grande et al. [211]	2021	CNP	No	40	No	EMG	Supervised	Yes
Lamichhane et al. [212]	2021	LBP	No	51	No	fMRI	Supervised	AURR
Lamichhane et al. [213]	2021	LBP	No	51	No	fMRI, MRI	Supervised	No
Larsson et al. [214]	2017	CP	No	2457	No	Demographical	Unsupervised	No
Lee et al. [215]	2019	cLBP	No	53	No	fMRI, Autonomic	Supervised	No
Levitt et al. [216]	2020	CLR	No	57	No	EEGs	Supervised	No
Mano et al. [217]	2018	cLBP	No	97	-	MRI	Supervised, Unsupervised	Yes
Mao et al. [218]	2020	IBS	No	68	-	MRI	Supervised	AURR
Miettinen et al. [219]	2021	CP	Neuropathic	277	No	Pain Records, Psychological, Demographical, Lifestyle, Comorbidities	Supervised, Unsupervised	No

Table 2.2 continued from the previous page

Authors	Year	Pain Context	Mech. Focus	Participant Number	Female Only	Data Source	Type of ML	Data Availability
Minerbi et al. [220]	2019	FM	No	156	Yes	DNA, RNA, Demographical, Anthropometrics, Comorbidities, Medication, Lifestyle	Supervised	Yes
Mo et al. [221]	2021	TN	No	126	No	MRI	Supervised	AURR
Morales et al. [222]	2021	OA	No	4796	-	MRI	Supervised	Yes
Olesen et al. [223]	2016	CPan	No	60	No	Demographical, Clinical	Supervised	No
Ozkan et al. [224]	2016	FM	No	86	-	Blood, Physiological, SRS	Supervised	No
Pinedo-Villanueva et al. [225]	2018	KRS	No	126064	No	QRE	Unsupervised	No
Richter et al. [226]	2021	cLBP	No	4420	No	Claims Data	Supervised	AURR
Russo et al. [227]	2020	CRPS	Neuropathic	29	No	QRE, Lab	Supervised	No
Shen et al. [228]	2019	cLBP	No	197	-	fMRI	Supervised	No
Shim et al. [229]	2021	LBP	No	6119	No	Demographical, Clinical	Supervised	AURR
Ta Dinh et al. [230]	2019	CBP, NP, CWP, PHN, PNP	Neuropathic	185	No	EEG	Supervised	AURR

Table 2.2 continued from the previous page

Authors	Year	Pain Context	Mech. Focus	Participant Number	Female Only	Data Source	Type of ML	Data Availability
Thieme et al. [231]	2015	FM	No	120	Yes	Physiological	Unsupervised	No
Tu et al. [232]	2020	cLBP, FM, MwA	No	230	No	fMRI, MRI	Supervised	AURR
Tu et al. [233]	2019	cLBP	No	50	No	fMRI	Supervised	No
Tuechler et al. [234]	2020	cLBP	No	263	No	QRE	Supervised	No
You et al. [235]	2021	CMP	No	109	No	fMRI, MRI	Supervised	No
Zhong et al. [236]	2018	TN	Neuropathic	46	No	MRI	Supervised	No
Zhou et al. [237]	2020	cLBP	No	57	No	MRI	Supervised	No
Prognosis								
Lin et al. [238]	2021	CPP	No	66	Yes	MRI	Supervised	No
Lotsch et al. [239]	2020	RA	No	288	No	Demographical, Clinical, Lab	Supervised	No
Ounajim et al. [240]	2021	SCS	No	103	No	Demographical, Health Measures, QRE	Supervised	No
Shih-Ping Hung et al. [241]	2022	TN, cLBP, OA	No	959	No	MRI	Supervised	Yes
Self-Management								
Frostholm et al. [242]	2018	CP	No	424	No	QRE	Unsupervised	No

Table 2.2 continued from the previous page

Authors	Year	Pain Context	Mech. Focus	Participant Number	Female Only	Data Source	Type of ML	Data Availability
Rahman et al. [243]	2018	FM, CBP, RA, OA, Headache	Neuropathic	782	No	Demographical, Clinical, App Data	Supervised	No
Rahman et al. [244]	2019	CP	No	879	No	Pain Records	Supervised, Unsupervised	No
Wang et al. [245]	2021	CP	No	30	-	IMU, sEMG	Supervised	Yes

Acronyms: Mech. Focus: Mechanism Focus, '-': not reported

Pain Context AS: Ankylosing Spondylitis, CBP: Chronic Back Pain, cLBP: Chronic Lower Back Pain, CLR: Chronic Lumbar Radiculopathy, CMP: Chronic Musculoskeletal Pain, CNP: Chronic Neck Pain, CP: Chronic Pain, CPan: Chronic Pancreatitis, CPP: Chronic Pelvic Pain, CRPS: Complex Regional Pain Syndrome, CWP: Chronic Widespread Pain, FM: Fibromyalgia, IBS: Irritable Bowel Syndrome, KRS: Knee Replacement Surgery, LBP: Lower Back Pain, MD: Musculoskeletal Diagnosis, MPS: Myofascial Pain Syndrome, MwA: Migraine without Aura, NP: Neuropathic Pain, OA: Osteoarthritis (KOA- Knee, HOA- Hip), PHN: Post-Herpetic Neuralgia, PIR: Pain Intensity Rating, PNP: Polyneuropathic Pain, RA: Rheumatoid Arthritis, SCS: Spinal Cord Stimulation, TMD: Temporomandibular Disorders, TN: Trigeminal Neuralgia

Data Source App Data: Application Use Information CT: Computerized Tomography DNA: Deoxyribonucleic Acid EEG: Electroencephalogram EMG: Electromyography fMRI: Functional Magnetic Resonance Imaging Health Measures: e.g., Quantitative Measures of Health & Scores IMU: Inertia Measurement Unit Lab: Laboratory Analysis Lifestyle: e.g., Dietary intake, smoking, alcohol consumption MRI: Magnetic Resonance Imaging PPS: Pressure Pain Sensitivity QRE: Questionnaire QST: Quantitative Sensory Testings RNA: Ribonucleic Acid sEMG: Surface Electromyography SQS: Sleep Quality Scale SRS: Sympathetic Response Skin Measurements US: Ultrasound

2.6.1.3 AI/ML Method Used

A total of 39 different ML algorithms were found to be used in the articles. SVM was the first choice among the algorithms, where LR and RF hold the 2nd and 3rd spots. In most cases, it was driven by the simplicity and recognition of these algorithms that led the authors to use the algorithms.

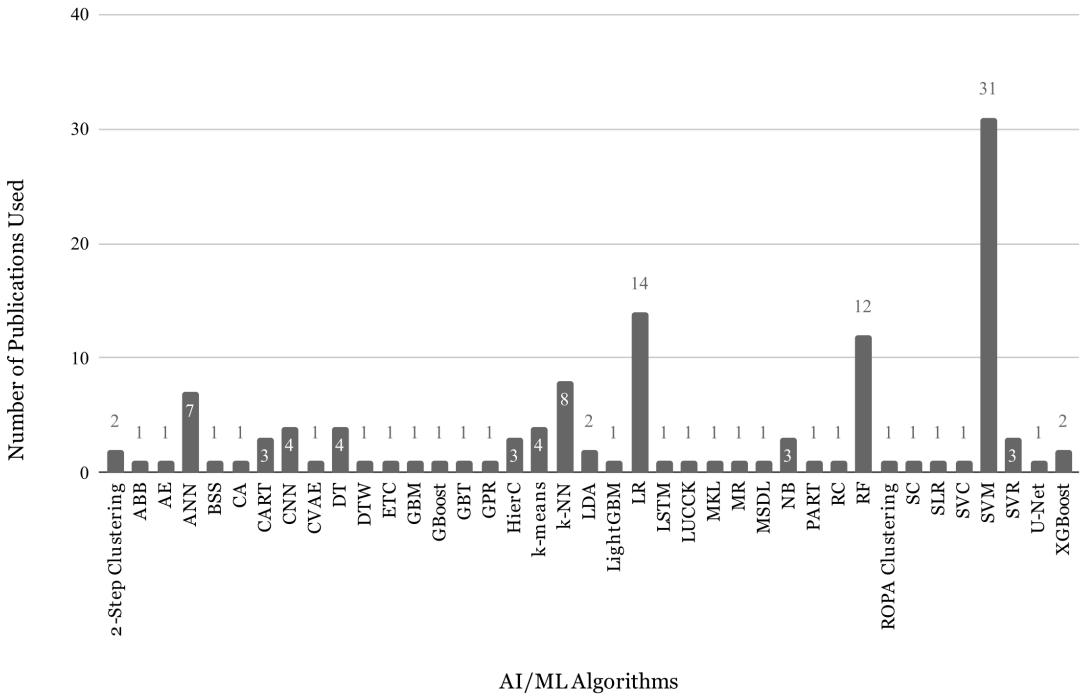


Figure 2.7: Column chart of AI/ML algorithm used in the publications. The abbreviations of the acronyms are provided as a footnote under Table 2.3.

All ML-relevant information, along with the context of CP and participant information, has been presented in Table 2.3.

Table 2.3: Information of included studies (focused on ML/AI algorithm used)

Authors	ML Algorithm	Reason of Use	Variables Used	Imbalanced Data	Imputation	Feature Selection	Feature Scaling	Training	Tuning	Metrics
Clinical Decision Support										
Almeida et al. [186]	CA	1	Numerical, Categorical	Yes	No	Yes	Yes	4	No	Accuracy
Fodeh et al. [187]	k-NN, DT, SVM, RF	1, 3	Text	Yes	No	No	NA	1	No	Accuracy, F1-Score, Sensitivity, AUC
Kang et al. [188]	2-Step Clustering	3	Numerical, Categorical	-	No	NA	Yes	NA	No	p-value
Pancino et al. [189]	AE, k-NN	3	Numerical, Images	Yes	No	No	NA	NA	Yes	Accuracy
Rogachov et al. [190]	MR	-	Numerical	Yes	No	Yes	NA	4	Yes	MSE, p-value
Santana et al. [191]	CNN, SLR, SVM, DTW, MSDL, ABB	2	Numerical	Yes	Yes	Yes	Yes	4	Yes	Balanced Accuracy, AUC, log-loss
Santana et al. [192]	RC, LR, SVC, k-NN, DT, RF, ETC, ANN, XGBoost	3	Numerical, Categorical	Yes	Yes	No	NA	4	Yes	Balanced Accuracy, AUC, log-loss, Precision, Recall
Diagnosis										
Alge et al. [193]	SVM	1	Numerical	Yes	No	Yes	Yes	4	Yes	Balanced Accuracy, Sensitivity, Specificity, PPV, NPV, AUC

Table 2.3 continued from the previous page

Authors	ML Algorithm	Reason of Use	Variables Used	Imbalanced Data	Imputation	Feature Selection	Feature Scaling	Training	Tuning	Metrics
Antonucci et al. [194]	SVM	1, 2	Numerical	No	No	NA	Yes	3	No	Accuracy, Sensitivity, Specificity, PPV, NPV
Bagarinao et al. [195]	SC	-	Numerical, Categorical	Yes	No	Yes	Yes	1	Yes	Accuracy, Kaplan-Meier plots
Bair et al. [196]	LR	-	Numerical	Yes	No	No	-	4	Yes	AUC
Barroso et al. [197]	SVM	1, 2	Numerical	Yes	No	NA	NA	4	Yes	Accuracy, Sensitivity, Specificity, MCC
Behr et al. [198]	SVM, LR	1, 2, 3	Numerical	Yes	No	No	Yes	4	Yes	Accuracy
Behr et al. [199]	LR, RF, LightGBM, XGBoost	1	Numerical, Categorical	No	No	Yes	Yes	4	Yes	Accuracy, F1-score, AUC
Bianchi et al. [200]	LR	1, 3	Numerical	No	No	Yes	Yes	3	No	Accuracy, Sensitivity, Specificity
Callan et al. [201]	ANN	-	Numerical	No	No	Yes	Yes	3	No	-
Caza-Szoka et al. [202]	CNN	1, 2	Time Series	Yes	No	Yes	Yes	3	Yes	AUC
Chen et al. [203]	LR	-	Numerical	Yes	No	Yes	No	4	Yes	Spearman Correlation Coefficient
Cheng et al. [204]	ROPA Clustering, k-means	6	Numerical, Categorical	No	No	No	Yes	NA	No	-

Table 2.3 continued from the previous page

Authors	ML Algorithm	Reason of Use	Variables Used	Imbalanced Data	Imputation	Feature Selection	Feature Scaling	Training	Tuning	Metrics
Gaynor et al. [205]	k-prototypes, SVR, RF	-	-	Yes	Yes	Yes	Yes	1, 4	Yes	AUC
Gudin et al. [206]	SVM	-	Images	Yes	No	No	NA	4, 3	Yes	Accuracy, Sensitivity, Specificity
Harte et al. [207]	k-means	-	Numerical, Categorical	Yes	No	No	NA	3	No	-
Holton et al. [208]	SVM	4	Numerical	Yes	No	No	NA	3	No	Accuracy
Ichesco et al. [209]	k-NN, SVM, LDA	-	Numerical	No	No	Yes	NA	4	No	Accuracy, Sensitivity, Specificity
Jimenez-Grande et al. [210]	SVM, k-NN, LDA	3	Numerical	No	No	Yes	Yes	4	Yes	Accuracy, Sensitivity, Specificity
Jimenez-Grande et al. [211]	SVM	3	Numerical	Yes	No	Yes	Yes	1, 4	Yes	Accuracy
Lamichhane et al. [212]	SVM	3	Numerical	Yes	No	Yes	NA	3, 4	Yes	Accuracy, Sensitivity, Specificity
Lamichhane et al. [213]	2-Step Clustering	3	Numerical, Categorical	-	Yes	No	Yes	NA	No	BIC
Larsson et al. [214]	SVM, SVR	-	Numerical	Yes	No	No	Yes	1, 3	No	Accuracy, Pearson correlation, RMS
Lee et al. [215]	SVM	-	Numerical	Yes	Yes	Yes	Yes	4	No	Accuracy

Table 2.3 continued from the previous page

Authors	ML Algorithm	Reason of Use	Variables Used	Imbalanced Data	Imputation	Feature Selection	Feature Scaling	Training	Tuning	Metrics
Levitt et al. [216]	SVM, CVAE	1	Numerical	Yes	Yes	Yes	NA	5	No	Accuracy, Sensitivity, Specificity
Mano et al. [217]	SVM	1	Numerical	No	No	Yes	Yes	4	No	Accuracy, Sensitivity, Specificity, AUC
Mao et al. [218]	HierC, CART, PART, RF, ANN	-	Numerical, Categorical	Yes	Yes	Yes	NA	1	No	Balanced Accuracy, F1-score, Sensitivity, Specificity, PPV, NPV, AUROC, Discriminant power, Youden's index
Miettinen et al. [219]	LR, SVM	-	Omics Data	Yes	No	Yes	Yes	3, 4	No	ROC, AUC
Minerbi et al. [220]	SVM, LR	-	Numerical	Yes	No	Yes	Yes	3, 4	No	Sensitivity, Specificity, AUC
Mo et al. [221]	CNN, LR	2, 3	Images	Yes	No	Yes	Yes	2	Yes	Sensitivity, Specificity, AUC
Morales et al. [222]	LUCCK, SVM	-	Numerical	Yes	No	Yes	Yes	4	No	F1-score, Sensitivity, Specificity, AUC
Olesen et al. [223]	SVM	-	Numerical, Categorical	Yes	No	No	NA	3	No	Accuracy

Table 2.3 continued from the previous page

Authors	ML Algorithm	Reason of Use	Variables Used	Imbalanced Data	Imputation	Feature Selection	Feature Scaling	Training	Tuning	Metrics
Ozkan et al. [224]	ANN	1	Numerical	Yes	No	No	-	4	No	Accuracy, Sensitivity, Specificity
Pinedo-Villanueva et al. [225]	HierC	1, 2, 5	Categorical	Yes	No	NA	NA	NA	No	-
Richter et al. [226]	BSS, SVM, RF	3	Categorical	Yes	Yes	Yes	NA	1	Yes	AUC
Russo et al. [227]	LR, DT, GBoost	-	Numerical, Categorical	Yes	No	No	Yes	1	No	AUC
Shen et al. [228]	SVM	-	Numerical	Yes	No	No	Yes	4	No	Accuracy, Sensitivity, Specificity
Shim et al. [229]	LR, k-NN, NB, DT, RF, GBM, SVM, ANN	1, 2	Numerical, Categorical	Yes	No	Yes	NA	1, 4	Yes	Accuracy, Sensitivity, Specificity, AUROC
Ta Dinh et al. [230]	SVM	1	Numerical	Yes	No	Yes	Yes	4	No	Accuracy, Sensitivity, Specificity
Thieme et al. [231]	k-means	-	Numerical	Yes	No	No	Yes	NA	No	-
Tu et al. [232]	SVR	-	Numerical	Yes	No	No	NA	4	No	R^2
Tu et al. [233]	SVM	-	Numerical	Yes, No, Yes, Yes	No	Yes	Yes	3	No	Accuracy, Sensitivity, Specificity, AUC
Tuechler et al. [234]	RF	3	Categorical	Yes	Yes	No	NA	1	Yes	Accuracy

Table 2.3 continued from the previous page

Authors	ML Algorithm	Reason of Use	Variables Used	Imbalanced Data	Imputation	Feature Selection	Feature Scaling	Training	Tuning	Metrics
You et al. [235]	MKL	3	Numerical	-	-	-	-	3	Yes	Accuracy, Specificity, ROC
Zhong et al. [236]	SVM	-	Numerical	No	No	Yes	Yes	3	No	Balanced Accuracy, Sensitivity, Specificity, AUC
Zhou et al. [237]	U-Net	1, 2, 3, 4	Images	Yes	No	NA	NA	1	No	IoU, Dice
Prognosis										
Lin et al. [238]	SVM	-	Numerical	No	No	No	Yes	3	Yes	Accuracy, Sensitivity, Specificity
Lotsch et al. [239]	CART, k-NN, SVM, ANN, NB	4	Numerical, Categorical	Yes	No	Yes	Yes	1	No	Balanced Accuracy, F1-score, Sensitivity, Specificity, PPV, NPV
Ounajim et al. [240]	LR, NB, ANN, SVM, CART, RFs, GBT	1	Numerical, Categorical	Yes	No	Yes	Yes	3, 4	Yes	Accuracy, Sensitivity, Specificity
Shih-Ping Hung et al. [241]	GPR	2	Numerical	Yes	No	No	NA	4	No	MAE, RMS
Self-Management										
Frostholm et al. [242]	HierC	3	Numerical, Categorical	Yes	Yes	No	NA	NA	No	1-way ANOVA

Table 2.3 continued from the previous page

Authors	ML Algorithm	Reason of Use	Variables Used	Imbalanced Data	Imputation	Feature Selection	Feature Scaling	Training	Tuning	Metrics
Rahman et al. [243]	k-NN, LR, RF, SVM	1, 3	Numerical, Categorical	Yes	No	No	-	4	Yes	Accuracy
Rahman et al. [244]	k-means, LR, RF, SVM	1	Numerical, Categorical	Yes	No	Yes	NA	4	No	Accuracy
Wang et al. [245]	RF, CNN, LSTM	6	Numerical, Time Series	Yes	Yes	Yes	NA	4, 5	Yes	Accuracy, F1-score, Precision, Recall

Acronyms: '-': not reported

ML Algorithm ABB: Ann4BrainsBatch, AE: Autoencoder, ANN: Artificial Neural Network, BSS: Best Subsets Selection, CA: Cluster Analysis, CART: Classification and Regression Tree, CNN: Convolutional Neural Network, CVAE: Conditional Variational Autoencoder, DT: Decision Tree, DTW: Dynamic Time Warping, ETC: Extra Trees Classifer, GBM: Gradient Boosting Machines, GBoost: Gradient Boosting, GBT: Gradient Boosted Tree, GPR: Gaussian Process Regression, HierC: Hierarchical Clustering, k-means: k-means Clustering, k-NN: K-Nearest Neighbors, k-prototype: k-Prototypes Clustering, LDA: Linear Discriminant Analysis, LightGBM: Light Gradient-Boosting Machine, LR: Logistic Regression, LSTM: Long Short-Term Memory Network, LUCCK: Learning Using Concave and Convex Kernels, MKL: Multiple Kernel Learning, MR: Multivariate Regression, MSDL: Marginal Space Deep Learning, NB: Naive Bayes Classifier, PART: Partial Decision Tree, RC: Random Chance, RF: Random Forest, ROPA: Rapid OPPERA Algorithm, SC: Supervised Clustering, SLR: Simple Linear Regression, SVC: Support Vector Classification, SVM: Support Vector Machine, SVR: Support Vector Regression, XGBoost: eXtreme Gradient Boosting

Reason of Use 1: ML used in same/similar field, 2: ML has shown good performance in this field, 3: Reasoning based on the nature of the algorithm and the data, 4: Increasing popularity/novelty, 5: Explainability and Interpretability 6: Others.

Training 1: train/test, 2: train/valid/test, 3: leave-one-out, 4: k-fold, 5: other, 6: NA (Clustering)

Tuning Yes, No, NA: Not Applicable

Metrics MSE: Mean Squared Error, AUC: Area Under Curve, ROC: Receiver Operating Characteristic Curve, AUROC: Area Under ROC Curve, PPV: Positive Predictive Value, NPV: Negative Predictive Value, IoU: Intersection over Union, DSC: Dice Similarity Coefficient, MAE: Mean Absolute Error, RMS: Root Mean Square Error, BIC: Bayesian Information Criterion, MCC: Matthews Correlation Coefficient, R^2: Squared Prediction-outcome Correlation

2.6.2 Gaps and Findings

In this section, the findings are presented, and the gaps are identified.

2.6.2.1 State of AI in CP Research

The application of AI in CP research has taken a significant leap over the past years. Publications by years shows how the number of publications is increasing over the years. The trend line in Figure 2.8 shows a rapid increase in the number of research in CP using AI/ML techniques. This indicates the potential of AI in the domain of CP.

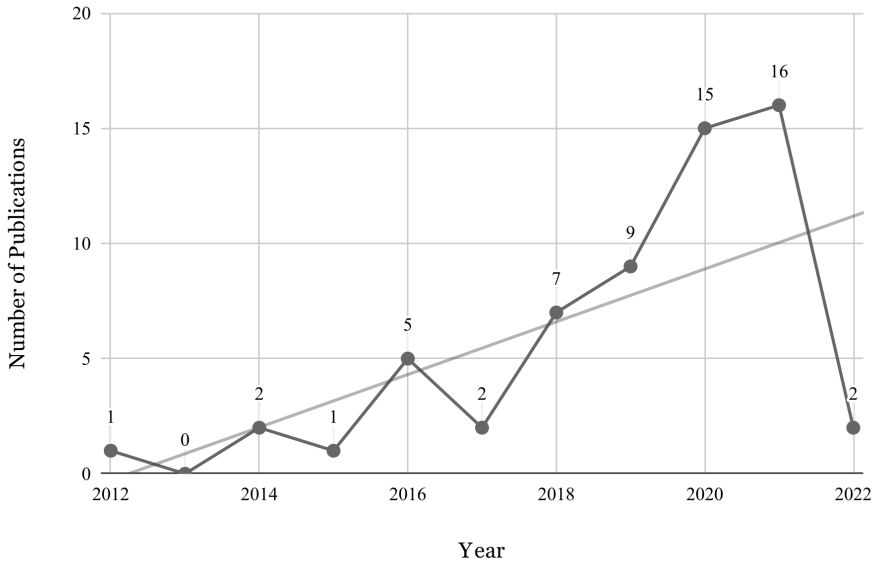


Figure 2.8: Line chart of the number of publications by years. For the year 2022, the publications were considered till the month of February (day of search).

2.6.2.2 Rationale of Choice

The rationale for using an algorithm should be informed and driven by proper reasoning. The rationale for the choice of the algorithms was clustered into six categories

(Figure 2.9). In many cases (23), the rationale was not reported. Also, the focus on explainability and interpretability was found to be extremely rare.

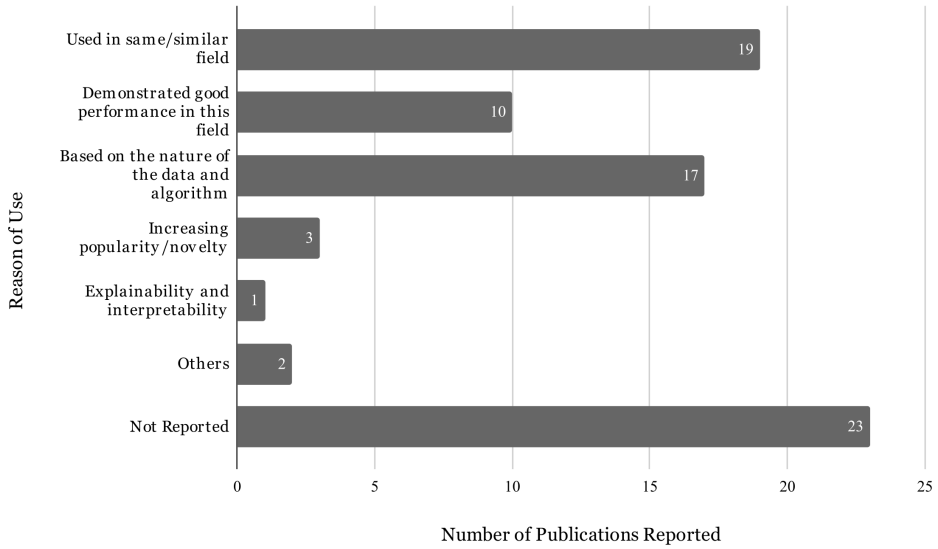


Figure 2.9: Rationale of selecting the AI/ML algorithm in CP research. Here the numbers are aggregated based on the six categories/rationales.

The spider or radar plot in Figure 2.10 depicts five important factors (handling data imbalance, missing data imputation, feature selection, feature scaling, and hyper-parameter tuning) in any AI/ML algorithm application. It has four axes representing if the answer was Yes, No, not reported (NR), and not applicable (NA).

A significant number of articles (22) did not handle data imbalance that can add bias and inconsistencies in algorithms performance. In this case, 'NA' was used where the data was balanced or handled using the algorithm. In terms of missing value imputation, only 10 articles mentioned it, others either did not use it or did not report it. Feature selection was used in most cases (32), and in a few cases (5), it was inherently done by the algorithm. Feature scaling is another very important aspect of

AI/ML algorithms, and the majority (55) of the articles either did it explicitly or it was handled by the AI/ML algorithm itself. However, hyperparameter tuning was not reported in most articles (34), indicating more improved performance could have been achieved.

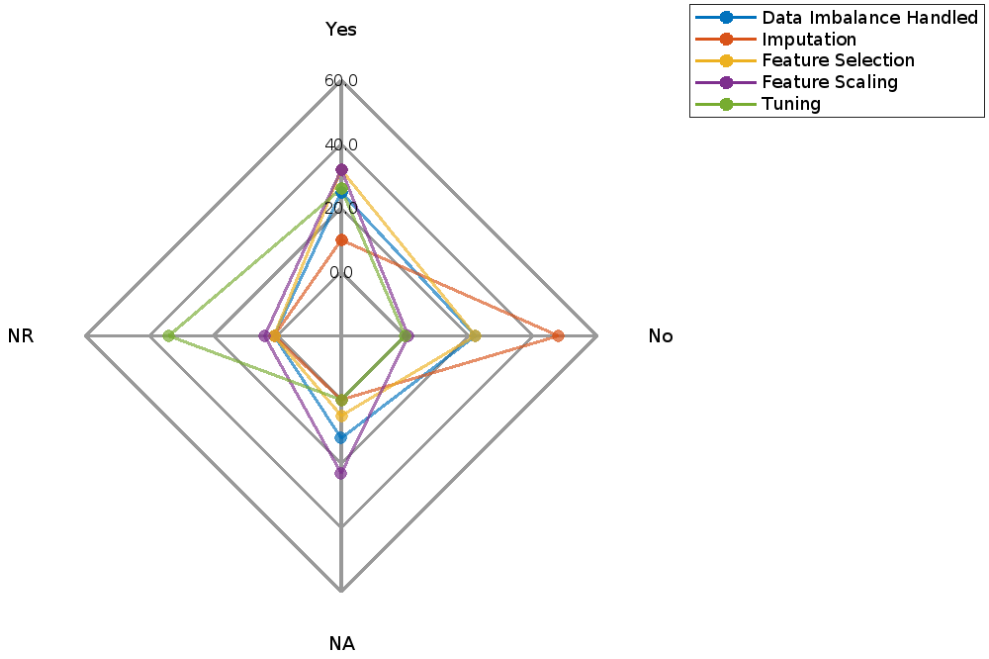


Figure 2.10: Spider plot of different ML-related aspects. Here, NA is 'Not Applicable' and NR is 'Not Reported'.

2.6.2.3 Future of AI in CP Research

Pain as a Continuum Though CP identification as a mechanism is being widely acknowledged as the optimal way of CP treatment and management, there is a significant gap of focus in work. Most of the works found focused on symptom-based or anatomical location-based pain identification and classification.

Only 7 articles out of 60 focused on Neuropathic pain. The other types of pain mechanisms (i.e., Nociceptive and Nociplastic) were not considered in any of the

reviewed articles. More emphasis should be put on identifying the pain mechanisms rather than anatomical location or symptoms-based pain classification.

Choice of Algorithms and Quality The choice of AI algorithm should be informed and should have a strong reason for using an existing algorithm or developing a new algorithm. As seen in Table 2.3, Figure 2.7 and 2.10, though varieties of algorithms were used, the emphasis on reasoning and proper reporting is expected for higher quality and transparency.

Focus on ML Algorithms The supervised algorithms were mostly used by researchers in the area of CP research. The other types of ML algorithms can have significant benefits and use cases. For example, Unsupervised Learning can reveal inherent trends in the data, which can help to minimize human bias. Another good candidate can be reinforcement learning, e.g., it can be used in a clinical setting to train an algorithm (i.e., decision support system) every time the physician or medical practitioner makes a decision which can get better with time.

Transparency and Reproducibilty In 57% of the publications, the model parameters and their tuning was not reported making the research irreproducible. The model parameters and tuning should be reported, especially considering the clinical application aspects.

2.6.2.4 Call for Open-sourced Pain Data

A lack of open-source data should not be overlooked. To get the best support from the scientific community and advance CP research rapidly, the availability of the

data is a major factor. Figure 2.11 demonstrates the status of data availability in the publications. It is understandable (and mentioned in the article in some cases) that the data might have sensitive information and/or have restrictions regarding publication. Yet, de-identified risk-free data should be shared or should be made more accessible by the researchers.

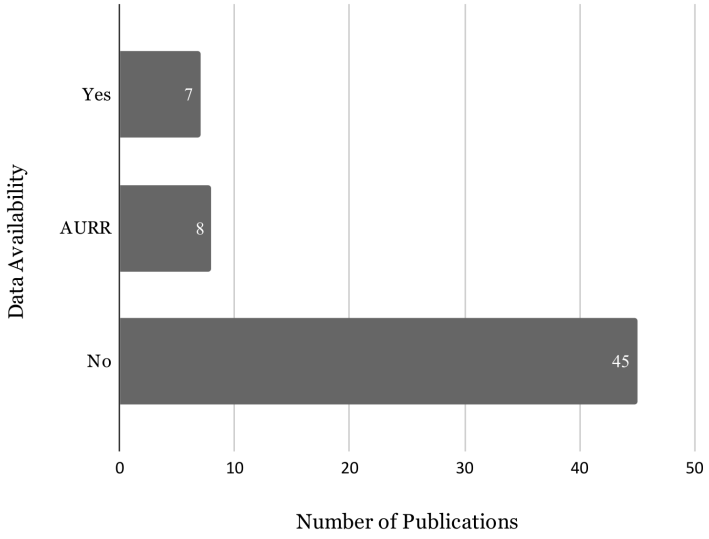


Figure 2.11: Dataset availability status in the articles. Here AURR stands for 'Available Upon Reasonable Request'.

Not sharing data and relevant information may also raise questions about the soundness and reproducibility of the work.

2.6.3 Limitations of the Scoping Review

The review presented aimed to survey a broad scope of studies to summarize the state of AI/ML methods used in CP research during the last 10 years. Though a comprehensive literature search developed through multiple iterations and by consulting with a

librarian and SMEs was performed on four research databases (MEDLINE, Web of Science Core Collection, IEEE Xplore, and ACM Digital Library), it is not possible to develop a perfect search query that can retrieve all relevant articles. For additional publications, reference lists were not checked, and this review did not consider the articles not indexed in these databases, which might overlook some not peer-reviewed conferences.

A strict set of criteria were determined before the review and used during the review. Although at least two reviewers administered each step of the review, the chance of discarding a relevant article was low, but the criteria might discard some applicable articles without having a significant reason. Additionally, the use of STAR-ML to include only high-quality publications were kept consistent by the agreement of at least two reviewers, yet, it might not be perfect and relevant works might be disregarded for achieving a low score in STAR-ML. Overall, the chances of human error were particularly low as the review established upon consensus in case of disagreement.

Chapter 3

Domain Data

This chapter discusses the datasets' description with population, explanation of the variables, data preprocessing, exploratory data analysis with descriptive statistics and visualizations. Additionally, it also presents the ethics approval for the study, data storage, and retention.

3.1 Data Description

3.1.1 Datasets and Study Population

Two datasets were obtained from University Health Network (UHN) [246] and processed accordingly to be able to apply AI/ML algorithms. Both the datasets contain mixed-type variables (numerical and categorical) and in a tabular format containing socio-demographics, consumption habits, pain characteristics, the impact of pain on various aspects of daily living, pain catastrophizing tendency, medical history, past and current pain treatment, patients' expectations, specific pain diagnosis established by pain

clinicians, etc. (check Appendix B.4 for details).

The datasets are comprised of two different cohorts. The first dataset was collected using DAta Driven Outcome System (DADOS) [247] platform and will be referred to as DADOSDD dataset. The second dataset is called Toronto Rehabilitation Institute's Clinic Dataset (TRICD).

The DADOSDD has 738 instances, and the TRICD has 201 instances. The data in the datasets were collected and prepared by UHN and de-identified by them as well. All study participants signed written consent that they understood and were willing to participate and share their data for research studies, which had been approved in accordance with the Institutional Review Board (UHN). A data sharing agreement was signed between UHN and McMaster University and an ethics application (MREB#: 5567) was approved at the McMaster Research Ethics Board (MREB) [248] for the data and the study (Section B.1 contains details).

3.1.2 Definition of Variables

The datasets are in a tabular format (originally in Microsoft Excel files) containing pain characteristics (duration, frequency, intensity, etc), impact on various aspects of daily living including sleep, specific pain diagnosis established by pain clinicians, psychological well-being (depression, anxiety) and pain catastrophizing tendency, health-related quality of life, medical history, consumption habits, past and current pain treatment, patient expectations, and socio-demographics [246]. However, the TRICD dataset does not have mechanistic CP classification or true labels. A comparison between the two datasets is provided in the Appendix (Table B.1).

3.1.2.1 Questionnaires in the Datasets

The validated questionnaires that are part of these datasets are listed below.

DADOSD

- Brief Pain Inventory (Short Form) [249, 250]

TRICD

- Brief Pain Inventory (Short Form) [249, 250]
- Pain Stages of Change Questionnaire (PSOCQ) [251]
- Pain Patient Profile (P3) [252]

3.1.3 Study Population and Subject Selection Criteria

The datasets had the same subject selection and screening criteria though they were not the same in terms of the parameters collected. The patients were screened by UHN while collecting the data and based on the consent provided. In order to be included in the study, the patient must be 18 (one exception in the DADOSD dataset with an age of 17.62 years) or older, have had pain for three months or more (CP), and accepted consecutive consultation request from referring doctors. Meeting these criteria, the patients had to answer self-administered and nurse-administered questionnaires (TRICD).

3.1.3.1 Population

The patients were referred to Tertiary Pain Clinics with different CP conditions. The DADOSD dataset was collected between November 2017 to October 2019, and the

TRICD dataset was collected between 2018 to 2020.

Age and Biological Sex The age range of the participants before and after data preprocessing is provided in Table 3.1. In both datasets, the number of female participants is higher than the number of males.

Table 3.1: Summary of age and sex information of the datasets

	Before Preprocessing		**After Preprocessing**	
	TRICD	DADOSD	TRICD	DADOSD
Minimum Age	19	17.62	19	17.62
Maximum Age	102	91.78	89	91.78
# of Male	71	248	70	154
# of Female	130	481	125	297
Male-Female Ratio	0.55	0.52	0.56	0.52

3.1.4 Datatype

Both datasets consist of mixed-type variables, i.e., both of them have numerical and categorical features.

3.2 Data Preprocessing

Both datasets have missing values, inconsistent and invalid inputs, and duplicate or redundant entries. This section describes the steps taken to preprocess the datasets to be able to apply the algorithms. Table 3.2 presents a summary of number of examples (i.e., patients) and features before and after preprocessing. For the DADOSD dataset,

patients' data were not considered if the CP mechanism labels were unavailable (to be able to measure the performance of the ML models).

Table 3.2: Number of instances and features

	Before Preprocessing		**After Preprocessing**	
	TRICD	DADOSD	TRICD	DADOSD
Number of Instances	201	738	195	451
Number of Features	60	146	28	44

A list containing reasons for removing the rows and columns can be found in the Appendix B.3.

3.2.1 Data Cleaning

The data cleaning was a series of steps led by the issues present in the data. Figure 3.1 shows the steps taken to clean the datasets. These steps were carefully chosen after the initial analysis of the datasets.

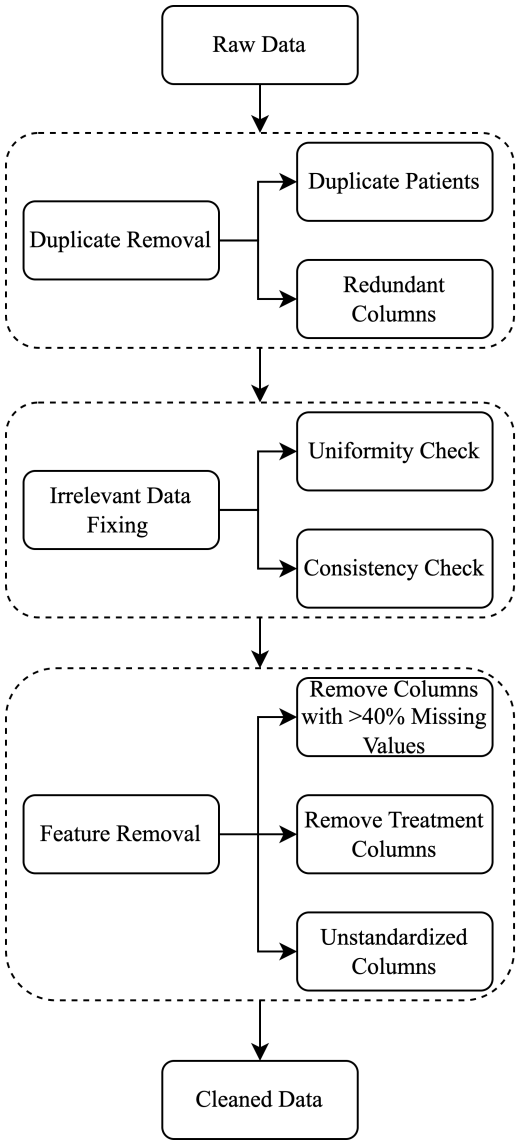


Figure 3.1: The steps to clean the data for analysis

3.2.1.1 Duplicate Removal

First, the features were renamed (i.e., spaces removed, readability enhancements) to be able to load the datasets in the Python (programming language) environment [253]. TRICD dataset had duplicate entries, which were removed.

DADOSD No duplicate rows were found.

TRICD 6 rows/instances were removed as they were duplicates.

3.2.1.2 Irrelevant Data Fixing

Unit Uniformity Check All the data fields were checked for irrelevant or inconsistent data units.

DADOSD The DADOSD data had multiple columns that had issues with units (e.g., percentages and decimals in a single column). These were fixed for every column having inconsistent units.

TRICD The TRICD data also had multiple columns that had issues with units (e.g., numbers to calculate percentage instead of the calculated percentage). Inconsistencies in units were fixed for all the columns where applicable.

Consistency Check Every column for both datasets was checked for consistency in terms of formatting.

DADOSD A few format fixes that were done are listed below:

- Pain_Feel column was fixed for semicolon issues and was separated into 15 one-hot encoded columns
- Patients' age was calculated from the date of birth and date of enrolment.
- Nicotine_Smoking, Alcohol_Consumption, Recreational_Drugs, Nicotine_Smoking_Amount columns were fixed for text formatting.

TRICD Format fixes that were done are listed below:

- Recalculated percentage for BPI_Pain_Severity, BPI_Relief, BPI_Pain_Interference

3.2.2 Encoding

The categorical columns have been coded by giving numerical values. LabelEnocoder is used to do that [254]. It was used to normalize labels and to transform non-numerical labels into numerical labels, e.g., no, yes, and unknown to 0, 1, and 2. The derived or encoded columns are kept, and the old columns are removed. Manual fixing was also done in some cases (e.g., 'None' and 'N/A' were coded to 0 for Recreational_Drugs column in DADOSD data).

3.2.3 Recoding

Some columns were recoded after consulting with the SMEs. It was done for three reasons, i.e., data were missing and the possible reason was known ('unknown' was introduced for Smoke_Categorized column in TRICD data for the subjects where there was no information), to address interpretability (Employment_Status was recoded to employed, unemployed, and retired in TRICD data), and coded inconsistently ('No' and 'Yes' were recoded to 0 and 1 for every columns where applicable).

3.2.4 Unstandardized Column Removal

The free-text answers were unstandardized, e.g., Pain_Worse_Reason, Pain_Better_Reason was removed after consulting with SMEs (Dr. Kumbhare and Dr. Samah Hassan). They were removed as there was no reasonable way to standardize them, i.e., these answers were free text and could not be categorized without adding bias. Additionally, the datasets are retrospective, and it is not possible to collect the data points again.

The features containing treatments were also removed to avoid chances of bias as recommended by the SMEs.

3.2.4.1 Missing Data Handling

Even after discarding the features with data quality issues, there were missing values to be handled. Both datasets have missing values in them. However, the causes of the missing data in the remaining fields in the datasets were unknown. The most likely cause was that the patient did not answer the questions where the values were missing.

Different approaches were taken to tackle missing data points for numerical and categorical features. The categorical missing values were imputed by the majority class of that specific feature. The missing values in the numerical features were imputed by the mean value of the existing values of that feature [255]. On the numerical columns, mean imputation was performed after the outliers had been removed, as mean imputation is sensitive to outliers.

Figure 3.2 and Figure 3.3 show the state of missing values for the raw DADOSD and TRICD datasets. This is a matrix visualization of the nullity of the datasets showing the positional information of the missing values.

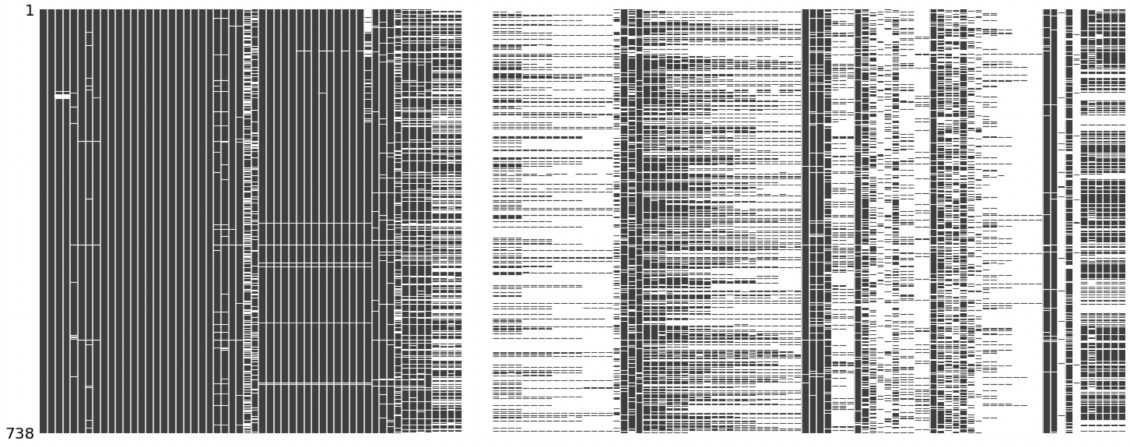


Figure 3.2: Missing values in raw DADOSD dataset. Every column represents a feature where white places indicate missing values.

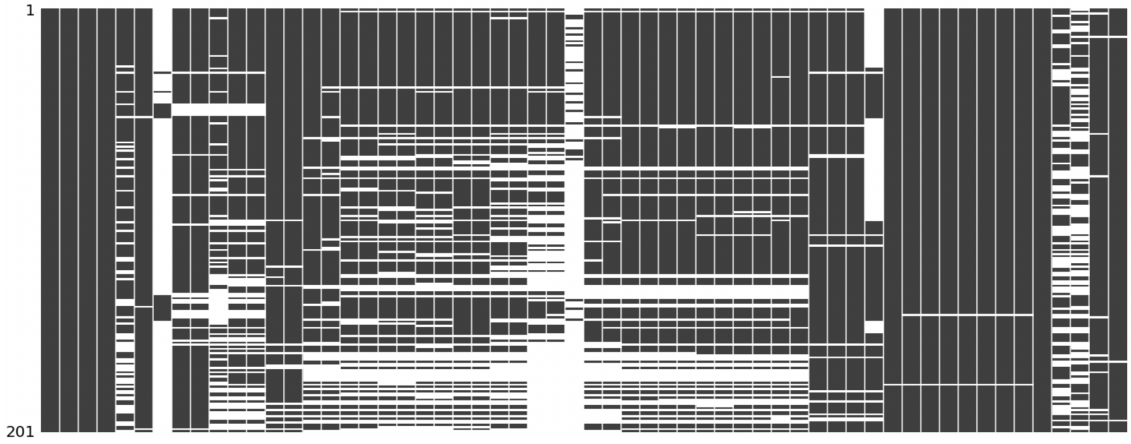


Figure 3.3: Missing values in raw TRICD dataset. Every column represents a feature where white places indicate missing values.

Three consecutive steps were taken to handle the missing data points.

1. Manual Fixes

2. Removal of the features with $\geq$40% missing values

3. Imputation (This step was taken after the outlier removal step discussed in Subsection 3.3.1)

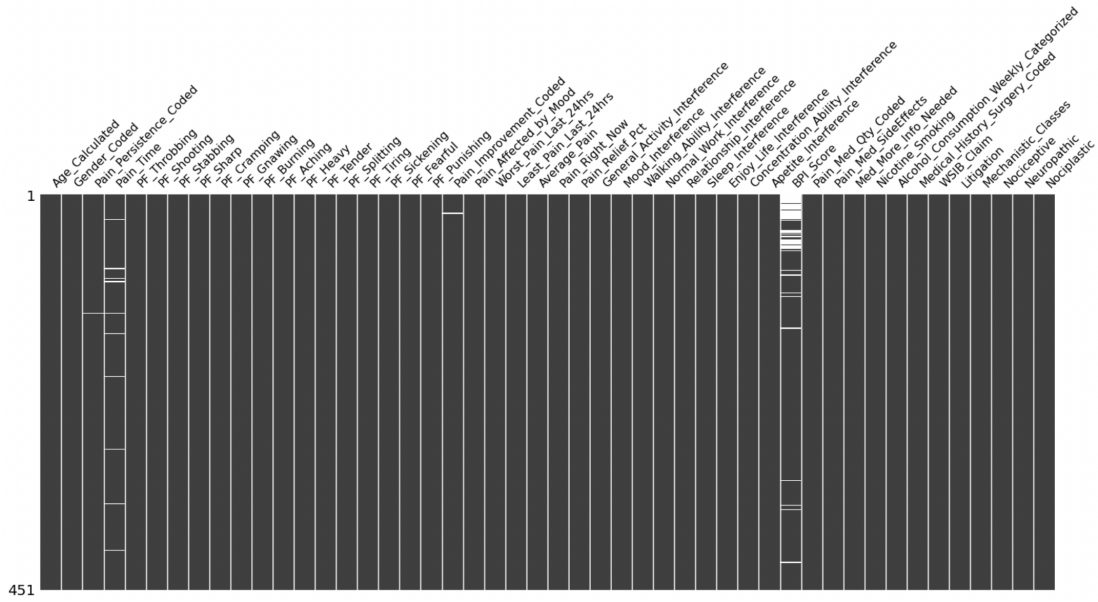


Figure 3.4: Missing values DADOSD dataset after outlier removal and before imputation. Every column represents a feature where white places indicate missing values.

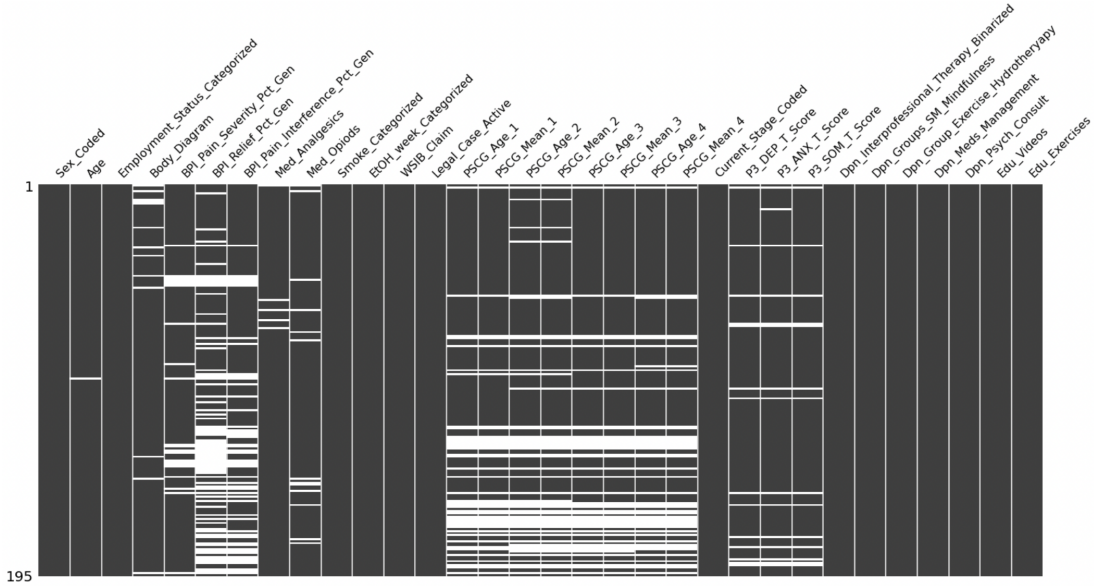


Figure 3.5: Missing values TRICD dataset after outlier removal and before imputation. Every column represents a feature where white places indicate missing values.

The first one was to fix the empty cells where they were intentionally left empty, e.g., empty cells in Nicotine_Smoking column in DADOSD data indicating the person does not smoke nicotine. For these types of cases, the empty cells were filled with appropriate values consulting with the responsible person at the UHN.

In the second step, features were dropped with $\geq$40% missing values after consulting with the domain experts, as data imputation for these features will not be helpful.

The resulting datasets still had empty cells or missing values which were dealt with after the outlier removal. Figure 3.4 and 3.5 show the state of missing values before imputation. In the third step, measures were taken with proper reasoning to impute the missing and removed data points (outliers).

3.3 Data Analysis

This section gives a summary of the raw and processed data. It also includes data visualization, data imputation, and feature scaling steps. Figure 3.6 shows the steps taken to explore and analyze the datasets after cleaning.

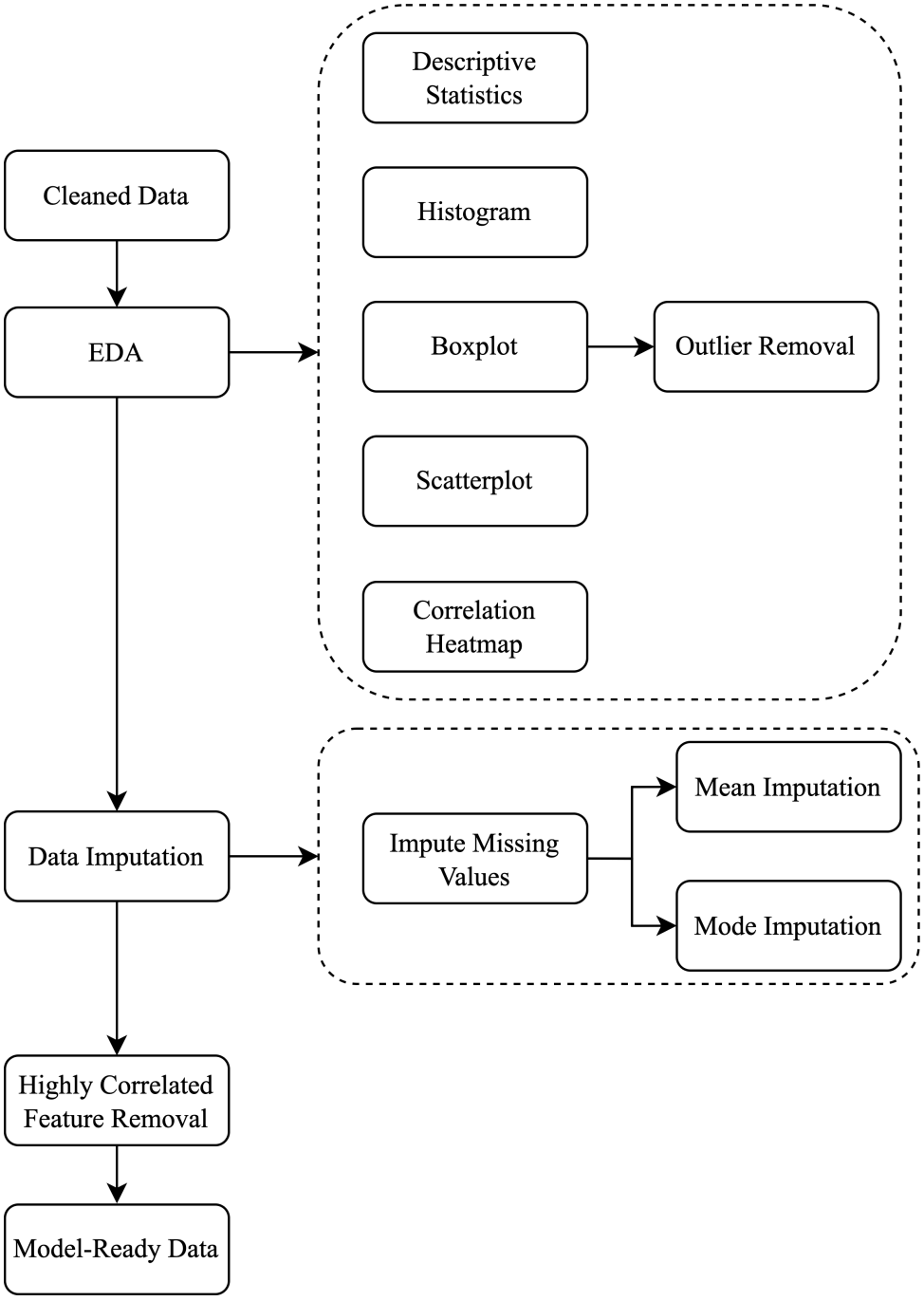


Figure 3.6: Cleaned data to Model-Ready data. These steps were taken to analyze the data and to make the cleaned datasets ready for AI/ML algorithms.

3.3.1 Exploratory Data Analysis

3.3.1.1 Descriptive Statistics

DADOSD The descriptive statistics for DADOSD dataset are given separately for numerical and categorical variables in Table 3.3 and Table 3.4 respectively.

Table 3.3: Descriptive statistics for numerical variables in DADOSD dataset (after outlier removal and imputation). Here, Count is the total number of values in the column and Std is standard deviation.

Variable Name	Count	Mean	Std	Min	25%	50%	75%	Max
Age_Calculated	451	50.89	15.38	17.62	39.86	50.74	61.32	91.78
BPI_Score	451	58.8	17.31	6	49.5	58.8	71	90

Table 3.4: Descriptive statistics for categorical variables in DADOSD dataset (after outlier removal and imputation). Here, Count is the total number of values in the column and Top indicates the category with the highest frequency.

Variable Name	Count	Unique	Top	Frequency
Gender_Coded	451	2	1	297
Pain_Persistence_Coded	451	2	1	340
Pain_Time	451	2	1	291
PF_Throbbing	451	2	1	262
PF_Shooting	451	2	1	232
PF_Stabbing	451	2	1	235
PF_Sharp	451	2	1	287
PF_Cramping	451	2	0	267

Continued on the next page

Table 3.4 continued from the previous page

Variable Name	Count	Unique	Top	Frequency
PF_Gnawing	451	2	0	308
PF_Burning	451	2	0	253
PF_Aching	451	2	1	347
PF_Heavy	451	2	0	282
PF_Tender	451	2	0	248
PF_Splitting	451	2	0	375
PF_Tiring	451	2	1	294
PF_Sickening	451	2	0	334
PF_Fearful	451	2	0	388
PF_Punishing	451	2	0	322
Pain_Improvement_Coded	451	3	2	235
Pain_Affected_by_Mood	451	2	1	229
Worst_Pain_Last_24hrs	451	11	8	114
Least_Pain_Last_24hrs	451	11	3	73
Average_Pain	451	11	7	107
Pain_Right_Now	451	11	7	89
Pain_Relief_Pct	451	11	0	76
General_Activity_Interference	451	11	7	84
Mood_Interference	451	11	8	79
Walking_Ability_Interference	451	11	10	67
Normal_Work_Interference	451	11	10	118

Continued on the next page

Table 3.4 continued from the previous page

Variable Name	Count	Unique	Top	Frequency
Relationship_Interference	451	11	8	68
Sleep_Interference	451	11	10	85
Enjoy_Life_Interference	451	11	10	114
Concentration_Ability_Interference	451	11	8	84
Apetite_Interference	451	11	0	94
Pain_Med_Qty_Coded	451	8	1	113
Pain_Med_SideEffects	451	3	0	208
Med_More_Info_Needed	451	2	0	288
Nicotine_Smoking	451	2	0	357
Alcohol_Consumption_Weekly_Categorized	451	4	0	252
Medical_History_Surgery_Coded	451	2	0	273
WSIB_Claim	451	2	0	432
Litigation	451	2	0	387

The descriptive statistics of the mechanistic classes or target variables are provided in Table 3.5. It is evident that the dataset is imbalanced, while 'Nociceptive (0)' being the majority and 'Neuropathic (1)' being the minority class.

Table 3.5: Descriptive statistics for mechanistic classes in DADOSD dataset. Here, Count is the total number of values in the column and Top indicates the category with the highest frequency.

Variable Name	Count	Unique	Top	Frequency
Mechanistic_Classes	451	4	0	180
Nociceptive	451	2	1	263
Neuropathic	451	2	0	318
Nociplastic	451	2	0	297

TRICD Similarly, the descriptive statistics for TRICD dataset are given separately for numerical and categorical variables in Table 3.6 and Table 3.7 respectively.

Table 3.6: Descriptive statistics for numerical variables in TRICD dataset (after outlier removal and imputation). Here, Count is the total number of values in the column and Std is standard deviation.

Variable Name	Count	Mean	Std	Min	25%	50%	75%	Max
Age	195	52.42	16.05	19	41	52	61	89
Body_Diagram	195	15.8	9.78	1.25	8	15	21.75	47
BPI_Pain_Severity_Pct_Gen	195	60.58	20.09	10	50	60.58	75	100
BPI_Relief_Pct_Gen	195	48.61	18.44	0	48.61	48.61	50	90
BPI_Pain_Interference_Pct_Gen	195	59.97	19.9	0	55.64	59.97	70.56	100
Med_Analgesics	195	1.52	1.46	0	0	1	2	6

Continued on the next page

Table 3.6 continued from the previous page

Variable Name	Count	Mean	Std	Min	25%	50%	75%	Max
Med_Opiods	195	0.6	0.72	0	0	0	1	2
PSCG_Age_1	195	63.16	16.2	23	54	63.16	71	100
PSCG_Mean_1	195	3.16	0.81	1.1	2.71	3.16	3.6	5
PSCG_Age_2	195	68.37	11.5	32	62	68.37	75.5	98
PSCG_Mean_2	195	3.41	0.57	1.6	3.1	3.41	3.7	4.9
PSCG_Age_3	195	58.42	15.42	17	53	58.42	67	100
PSCG_Mean_3	195	2.92	0.77	1	2.7	2.92	3.3	5
PSCG_Age_4	195	58.61	15.11	17	51	58.61	69	100
PSCG_Mean_4	195	2.94	0.76	0.9	2.65	2.94	3.43	5
P3_DEP_T_Score	195	48.08	8.98	31	41	48.08	54	70
P3_ANX_T_Score	195	45.08	8.87	31	38.5	45.08	51	70
P3_SOM_T_Score	195	46.84	8.85	25	42	46.84	53	67

Table 3.7: Descriptive statistics for categorical variables in TRICD dataset (after outlier removal and imputation). Here, Count is the total number of values in the column and Top indicates the category with the highest frequency.

Variable Name	Count	Unique	Top	Frequency
Sex_Coded	195	2	1	125
Employment_Status_Categorized	195	4	3	54
Smoke_Categorized	195	3	0	121
EtOH_week_Categorized	195	3	0	81

Continued on the next page

Table 3.7 continued from the previous page

Variable Name	Count	Unique	Top	Frequency
WSIB_Claim	195	2	0	187
Legal_Case_Active	195	2	0	178
Current_Stage_Coded	195	8	1	57
Dpn_Interprofessional_Therapy_Binarized	195	2	0	152
Dpn_Groups_SM_Mindfulness	195	2	0	139
Dpn_Group_Exercise_Hydrotherapy	195	2	0	149
Dpn_Meds_Management	195	2	1	138
Dpn_Psych_Consult	195	2	0	188
Edu_Videos	195	2	0	109
Edu_Exercises	195	2	1	126

Unlike the DADOSD dataset, the mechanistic class or target variables are not available in the TRICD dataset. Therefore, it remains undetermined if the dataset is imbalanced or not.

3.3.1.2 Visualization and Analysis

Histogram and Barchart The histogram for the numerical features and the barchart for the categorical features are provided together. The categorical features can be distinguished from the gaps between the rectangles.

DADOSD Data The histogram and barchart of the cleaned DADOSD data before outlier removal and missing data imputation is shown in Figure 3.7.

Figure 3.8 shows the histogram and barchart of the DADOSD dataset after outlier removal and missing value imputation. No significant changes in the distribution of the variables were observed compared to Figure 3.7.

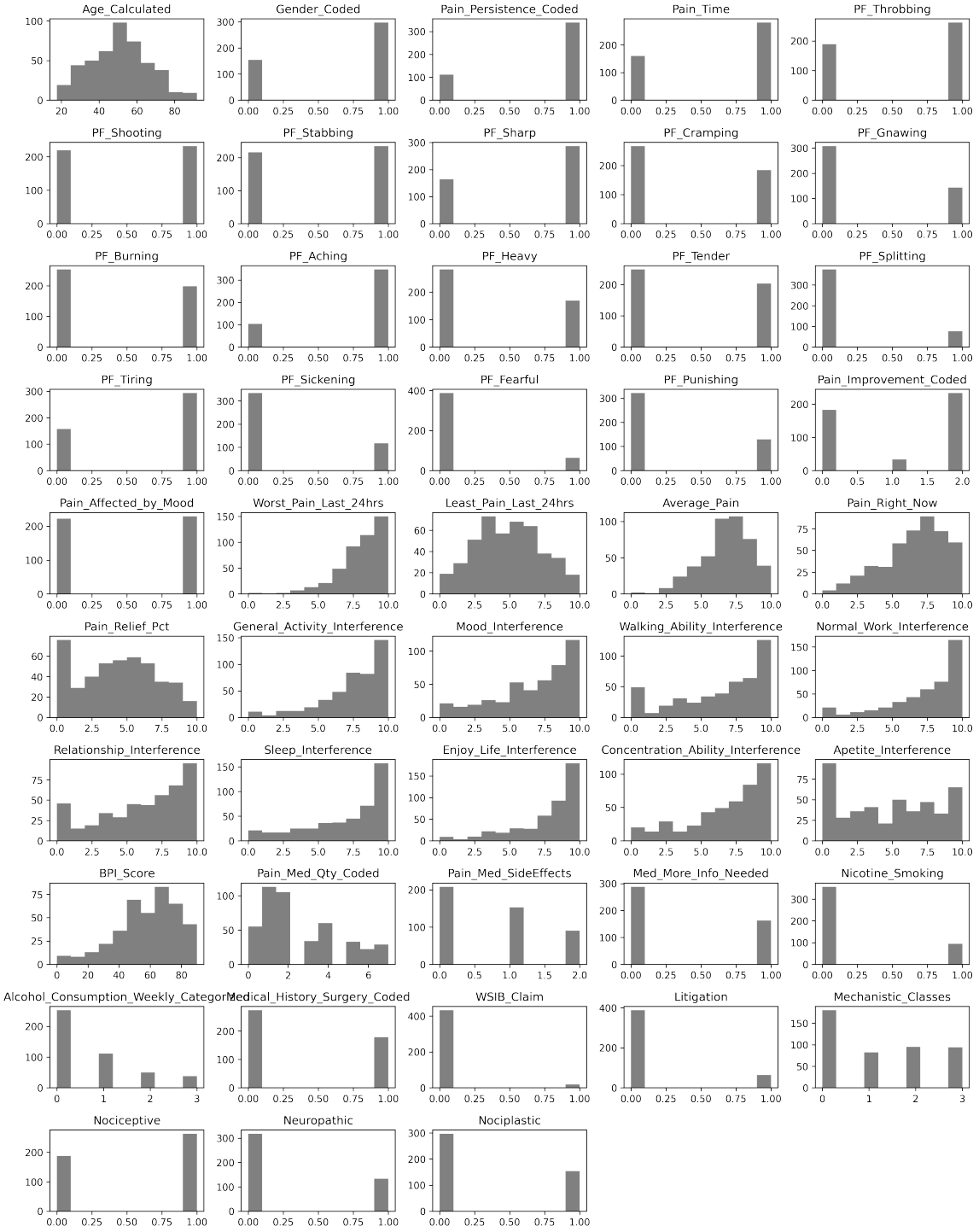


Figure 3.7: Histogram and barchart of the cleaned DADOSD dataset before outlier removal

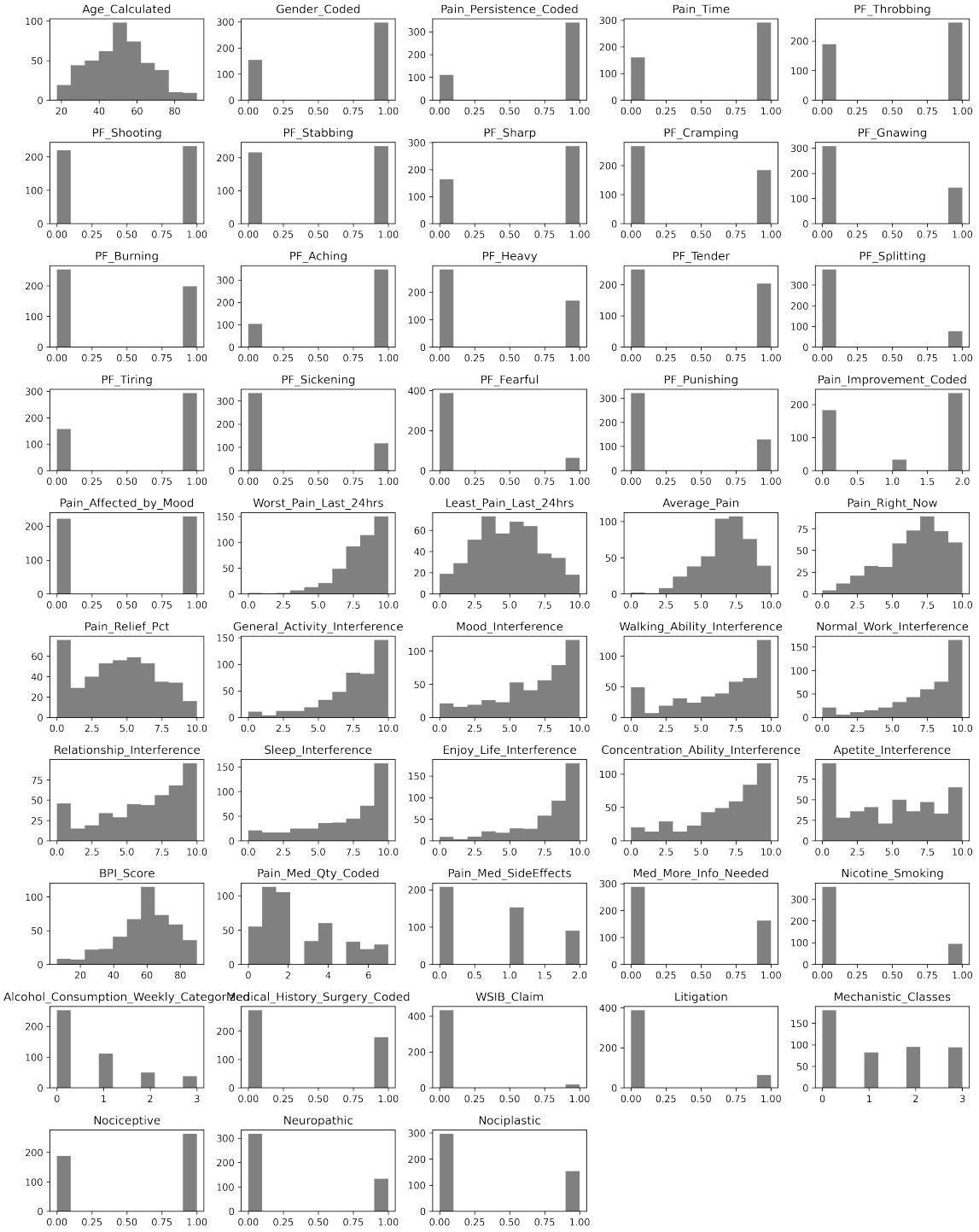


Figure 3.8: Histogram and barchart of the DADOSD dataset after outlier removal

TRICD Data The histogram and barchart of the cleaned TRICD data before outlier removal and missing value imputation is shown in Figure 3.9.

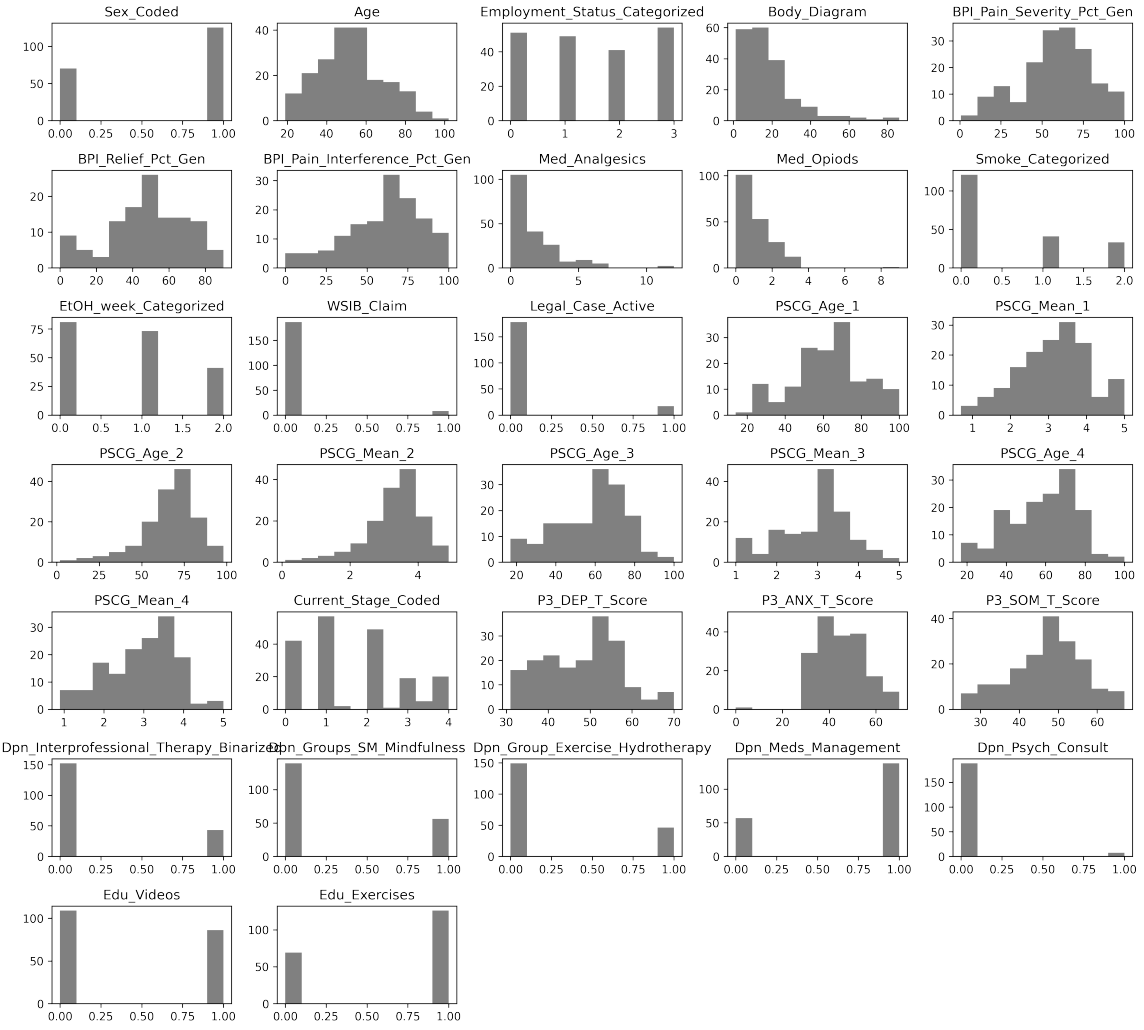


Figure 3.9: Histogram and barchart of the TRICD dataset before outlier removal

Figure 3.10 shows the histogram and barchart of the TRICD dataset after outlier removal and missing value imputation. Similar to the DADOSD dataset, no significant changes in the distribution of the variables were observed compared to Figure 3.9.

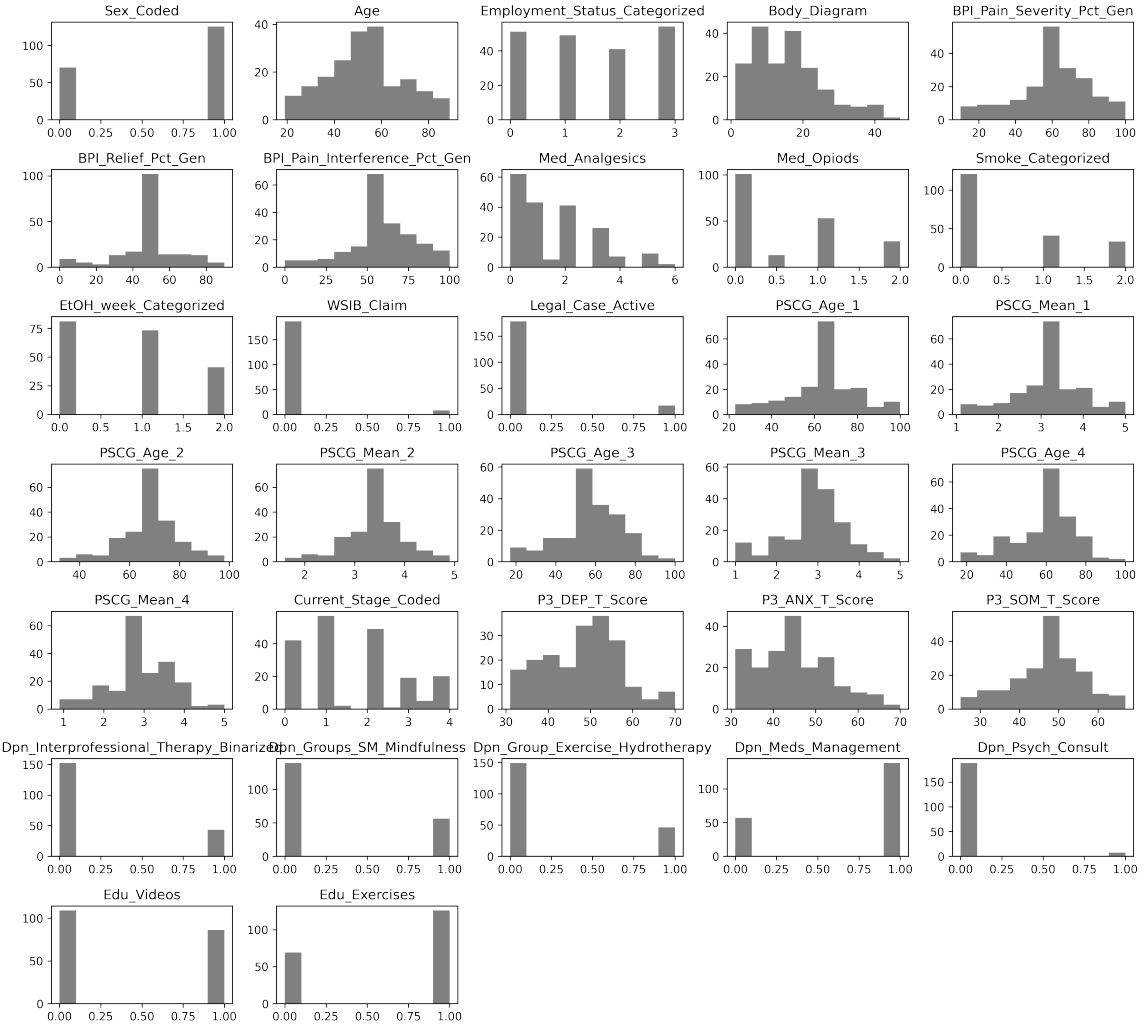


Figure 3.10: Histogram and barchart of the TRICD dataset after outlier removal

3.3.1.3 Outlier Detection and Removal

As the numerical features present in the datasets are mostly normally distributed (graphically assessed from the histograms), interquartile range (IQR) was used to identify outliers from the density distribution of the numerical features, which is visualized using Boxplot.

DADOSD Data Figure 3.11 shows outliers in diamond-shaped dots. The outliers or data points beyond interquartile range for 'BPI_Score' were removed.

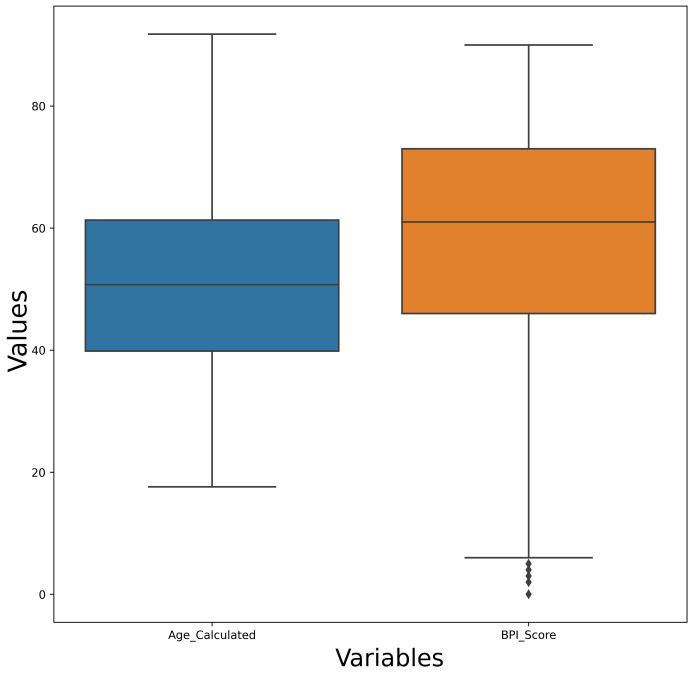


Figure 3.11: Boxplot to visualize outliers in DADOSD dataset. Here the outliers are indicated by diamond-shaped dots.

TRICD Data Similarly, Figure 3.12 shows outliers in diamond-shaped dots for TRICD Dataset. The outliers or data points beyond interquartile range were removed for all the applicable features.

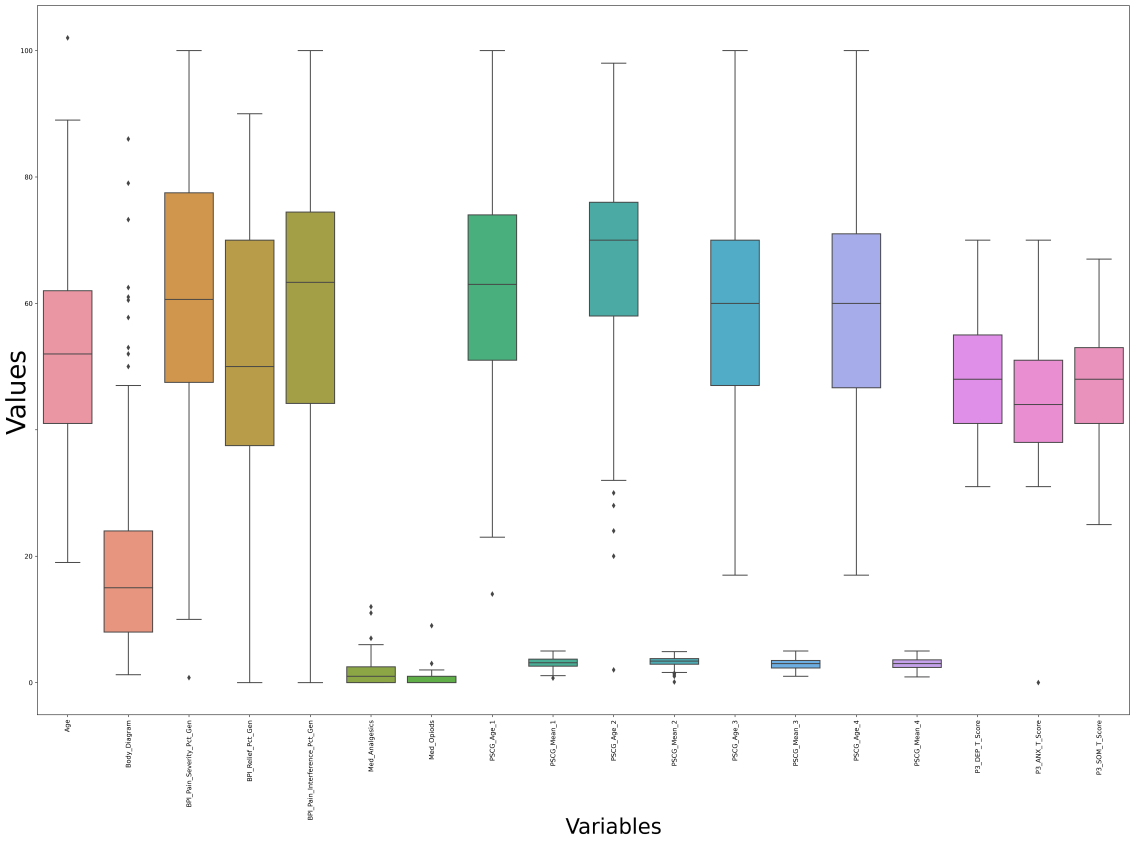


Figure 3.12: Boxplot to visualize outliers in TRICD dataset. Here the outliers are indicated by diamond-shaped dots.

3.3.1.4 Data Imputation

The missing values were imputed using mean and mode values for the numerical and categorical columns, respectively. As both imputation techniques are sensitive to outliers, they were performed after the outlier removal.

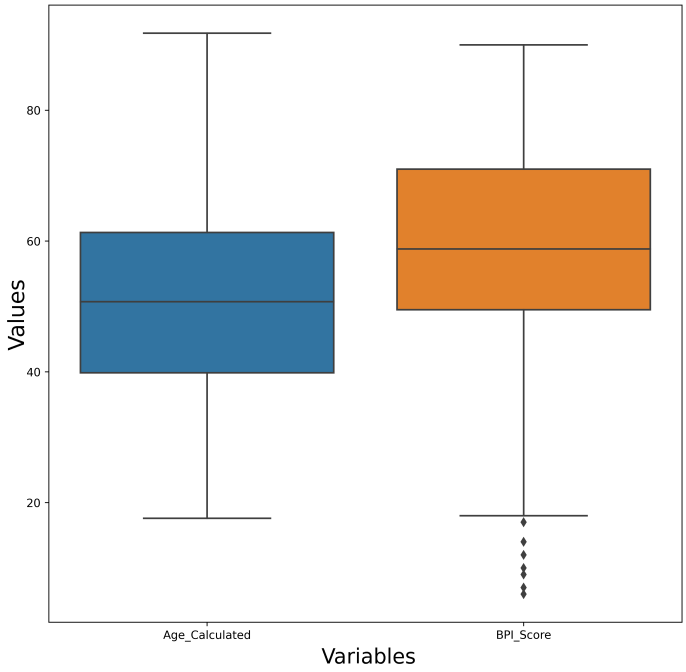


Figure 3.13: Boxplot after outlier removal and imputation in DADOSD dataset. Mean imputation was done for numerical variables.

DADOSD Data Figure 3.13 shows the distribution of the numerical features after outlier removal and data imputation.

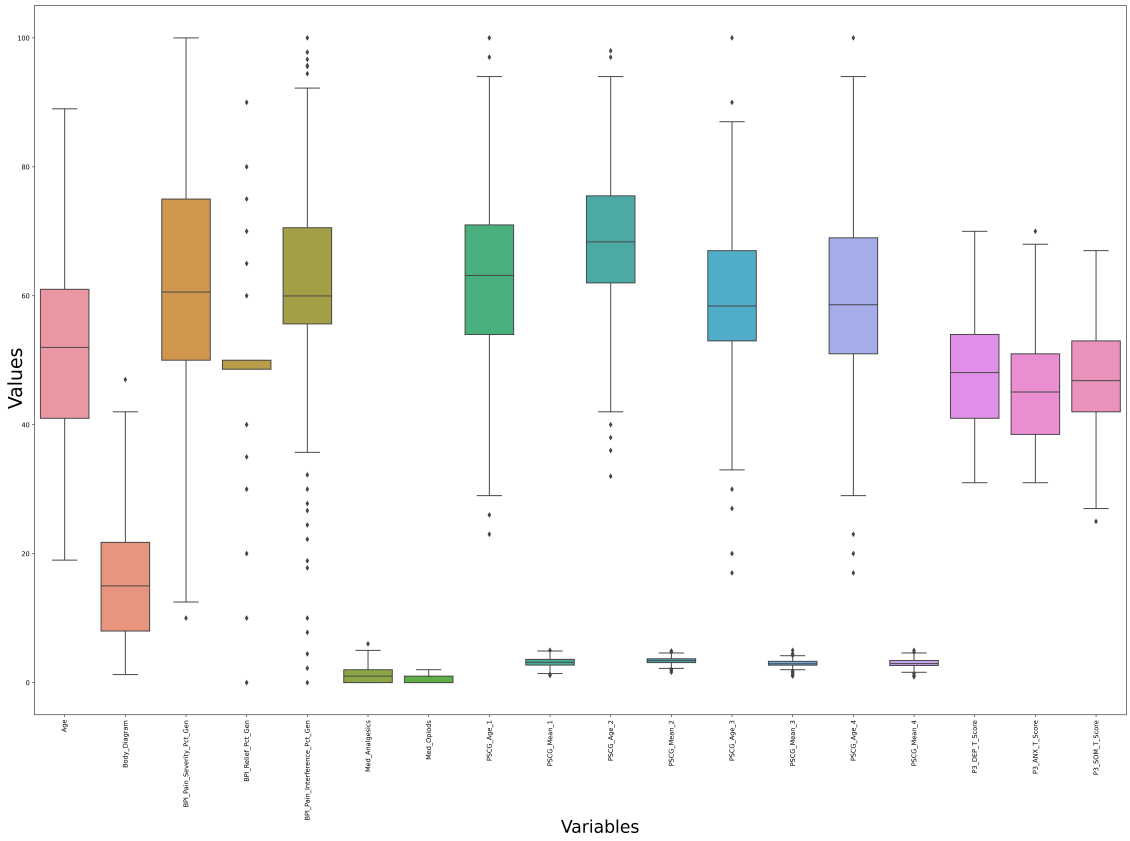


Figure 3.14: Boxplot after outlier removal and imputation in TRICD dataset. Mean imputation was done for numerical variables respectively.

TRICD Data Figure 3.13 shows the distribution of the numerical features after outlier removal and data imputation. From the figures, it can be observed that new outliers have appeared. It is because the distribution after removing outliers is not exactly the same as before, and it renders its own outliers.

However, for the categorical variables, Mode imputation was performed for both datasets.

Correlation Heatmap Pairwise correlations have been computed for the features in the datasets to remove redundant features. The correlation between each pair of

variables can be found in Appendix C.2 for both datasets.

DADOSD No strong correlation was found between the variables.

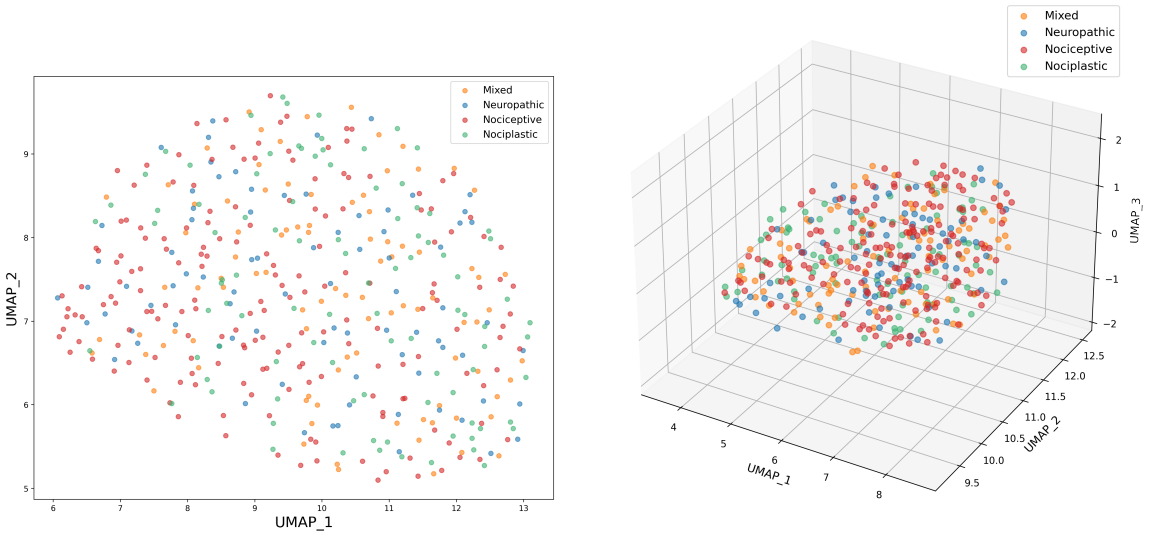
TRICD Highly correlated columns were removed (i.e., PSCG_Mean_1, PSCG_Mean_2, PSCG_Mean_3, PSCG_Mean_4).

3.3.1.5 Feature Scaling

Feature scaling was performed to reduce the chances of bias toward a particular feature that had values higher in magnitude. The numerical features were scaled using standard scalar as it is robust to outliers. Only for the AE, the min-max scalar was used, as ANN tends to do well when the features are normalized.

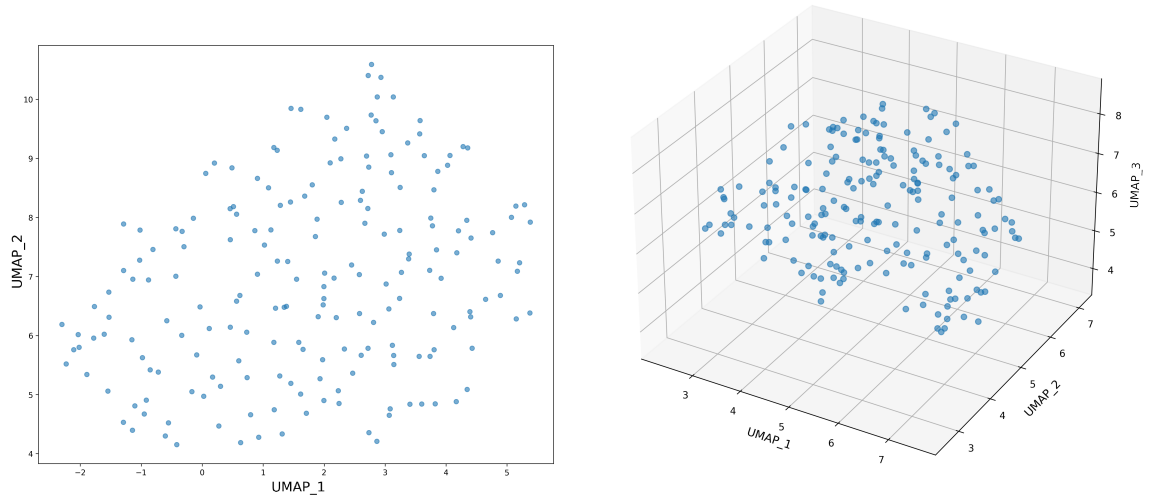
3.3.2 Data Visualization

UMAP was used to reduce the dimension to 2 and 3 to show the datapoints. For the DADOSD dataset, the datapoints were colored as per the true labels in the dataset. As the TRICD dataset does not have true labels, the datapoints were not colored.



(a) Two-dimensional scatterplot (b) Three-dimensional scatterplot

Figure 3.15: The datapoints in DADOSD dataset are visualized as a scatterplot using UMAP. The datapoints are colored based on CP mechanism labels present in the dataset.



(a) Two-dimensional scatterplot (b) Three-dimensional scatterplot

Figure 3.16: The datapoints in TRICD dataset are visualized as a scatterplot using UMAP. The datapoints are single-colored, as CP mechanism labels are not present in the dataset.

In Figure 3.15, it can be observed that the datapoints are not discernible as per their true classes based on their local and global structures. The axes in these figures represent UMAP embedding dimensions.

3.4 Secondary Use of Data

This work only involved secondary use of data, and the data collection was not originally intended for this thesis. The original purpose of the data collection was for the use of clinical staff, i.e., medical doctors, nurses, chiropractors, physiotherapists, and pharmacists. Details about the ethics approval, and data storage & retention can be found in Appendix B.1.

3.5 Availability of the Datasets

The datasets are not publicly available. It might be available on request from UHN.

Chapter 4

Machine Learning Model

This chapter focuses on the ML/AI methods and their developments. It also provides relevant information about why the algorithm or the architecture was used for this work. This is to be noted that Model-Ready datasets (Figure 3.6) were used in this and the following sections.

4.1 k-prototypes

k-prototypes has demonstrated exemplary performance in clustering mixed-type or heterogeneous data. For this reason, it has been used to cluster the datasets [97].

k-prototypes only accepts floating-point values for numerical data and integer values for categorical data. After scaling the numerical features, the data were fed to the k-prototypes algorithm. k-prototypes requires numerical and categorical data to be provided in separated arrays. Thus, the indexes of the categorical variables were given to the algorithm. The initialization of centroids was done using the centroid initialization function proposed by 'Huang'. The distance function used for numerical

features was euclidean, and the similarity measure used for categorical features was matching distance. The weighing factor 'gamma' determines the relative importance of numerical vs. categorical variables cite Huang [1997]. It was automatically calculated from the data.

To find the optimal number of clusters, elbow method was used. For cluster numbers 2 to 10 the cost was calculated to identify the optimal number of clusters. Additionally, average silhouette scores were calculated for the cluster numbers 2 to 10 to see where the highest value indicates optimal cluster numbers.

The model was trained using the entire dataset with all the features after preprocessing and scaling (Model-Ready data). For a given number of clusters, k-prototypes computed cluster centroids and predicted cluster index for each sample in the dataset.

4.2 Semi-supervised Learning (SVM)

As the datasets are relatively small in size, the Semi-supervised SVM (i.e., SVC) was applied. The other two Semi-supervised models were used to compare the results with the Semi-supervised SVM's result.

Semi-supervised Learning was applied using the self-training mechanism. SVC was used to function as a Semi-supervised classifier using self-training, allowing it to learn from unlabeled data. It iteratively predicts pseudo-labels for the unlabeled data and adds them to the training set. The classifier continues iterating until either maximum iteration is reached or no pseudo-labels are added to the training set in the previous iteration.

The dataset was split into a train and test set where 80% data were randomly selected and put into training set, and the rest were put into test set (hold-out). The

training set was again split into two parts where 30% were randomly selected, and true labels were kept. For the rest 70% data, the true labels were removed and replaced with -1 as an identifier of unlabeled examples. Therefore, unlabeled points along with the labeled data were provided to train the model. The number of examples in each subset can be found in Table 4.1.

Table 4.1: Model training and test split (DADOSD dataset)

	Training Set	**Test Set**	**Split**
Total	360	91	80/20
Unlabeled	120		70/30
Labeled	240		

In each iteration, the base classifier (SVC) predicts labels for the unlabeled examples and adds a subset of these labels to the labeled dataset. The selection criterion determines this subset. This selection is done by choosing the $k = 10$ best samples from the prediction probabilities. The algorithm iterates and predicts labels until all samples have labels or no new samples are selected in that iteration, or the maximum number of iterations is reached.

For the DADOSD dataset, regular 4 classes (mechanistic classes, i.e., Neuropathic, Nociceptive, and Nociplastic, including mixed CP as a separate class where two or more conditions are present) were used for model training and testing. For the TRICD dataset, Semi-supervised Learning could not be explored due to the absence of labels.

4.3 FCM

Clustering algorithm performs better in lower dimensional space. Due to the ability to bring down the high-dimensional data into a lower-dimensional space by extracting latent features, AE was used as a feature extraction technique for the FCM algorithm.

To check the overlapping tendency of the clusters, FCM was used as it can handle overlapping clusters by indicating the probability of cluster memberships for every data point.

The FCM algorithm was implemented using the features extracted by AE.

4.3.1 Feature Extraction

The overall AE architecture had an encoder and a decoder component to it. The bottleneck in the middle of the encoder and decoder ensured only the core structured part of the information could go through from where the original data could be reconstructed. In other words, the high-dimensional data were fed to a ANN with a narrow bottleneck layer in the middle containing the latent representation of the input features.

The encoder took the Model-Ready data as input and put it through a few hidden layers with a LeakyReLU activation function. For the decoder, the architecture was the same as the encoder but reversed. The decoder took the bottleneck as input and output reconstructed inputs. In the first hidden layer of the encoder and the decoder, $L2$ regularization was applied to the output of the activation function (LeakyRELU) during optimization.

A sigmoid activation function was used at the bottleneck layers, where it outputs another value between 0 and 1. Sigmoid is non-linear, continuously differentiable, and

has a fixed output range. The goal was to search for an encoder and decoder that minimizes the reconstruction error done by gradient descent over the parameters of the network.

The Model-Ready data were split into a train and validation set (80/20 random split). It was the same split used for the Semi-supervised Learning algorithms (DADOSD dataset). However, the test set was used as validation data while training the model. Similar to the DADOSD dataset, the TRICD dataset was also split into training and validation set. Table 4.2 holds the information of the training and validation set information for both datasets.

Table 4.2: AE training and validation split

Dataset	**Training Set**	**Validation Set**	**Split**
DADOSD	360	91	80/20
TRICD	156	39	80/20

During training, at each iteration, the AE architecture (encoder followed by the decoder) with batches of data and compared the encoded-decoded output with the initial data and backpropagated the error through the architecture to update the weights of the network.

No class information was provided as AE is unsupervised and does not need class information to train its parameters. However, the model was fit using the Adam version of stochastic gradient descent while it minimized the mean squared error (MSE), given the reconstruction of the data, which was a type of multi-output regression problem. Early stopping was used to stop training when MSE stopped improving.

4.3.2 FCM Implementation

The FCM needs the apriori specification of the number of clusters. The number of clusters (e.g., $k = 3$) was given to it as there are three mechanistic classes of CP. All the latent representations of the features were fed to the FCM model. As FCM has difficulties finding optimal cluster centers, k-Means++ was used to find the cluster centers. FCM was then trained with k-Means++ center initialization. However, FCM updated cluster membership and centers in each initialization while training.

With the fitted model, the same extracted features were given to predict the membership probabilities for all the clusters for all the instances (soft clustering). FCM outputs the cluster centers and can also provide us labels like the usual/regular clustering methods by putting an instance in the cluster where it has the highest probability of belonging (used for visualization: Figure 6.4).

Chapter 5

Methodology

The focus of this chapter is the application of the ML/AI models discussed in the previous chapter (Chapter 4). In this chapter, how the models were applied and evaluated are presented.

5.1 Identifying Distinguishable Clusters

At first, clustering was performed to check if CP data reveal distinguishable and clinically meaningful clusters. Therefore, k-prototypes algorithm was applied to the Model-Ready data. Figure 5.1 shows how clusters were applied to Model-Ready data.

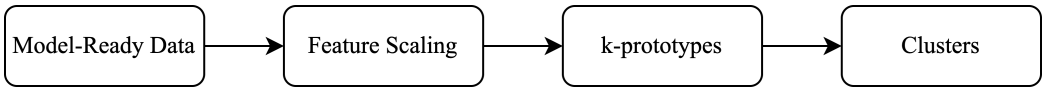
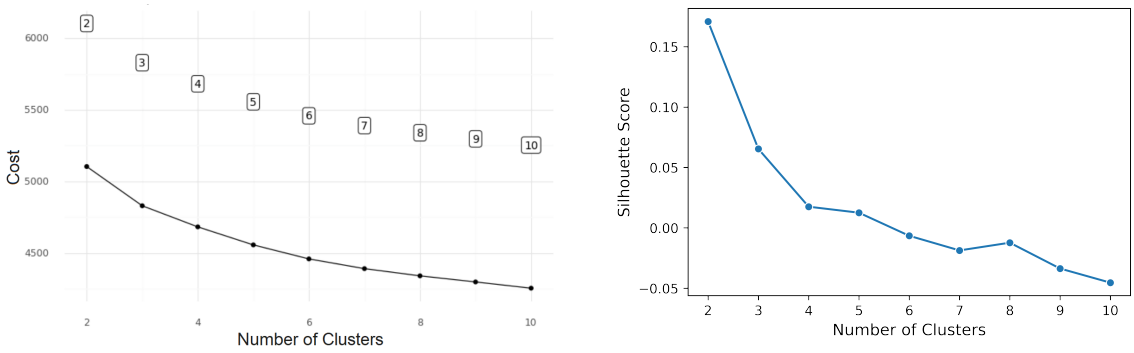


Figure 5.1: Applying k-prototypes to the Model-Ready data

5.1.1 Finding the Optimal Number of Clusters

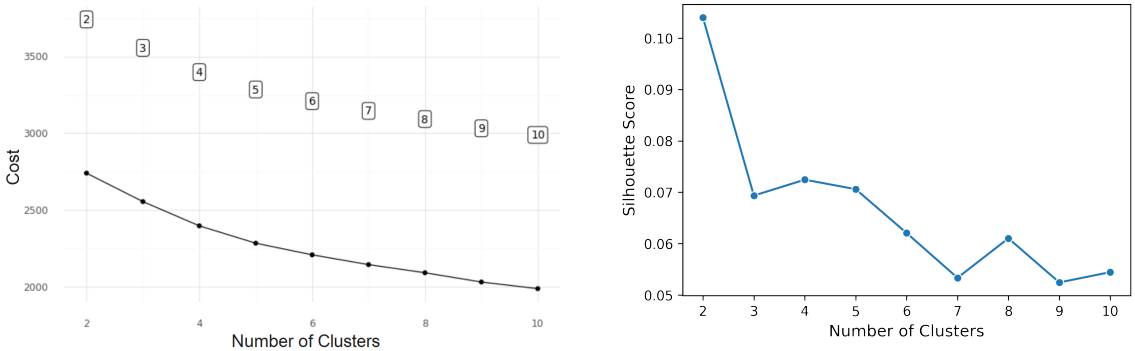
To find the optimal number of clusters, elbow plot and silhouette score were calculated.



(a) Elbow plot (b) Average silhouette score

Figure 5.2: Optimal number of clusters in DADOSD dataset using elbow method and average silhouette score. A clear hinge or elbow is expected to elect the optimal number of clusters using the elbow method. On the other hand, a prominent peak is sought which indicates a high average silhouette score indicates better-defined clusters.

For the DADOSD dataset, Figure 5.2a shows the elbow plot for 2 to 10 clusters, and Figure 5.2b shows the average silhouette scores for clusters 2 to 10.



(a) Elbow plot (b) Average silhouette score

Figure 5.3: Optimal number of clusters in TRICD dataset using elbow method and average silhouette score. A clear hinge or elbow is expected to elect the optimal number of clusters using the elbow method. On the other hand, a prominent peak is sought which indicates a high average silhouette score indicates better-defined clusters.

For the TRICD dataset, Figure 5.3a shows the elbow plot for 2 to 10 clusters, and

Figure 5.3b shows the average silhouette scores for clusters 2 to 10.

5.1.2 Cluster Validation

Though there is no indication of the optimum number of clusters from the elbow plot or the silhouette score, cluster 3 and 4 are explored. The number of clusters 3 was tried with the notion of 3 CP mechanisms or categories where 4 clusters were also tried to see if mixed pain itself makes a cluster or not.

Performance Evaluation The clustering result was assessed using Rand Index (RI), Adjusted Rand Index (ARI), (Adjusted Mutual Information (AMI), homogeneity and completeness, Fowlkes-Mallows Index (FMI), silhouette coefficient, and Davies-Bouldin Index (DBI). However, for the TRICD dataset, only the silhouette coefficient and DBI could be calculated in the absence of true labels.

5.2 Semi-supervised Learning in Identifying CP Mechanisms

Semi-supervised Learning was applied to the Model-Ready data after scaling (standard scaler) the numerical features. It was performed to inspect if some information about the true classes can help the algorithm to identify the CP mechanisms.

Two types of Semi-supervised methods were used. Figure 5.4 shows how the algorithms were applied to the datasets to compare the results.

First, SVC was trained using the dataset to tune the model parameters. To do so, the data was split into training set and test set (80-20 split, details were provided in

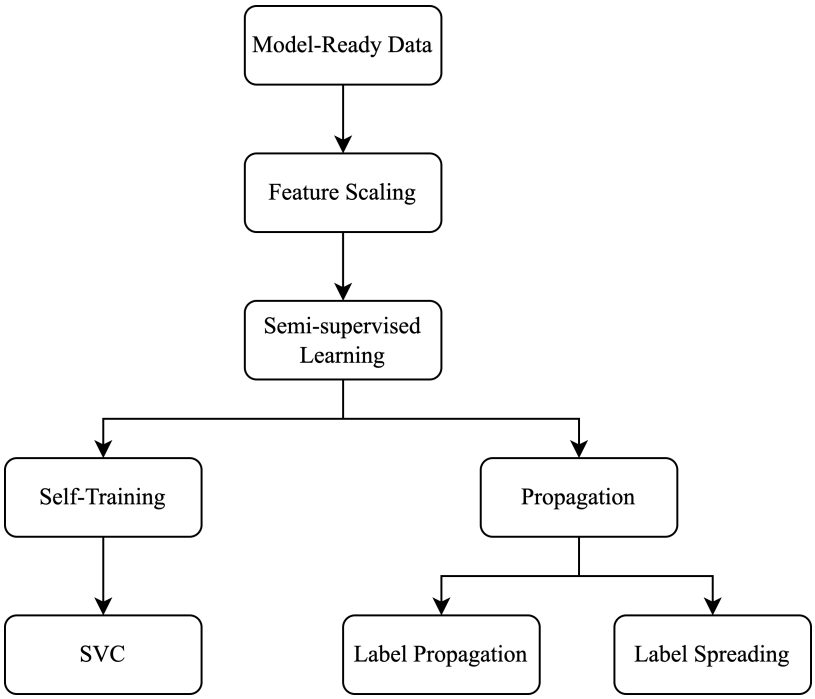
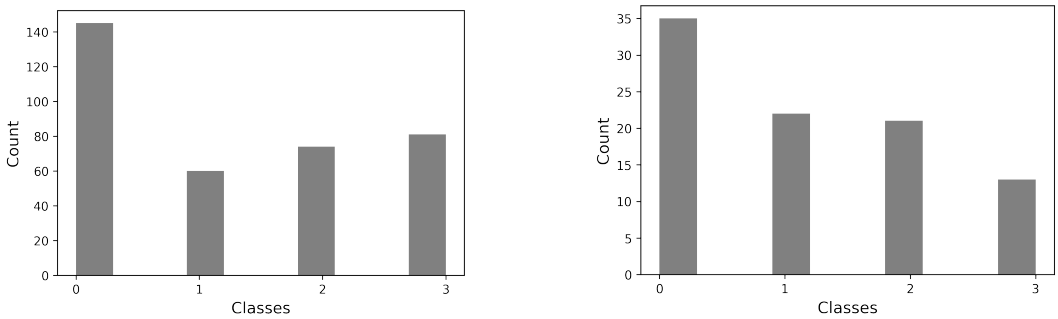


Figure 5.4: Applying Semi-supervised Learning to the Model-Ready data

Table 4.1). The parameters were tuned using a grid search generating the best result. After that, the tuned parameters were provided to the SVC for the training using the self-training mechanism.

For the propagation techniques, Label Propagation and Label Spreading were applied to the Model-Ready data after feature scaling (standard scalar). The same train and test sets were used for these algorithms as well. Similarly, the same set of unlabeled data was used for propagating the labels during the training of both algorithms.



(a) Training set (b) Hold-out test set

Figure 5.5: Class distribution in training and test set (DADOSD dataset)

For the DADOSD dataset, 4 classes were present where mixed pain was an individual class. However, for the TRICD dataset, Semi-supervised algorithms could not be applied due to the lack of true labels (mechanistic classes).

Performance Evaluation The models were tested by predicting the hold-out test set separated at the first train-test split. The model performances were reported as accuracy, precision, recall, and F1-score for all the classes along with aggregated scores.

Though training accuracies are reported, they are not very useful other than indicating model overfitting or underfitting, especially in this case of an imbalanced multi-class classification problem. Figure 5.5 shows the distribution of the classes in the training and test set. However, while calculating training accuracy, only the unlabeled examples (70% of the training dataset) were considered.

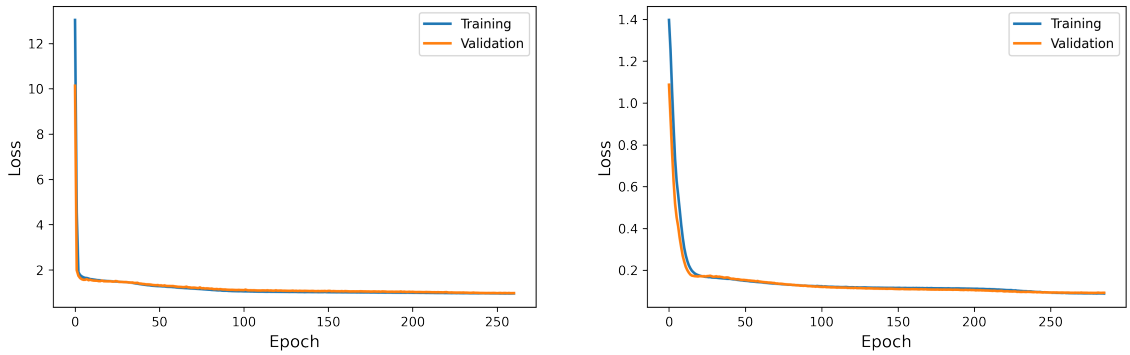
5.3 Unsupervised Learning in Identifying CP Mechanisms

AE was used for feature extraction, while FCM was used to cluster the features and find the overlapping clusters. A generalized diagram of the approach is given in Figure 5.7.

AE was employed to extract the latent data features while preserving the ability to reconstruct the data from the encoded latent features (bottleneck). Model-ready data was scaled (min-max scaler) and split into training and validation sets to feed the AE model (encoder-decoder). Details about the split were provided in Table 4.2. However, the AE model was validated using the validation set while calculating the reconstruction error in terms of MSE.

Figure 5.6. From the plot of the loss during training, it can be observed that the model has comparable performance on both train and validation sets. Early stopping was used to stop the training when no validation loss improvement in terms of MSE was observed.

For the DADOSD data, the best training loss was found 0.957, and the validation loss was 0.973. For the TRICD data, the best training loss was found 0.0892, and the validation loss was 0.089.



(a) AE training and validation on DADOSD Dataset

(b) AE training and validation on TRICD Dataset

Figure 5.6: AE model loss on the training and validation datasets over training epochs

Then the entire dataset was given to the Encoder of the trained AE to get the latent representation of the input data or the encoded features. The extracted features were then fed to the FCM algorithm. The number of clusters set for the FCM was 3 from the notion of 3 pain mechanisms. FCM then tries to put similar data points into clusters while giving the 'probability of belonging to' for all the 3 clusters.

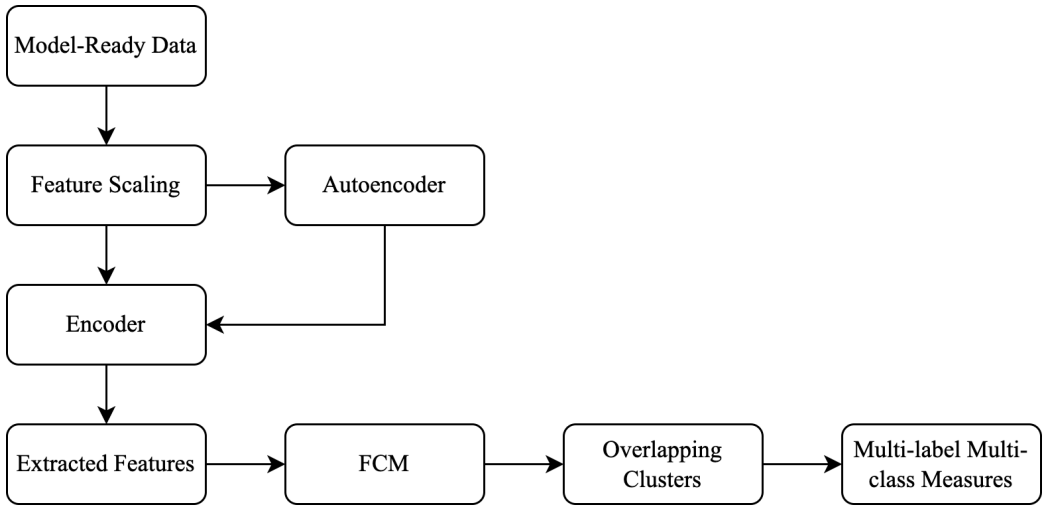


Figure 5.7: Applying AE and FCM to the Model-Ready data

As FCM is unsupervised, the cluster it generates does not have labels aligned with

true labels. Additionally, the values in each cluster hold the probability of belonging to that cluster for the instances. These probabilities were converted to binary based on a threshold, i.e., a probability value above the threshold became 1 and 0 otherwise. It was done to consider and evaluate the clustering result as a multi-label problem and compare them with one-hot encoded true labels (DADOSD dataset).

As the alignment with the true classes is not known, hamming loss was utilized to find the alignment. For all the six possible orders or combinations of the classes (Appendix C.4), the hamming loss was calculated. The lowest hamming loss indicated the best alignment indicating the cluster represents a particular pain type. Then the clusters were renamed accordingly. In this way, the result was also evaluated using MLCM by comparing them with the true labels using weighted average accuracy, precision, recall, and F1-score.

5.3.1 DADOSD Dataset

For the DADOSD data, the AE has an input layer with a dimension of 44 matching the Model-Ready dataset dimension. Figure 5.8 depicts the architecture of the AE. It has three hidden layers with a dimension of 22, 11, and 6. The bottleneck layer has 3 dimensions. The decoder has the same architecture as the encoder but is reversed.

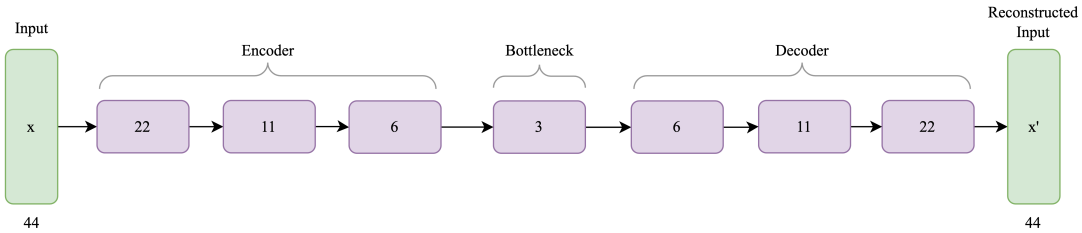


Figure 5.8: AE architecture for the DADOSD dataset

After training the dataset, the encoder was used to extract the features for use in FCM. Then the extracted features were used in FCM to identify clusters.

5.3.2 TRICD Dataset

For the TRICD data, the AE has an input layer with a dimension of 28 matching the Model-Ready dataset dimension. Figure 5.9 depicts the architecture of the AE. It has two hidden layers with a dimension of 14, and 7. The bottleneck layer has 3 dimensions. Similar to the AE for DADOSD dataset, the decoder has the same architecture as the encoder but is reversed.

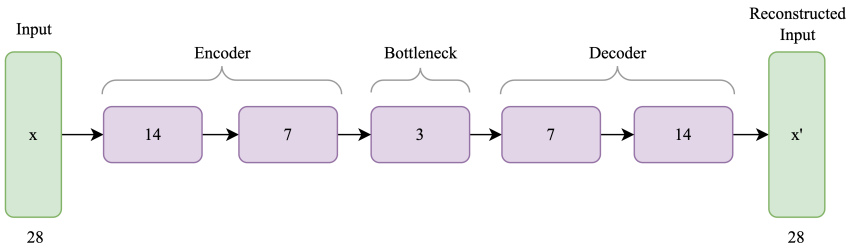


Figure 5.9: AE architecture for the TRICD dataset

After training the dataset, the encoder was used to extract the features for using in FCM. Then the extracted features were used in FCM to identify clusters.

5.3.3 Performance Evaluation

As FCM is unsupervised, the cluster it generates does not have labels aligned with true labels. To evaluate the performance of the FCM model in terms of multiple labels, the proper alignment of the clusters with true labels is necessary. However, the probabilities need to be binarized to be able to evaluate the performance according to the true labels (one-hot encoded) using hamming loss and MLCM.

Aligning the FCM Clusters with True Labels For the DADOSD dataset, a threshold of 0.3 was used on the probability given by FCM, i.e., a probability of 0.3 was converted to 1 and 0 if it was below 0.3. A threshold of 0.3 was chosen to give equal importance to the 3 mechanistic classes. Then the proper alignment of the clusters with the true pain classes was identified using hamming loss, which also indicated overall multi-label classification performance.

Calculating Performance Measures The lowest hamming loss indicated the best alignment indicating the cluster represents a particular pain type. Then MLCM was used to evaluate performance in terms of precision, recall, F1-score for individual classes in addition to micro, macro, and weighted scores. Weighted scores are particularly useful for imbalanced datasets like DADOSD.

However, for the TRICD dataset, the true labels are not available, and thus multi-label performance measures could not be computed.

Chapter 6

Results

In this chapter, the experimental results of the methods used are presented. At first, the result of the k-prototypes models for the two datasets is presented for the different number of clusters ($k = 3, 4$). Then the results from the Semi-supervised Learning approaches on the DADOSD dataset are demonstrated. In the end, AE training information and the result from FCM are shown.

6.1 Unsupervised Learning

k-prototype was performed to inspect if the data have inherent clinically meaningful clusters. As no indication of the optimal number of clusters was found from the elbow plot and silhouette score, cluster 3 and 4 was tried with the notion of 3 CP mechanisms, and if a 4th cluster exists representing the mixed type (as discussed in Section 5.1).

The clustering result was assessed using Rand Index (RI), Adjusted Rand Index

(ARI), (Adjusted Mutual Information (AMI), homogeneity and completeness, Fowlkes-Mallows Index (FMI), silhouette coefficient, and Davies-Bouldin Index (DBI) where applicable.

A list comprising of the range of values for the measures and interpretation is given below (more could be found in Subsection 2.1.5.1. However, true class information or label is not needed to calculate the silhouette score and DBI.

RI $[0, 1]$: 0 indicates that the two data clusterings do not agree, and 1 indicates perfect agreement.

ARI $[-1, 1]$: Lower score refers to poor agreements, a negative score refers to the agreement being less than expected from a random result, and 1 is perfect agreement.

AMI $[-1, 1]$: Bad or independent labeling results in negative scores, random label assignments score close to 0, and 1 indicates perfect assignments.

Homogeneity and Completeness $[0, 1]$: Higher value indicates a better clustering result.

FMI $[0, 1]$: A value close to 0 indicates largely independent label assignments, whereas a value close to 1 indicates significant agreement.

Silhouette Score $[-1, 1]$: A higher score indicates better defined and dense clusters. A negative score refers to incorrect clustering, and a score close to 0 indicates overlapping clusters.

DBI $[0, \mathbb{R}^+]$ 0 is the lowest possible score. Scores closer to 0 indicate better partitioning, referring to the clusters being further apart. Here $\mathbb{R}^+$ is a positive real

number. A value larger than 0 denotes that the clusters are closer to each other or overlapping.

6.1.1 k-prototypes

Table 6.1: k-prototypes on DADOSD Dataset ($k = 4$)

Number of Clusters	**Performance Measure**	**Score**
4	RI	0.618
	ARI	0.023
	AMI	0.032
	Homogeneity	0.04
	Completeness	0.038
	FMI	0.282
	Silhouette Score	0.018
	Davies-Bouldin Index	4.52

Table 6.2: k-prototypes on DADOSD Dataset ($k = 3$)

Number of Clusters	**Performance Measure**	**Score**
3	Silhouette Score	0.065
	Davies-Bouldin Index	3.318

Table 6.3: k-prototypes on TRICD Dataset ($k = 4$)

Number of Clusters	**Performance Measure**	**Score**
4	Silhouette Score	0.072
	Davies-Bouldin Index	2.760

Table 6.4: k-prototypes on TRICD Dataset ($k = 3$)

Number of Clusters	Performance Measure	Score
3	Silhouette Score	0.069
	Davies-Bouldin Index	3.004

6.2 Semi-supervised Learning

Semi-supervised Learning was performed to inspect whether the algorithm could identify the CP mechanisms with some information about the true classes. Failure to identify the mechanisms/classes with good performance would refer that the classes are not distinguishable, in other words, overlapping. It would also indicate whether the mixed class should be considered a separate class.

All these performance evaluations were done for DADOSD dataset with respect to the 'Mechnistic_Classes' that has 4 classes, i.e., Nociceptive, Neuropathic, Nociplastic, and Mixed (details could be found in Section 5.2). The performance measures were described in Subsection 2.1.5.2.

The models' training and testing accuracy are reported to indicate the model fit (optimal, overfitting, and underfitting). An insignificant difference between training and testing accuracy is considered optimal and expected.

6.2.1 Semi-supervised SVM

The training accuracy was 40.42%, and the testing accuracy was 38.46%. Table 6.5 contains scores of the performance measures, and Figure 6.1 shows the normalized confusion matrix.

Table 6.5: Report showing the classification metrics for Semi-supervised SVC

	Precision	Recall	F1-score	Support
0	0.4	0.83	0.54	35
1	0.38	0.14	0.2	22
2	1	0.05	0.09	21
3	0.2	0.15	0.17	13
Accuracy			0.38	91
Macro Average	0.49	0.29	0.25	91
Weighted Average	0.5	0.38	0.3	91

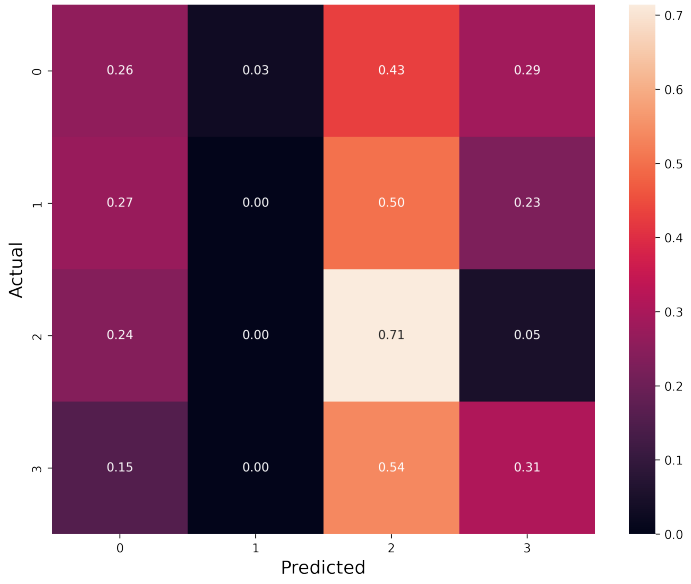


Figure 6.1: Confusion matrix of Semi-supervised SVC (DADOSD Dataset)

6.2.2 Label Propagation

The Label Propagation model achieved a training accuracy of 37.92%, and a classification accuracy of 36.26% on the hold-out test set.

Table 6.6: Report showing the classification metrics for Label Propagation

	Precision	Recall	F1-score	Support
0	0.43	0.66	0.52	35
1	0.00	0.00	0.00	22
2	0.27	0.48	0.34	21
3	0.00	0.00	0.00	13
Accuracy			0.36	91
Macro Average	0.17	0.28	0.22	91
Weighted Average	0.23	0.36	0.28	91

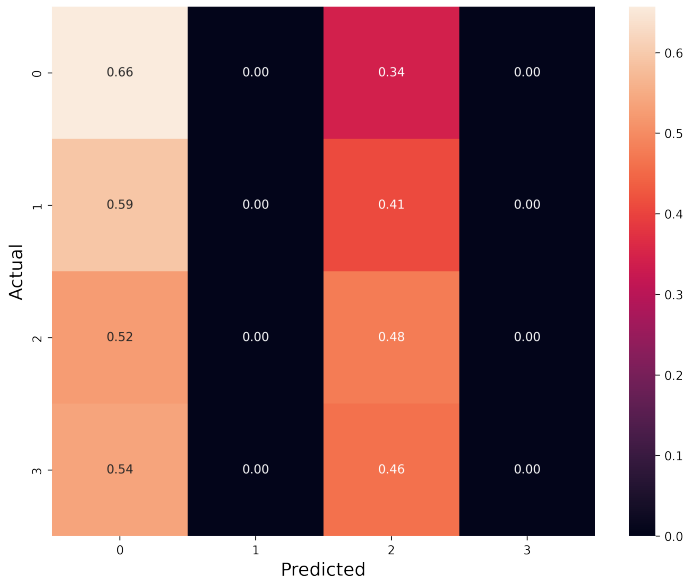


Figure 6.2: Confusion matrix of Label Propagation (DADOSD Dataset)

Table 6.6 contains scores of the performance measures, and Figure 6.2 shows the normalized confusion matrix.

6.2.3 Label Spreading

The Label Spreading model achieved a training accuracy of 35.42%, and a classification accuracy of 35.16% on the hold-out test set.

Table 6.7: Report showing the classification metrics for Label Spreading

	Precision	**Recall**	**F1-score**	**Support**
0	0.51	0.54	0.53	35
1	0.00	0.00	0.00	22
2	0.28	0.52	0.36	21
3	0.14	0.15	0.15	13
Accuracy			0.35	91
Macro Average	0.23	0.31	0.26	91
Weighted Average	0.28	0.35	0.31	91

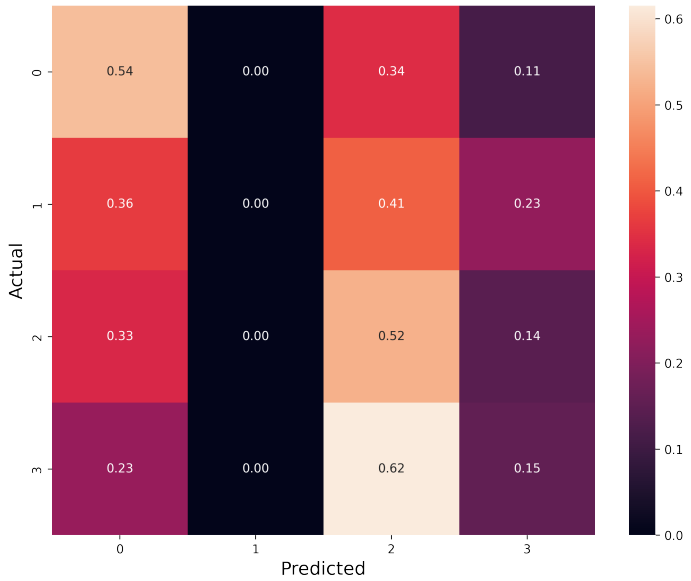


Figure 6.3: Confusion matrix of Label Spreading (DADOSD Dataset)

Table 6.7 contains scores of the performance measures, and Figure 6.3 shows the normalized confusion matrix.

6.3 Overlapping Clustering

Overlapping or Soft Clustering was tried to explain the overlapping tendency of the data, identify and quantify the CP mechanisms, and investigate the performance of multi-label classification in terms of the true labels in DADOSD dataset.

The features extracted by AE were used in FCM. The performance measures were calculated by comparing with one-hot true labels for the DADOSD dataset.

6.3.1 FCM

Regular clustering was done to compare the result with previous results, and soft clustering was performed to identify and quantify co-existing CP mechanisms and observe the performance of multi-label classification.

6.3.1.1 Regular Clustering

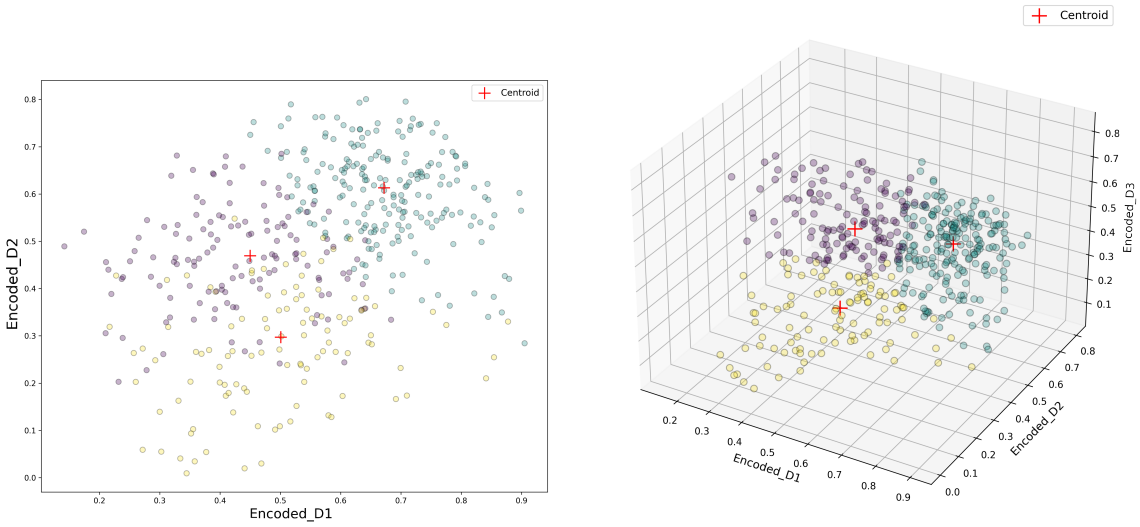
Table 6.8: FCM on AE extracted features DADOSD dataset ($k = 3$)

Number of Clusters	**Performance Measure**	**Score**
3	Silhouette Score	0.158
	Davies-Bouldin Index	2.102

Table 6.9: FCM on AE extracted features TRICD dataset ($k = 3$)

Number of Clusters	**Performance Measure**	**Score**
3	Silhouette Score	0.276
	Davies-Bouldin Index	1.474

Figure 6.4 shows the AE extracted features from the DADOSD dataset are visualized as a scatterplot with FCM cluster labels (regular clustering). The dots are colored based on FCM cluster labels before alignment with CP mechanisms. However, these plots are only for visualizing FCM's regular/hard clustering result.



(a) Two-dimensional scatterplot (b) Three-dimensional scatterplot

Figure 6.4: The AE encoded features from DADOSD dataset visualized as a scatterplot with FCM cluster labels (regular clustering). The dots are colored based on FCM cluster labels before alignment with CP mechanisms.

6.3.1.2 Soft Clustering

Table 6.10 presents a few examples of FCM soft clustering result, which indicates the probability of belonging to a cluster (i.e., CP mechanisms). The higher the probability, the more dominant the CP mechanism.

Table 6.10: Example of FCM outputs

Example	Nociplastic	Nociceptive	Neuropathic
1	0.03	0.68	0.29
2	0.58	0.04	0.38
3	0.02	0.00	0.97
4	0.05	0.48	0.46
5	0.36	0.02	0.63

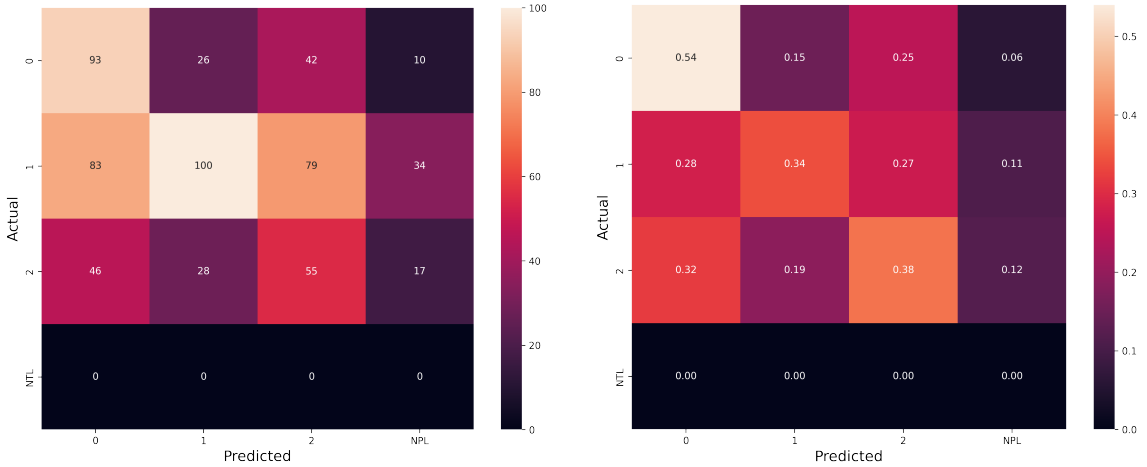
Hamming loss and Cluster Alignment In multi-label classification, hamming loss penalizes only the individual labels which were predicted wrong. Lower hamming loss states better classification (Subsection 2.1.5.3).

Therefore, hamming loss was used to evaluate multi-label classification performance and also to find the proper alignment of the FCM-generated clusters with respect to the true label. For all the combinations, hamming loss was computed and the lowest hamming loss of 0.43 was found for combination E (Appendix C.4), i.e., Nociplastic, Nociceptive, and Neuropathic.

Other Multi-label Classification Measures A multi-label classification was performed after the probabilities resulted from FCM were converted to binary based on a threshold, i.e., a probability of 0.3 or above was converted to 1 and 0 if it was below 0.3 (Subsection 5.3.3).

Figure 6.5 illustrates the raw and normalized MLCM. Details about MLCM were presented in Subsection 2.1.5.3.

Table 6.11 demonstrates the precision, recall, and F1-score based on the results calculated from one-vs-rest confusion matrix. The scores of the performance measures with respect to the individual classes increased compared to the Semi-supervised Learning approaches.



(a) Raw MLCM (b) Normalized MLCM

Figure 6.5: Multi-label confusion matrix of FCM using AE features. The normalized matrix demonstrates the percentage of true and false predictions. In contrast, the raw confusion matrix shows the actual number of counts of the true and the false predictions and indicates the size of each class. Here, NTL is No True Label, and NPL denotes No Predicted Label.

Table 6.11: Overall multi-label classification performance

	Precision	**Recall**	**F1-score**	**Weight**
0	0.42	0.54	0.47	171
1	0.65	0.34	0.44	296
2	0.31	0.38	0.34	146
Micro Average	0.40	0.40	0.40	613
Macro Average	0.46	0.42	0.42	613
Weighted Average	0.50	0.40	0.43	613

Chapter 7

Discussion

In this chapter, the results from the experiments are analyzed. This chapter also discusses how the results of Unsupervised and Semi-supervised Learning led toward viewing CP as a continuum, which was validated by soft clustering. Additionally, the generalizability of the results, explainability, possible implications, and limitations of this work are discussed.

7.1 Interpreting the Results

In this section, the results are interpreted and discussed in detail. At first, the trends of the datasets are discussed from the results of regular clustering. After that, if mixed pain should be considered a separate pain mechanism is explained in the light of the results from the Semi-supervised Learning models. Considering the outcome of Unsupervised and Semi-supervised Learning, the result of soft clustering to identify co-existing CPs is explained.

7.1.1 The Trend in Clustering CP Data

To find the optimal number of clusters, the cost function (sum distance of all points to their respective cluster centroids) that combines the calculation for numerical and categorical variables was used and plotted for different numbers of clusters ($k = 2$ to 10). For DADOSD dataset, the elbow method did not show any inflection point or elbow indicating the optimal number of clusters. Similarly, silhouette score was calculated for different numbers of clusters ($k = 2$ to 10). The average silhouette score did not clearly identify a number of clusters that is appropriate to interpret and validate the consistency within the clusters.

The same trend was observed for the TRICD dataset. No clear indication for the optimal number of clusters was found from the elbow plot and silhouette score plot.

However, for the number of clusters 3 and 4, clustering performance measures were tested. 3 clusters were tried to investigate if they are relatable to the 3 pain mechanisms, whereas 4 clusters were tested to see if the mixed-type itself constructs another cluster. All the relevant performance measures (RI, ARI, AMI, Homogeneity, Completeness, FMI, Silhouette Score, DBI) can only be calculated when true labels are available. However, in the absence of true labels, only silhouette score and DBI could be calculated.

For the DADOSD dataset, true labels were available for the mechanistic classes where mixed-type was labeled as a separate class (0- Nociceptive, 1- Neuropathic, 2- Nociplastic, 3- Mixed). Where RI indicated a good clustering similarity based on the true labels, ARI of 0.023 indicated that showed poor agreements in clustering. AMI was found to be close to 0 (0.032), which indicates the clusters are not pure or random label assignments. Similarly, the FMI value of 0.282 indicates the label assignments

were independent to some extent. Additionally, the clusters do not satisfy Homogeneity and Completeness measures. The scores indicate that the clusters contain instances that are members of multiple classes or all the instances that are members of a given class are members of different clusters. The silhouette score indicates the clusters are overlapping and not well-defined. Also, a DBI value of 4.52 refers that the clusters are not further apart.

For 3 clusters, as true labels were not available, only silhouette score and DBI could be calculated. The scores are similar to 4 clusters. Both silhouette score and DBI slightly improved, referring marginally more defined than that of 4 clusters.

Silhouette score and DBI were also calculated for the TRICD data. For the number of clusters 3 and 4, the differences were insignificant. Silhouette score indicated the overlapping tendency of the clusters where the DBI resonated that the clusters were not further apart.

7.1.2 Should Mixed-Pain be Considered Separately?

Semi-supervised algorithms could only be applied to DADOSD data where mechanistic classes (4 classes) were used as true labels. The performance measures were evaluated in terms of precision, recall, and F1-score, including their macro and weighted averages as a multi-class classification problem.

In terms of training the models, none of the models was overfitted or underfitted as the difference between the training and testing accuracy was insignificant.

In the case of Semi-supervised SVM, for Nociceptive, while F1-score was good where a recall was observed. While the precision was perfect for Nociplastic but the recall was not good as it misclassified Nociceptive as Nociplastic. However, Neuropathic and

Mixed-type were often misclassified as Nociplastic. The macro and weighted F1-score was fairly low (0.25 and 0.3).

In Label Propagation, it did not predict Neuropathic and Mixed at all. The individual recall for Nociceptive and Nociplastic was good, where the precision was on the lower side. Overall, the classifier performed similarly to the Semi-supervised SVM.

The same trend was demonstrated by the Label Spreading model as well. It could not classify Neuropathic at all. However, Label Spreading achieved marginally better macro and weighted F1-score (0.26 and 0.31), which is still not satisfactory with regard to the true labels.

7.1.3 Identifying Co-existing CPs

The reconstruction loss or validation loss for AE was very good for both datasets. For the DADOSD dataset, the validation loss or MSE was 0.983, and for the TRICD dataset, the validation loss was 0.089. Both models were found to be good at reconstructing the data from the encoded latent representation by the encoder.

Both hard/regular and soft/overlapping clustering were performed on features extracted by AE using FCM. Hard or regular clustering was done to observe the clustering ability and compare it with previous results. For 3 clusters, silhouette scores slightly increased for both datasets (Subsection 6.3.1.1) compared to clustering performed by k-prototypes using the Model-Ready data. DBI also decreased in both cases. So, the clusters were slightly more defined and a little further from each other. Overall, the trend is very similar, and the overlapping tendency is still present.

In overlapping clustering, FCM suggested the probability of belonging for each of the 3 clusters for every instance. This empirically quantifies the pain mechanisms

present in patients where the highest probability is the dominant pain mechanism in that patient.

In order to find the best alignment or which cluster represents which mechanism and to calculate the multi-label performance measures, a threshold of 0.3 was used to binarize the probability. The 0.3 threshold was chosen to give the same gravity to all three pain mechanisms. The hamming loss was significantly better than random chance (0.43). It also indicated the clusters related to the pain mechanisms. The other multi-label measures were also found to be better than the performance of Semi-supervised Learning.

7.2 Why Were the Clusters Not Distinguishable?

For both datasets, the result from ARI and silhouette scores indicate that the clusters are not separable and thus overlapping. With the aid of class information, the Semi-supervised models also could not identify the classes where it had the most difficulty with Mixed classes, i.e., Mixed class was often misclassified as Nociplastic. This result is not surprising as the features might be shared among the pain mechanisms, which makes it a Mixed type, and the model failed to find a distinguishable pattern. From these observations, it can be stated that the CP patients' data are not clusterable while revealing distinct pain phenotypes. Additionally, the mixed type should not be considered as a separate class or cluster. The literature also suggests that CP mechanisms co-exist with a prevalence of about 60% or more. This work resonates with the same information.

Overlapping clusters indicated that the pain mechanisms co-exist. The result of hamming loss and the other multi-label performance measures showed that the clusters

could reveal the pain mechanisms. Therefore, the probability score of belonging to the clusters quantifies the pain mechanism present in a patient where the highest probability indicates the dominant pain mechanism in that patient.

7.3 Chronic Pain is a Continuum

From the notion of pain as a continuum, it can be concluded that the pain mechanisms are like a spectrum or continuum where all the mechanisms can be present while one being more dominant than others.

As CP treatment is most effective if the underlying cause or mechanism can be identified, the diagnosis should focus on the mechanism instead of anatomical location-based or symptom-based treatments. Instead of looking at CP as a distinct entity, it should be viewed as a continuum.

This work suggests that CP mechanisms co-exist in a patient, and it can be indicated with a reasonable estimate of which mechanisms are present in a patient and which is the dominant CP mechanism. This is a significant finding while having the potential to aid clinicians in identifying the underlying cause of CP faster and improve the diagnosis and treatment.

7.4 Generalizability

The generalizability of the results cannot be fully ascertained, in part, because of the limited size of the available datasets for model development. However, the results support using AE and FCM as a pipeline of methods to identify and quantify the pain mechanisms present in a patient. It was encouraging that although the datasets were

different, they produced similar results. The results also emphasize that 3 clusters could, and did, represent the clinically meaningful 3 pain mechanisms.

The architecture of the AE was different in addressing the different dimensions of the data. However, the same dimension of bottleneck, which reduced the data to the salient features, suggests a commonality in the data supporting generalizability as FCM takes the extracted features and does not require tuning.

7.5 Explainability

Explainability is a crucial aspect when it comes to medical applications. It is not sufficient to only identify the CP mechanisms. It is also important to indicate the dominant mechanism, so that it becomes more useful for treatment and management purposes. This work quantifies the co-existing CP mechanism, which suggests the dominant mechanism. It was done by indicating the probability of pain mechanisms present.

Table 6.10 demonstrates the result by FCM. The probabilities quantify the CP mechanisms where a higher number indicates the dominant mechanism. For instance, for the patient in example 1, Nociceptive (0.68) is the dominant CP, whereas Neuropathic is present with 0.29 magniture and Nociplastic (0.03) is close to 0.

Therefore, this method not only quantifies the CP mechanisms but also specifies the dominance of one mechanism over the other. It resembles a vector having three dimensions (each dimension representing a pain mechanism), where the magnitudes quantify the mechanism.

Another aspect of explainability was presented by the data used. Patient-reported data and medical history were utilized in this work. As input data can be reconstructed

from the latent representation of the AE, it is possible to see what values were present in the features that produced the result. This can also aid medical experts in understanding the context.

7.6 Impact and Novelty

From the literature review (Chapter 2), it can be observed that when data in context contain one or two CP conditions (not mechanisms) or one to two CP conditions with healthy participants, they might be distinguishable from one another. In the case of this research, the dataset contained all the pain mechanisms which have been used to validate the clusters. No published literature has tried to identify clinically meaningful clusters where it reveals CP mechanisms with quantification and a direction toward the dominant mechanism (conceptually, a pain vector, where the magnitudes quantify the co-existing CPs). A *pain vector* would be more useful from the perspective of patients, and the healthcare system [19].

As there is a lack of gold standard in identifying CP mechanisms, a set of criteria was nominated by Delphi study [19]. However, it is tough to consider and process all the information/criteria together. In the case of overlapping conditions, it might become more challenging for a physician. Moreover, the clinical decisions might not always be consistent due to the presence of possible bias from the clinicians (e.g., different opinions, level of expertise, mental and physical state, etc.). As this method is data-driven, it can possibly avoid bias, and it can help unify clinicians' thoughts.

7.7 Possible Implications

The number of CP patients is rising, and it increases the use of healthcare resources. As a result, wait times increase, making CP more difficult to manage. Delays and sub-optimal management of CP decrease the patient's perception of getting better, which is an important factor in CP management. Additionally, it takes a significant amount of time for the physicians to be able to identify the underlying cause, which makes it more resource-intensive.

This thesis suggests a way of identifying and quantifying the pain mechanisms present in a patient with a justifiable success rate with only patient-reported history and questionnaires data. It also avoids the cognitive bias that might be present in a medical practitioner's diagnosis which makes it fully data-driven.

As it could suggest the pain mechanism without any pathological or imaging test, it arguably takes an hour or two to answer the questionnaires for the patients. The acquired data can be utilized to fast-track the treatment by identifying and quantifying CP, which can aid physicians in making rapid decisions and directing the patients' to the best possible treatments faster. As a consequence, this work can help in decreasing the burden on the healthcare induced by CP patients while reducing the total cost.

7.8 Limitations

This work exhibits a few limitations. Both datasets had many missing points. In most cases, the reasons were unknown, and they could not be fixed as the datasets are retrospective. During the cleaning phase, data were lost due to several reasons. Additionally, features were also removed where more than 40% values were missing.

The impact of the removed features is not known; thus, working with datasets with less or no missing information should be targeted in the future.

For some features, the missing data points were imputed. As Unsupervised Learning was involved, the goal was to keep as much information as possible to minimize or avoid bias. Though the impact of imputation was marginal according to the visualization, the confidence would have been higher if the data points were not missing or the pattern(s) or reason(s) for missing data were known.

Although questionnaires are found to be effective in CP diagnosis, treatment, and management; the answers or the data collected from them are subjective. The questions in the datasets are from several validated questionnaires, but the combination of them was not validated. Validated questionnaires are tested for reliability and validity, which is efficient in both research, and clinical settings [256]. As similar results were observed from two different datasets, the models' performances are expected to remain unaffected. Yet, a set of validated questionnaires could reduce any effects due to subjectivity.

Additionally, the dataset (DADOSD) is imbalanced in terms of the distribution of patients with pain mechanisms. Though measures were taken to avoid the impact of the dataset imbalance on the result, fewer variations were present from the underrepresented groups. From the algorithm's perspective, the underrepresented or minority class might affect the models negatively due to possible low variations or patterns to learn.

The SME reviewed the doctor's diagnosis established at the clinic while adding the CP mechanism labels. Though relevant Delphi studies were taken into account with clinical knowledge, the true labels might contain bias from the doctors and the

SME who labeled the instances indicating CP mechanisms. Though the anticipated influence on the true labels and ML model's performance is minimal, it would have been superior if multiple doctors or SMEs had reviewed the labelings and reached a consensus.

Both datasets are relatively small in size. The datasets are different considering the set of available features (e.g., body diagram is present in TRICD data but not in DADOSD data). Though while analyzing the datasets, a similar trend was observed, the aspect of generalizability depends largely on the availability of a similar dataset from a different cohort.

Chapter 8

Conclusion

In this chapter, the thesis contributions are summarized, and possible avenues for future research are discussed.

8.1 Summary of Contributions

Chronic pain is a vast area of medical science, and it is complicated to diagnose and manage. As it is one of the leading causes of accessing the healthcare system and disability, it has become a global burden. Although the most effective way to treat CP is to identify the underlying cause or mechanism, it is often unattainable in a clinical setting. Most of the time, the pain mechanisms co-exist, making identification more challenging.

Although several existing works attempted to cluster CP conditions, there is a complete lack of work that helps identify co-existing CP mechanisms while quantifying them. This work attempted to identify the existing CP mechanisms (i.e., Nociceptive, Neuropathic, and Nociplastic) within a patient using Unsupervised Learning while

quantifying it without the help of diagnosis and treatment information.

This work involved working with patient-reported history and questionnaire data. Two datasets were used where the first dataset had labels indicating the mechanism(s) present in each patient, and a relatively small dataset was used for observing the trends compared to the first dataset. After preprocessing the datasets, Unsupervised Learning or clustering technique (k-prototypes) was applied where no class information was provided to the algorithm. Significant overlaps were observed where no optimal number of clusters was revealed. The overlaps indicate that CP mechanisms cannot be discerned or classified as distinct disorders. Additionally, it was shown that mixed pain mechanisms do not make a separate cluster or class.

From the results of Semi-supervised Learning, the same characteristics were observed. The classes or mechanisms were not classified with good performance even though the algorithm had the privilege to leverage some true class information. It was evident that the classes were overlapping. Therefore, from the results from k-prototypes and Semi-supervised Learning models, it became clear that the CP mechanisms co-exist, and rather than distinct entities, CP should be considered as a continuum.

With the help of Autoencoder, hidden features were extracted, and dimension was reduced. An overlapping clustering technique was employed, which revealed that 3 CP mechanisms could be identified with good performance (hamming loss of 0.43) while explaining the overlaps. Additionally, the pain mechanisms were also quantified, and the dominant CP mechanism was indicated in each patient in the data, which is considered significant in CP treatment and management.

This work is data-driven proof that CP should be considered as a continuum. CP

is a spectrum where all the CP mechanisms can also co-exist in a patient. Additionally, one CP mechanism can be more dominant than the others. However, validation of this work with data from a larger cohort, in addition to other clinical features, might improve the performance of the current pipeline of the algorithms. Furthermore, clinical validation is necessary before utilizing it in a real-world setting, though the same trend as found in this work is expected to be observed. Therefore, rather than trying to identify distinct CP phenotypes or mechanisms, CP should be considered a continuum where all CP mechanisms co-exist.

8.2 Future Directions

In this section, future directions and recommendations for future works are discussed. While emphasizing on minimizing the limitations (Section 7.8), other steps are documented here.

Firstly, the efficacy of the method of clustering and identifying CP mechanisms can be further assessed and evaluated with patients in a clinical setting. It can be done by testing the algorithm's performance against clinically set standards.

Secondly, the focus will also be on improving the overall accuracy (hamming loss) and extending the model presented in this thesis. This can be pursued from the standpoints of data and ML models. Validated questionnaires could be beneficial for the model in particular. This work can also be extended and tested for generalizability with a larger dataset from a different cohort. From the algorithms' point of view, the categorical features could be converted into numerical/continuous, which could open the window for other clustering algorithms.

Additionally, a validated set of questionnaires can be used with data governance

for data gathering and storing. Steps should be taken toward having an ideal dataset consisting of all the clinically important features. Data collected with the goal of pain mechanism identification might help increase the model's performance. Researchers can focus on improving the algorithm's performance by having clinical tests such as QST along with data comprised of a set of validated questionnaires. As criteria were suggested through Delphi study towards the establishment of a gold standard to distinguish CP mechanisms [19], it might be the next reasonable step to see if the criteria come out ahead in explaining the mechanisms and their overlaps or co-existence. In addition to that, the true labels could be reviewed by multiple experts while considering the suggested criteria to have more unified labeling and minimize bias.

Moreover, variables could be prioritized by clinical importance, possibly derived from a consensus of the SMEs. The ML model's performance can be tested by training it on the selected features. A focus could be on ranking the features based on contribution or impact on the result, which can be helpful in the clinical setting. However, bias could be present in this approach and should be carefully addressed.

Explainability is another significant aspect of ML in a clinical setting. Here, the dominant CP mechanism was indicated with a magnitude, but the next step might be to identify what are the driving factors behind the magnitude or the level of the magnitude. In terms of the AE, even though the data can be reconstructed from the AE-extracted features, it would be more intuitive if explainable components could be added to the output of FCM. So, adding explainability with FCM's output or explaining the clusters could be the step towards explainability.

More research is needed in the field of CP especially focusing on CP mechanisms.

Anonymized data should be open-sourced to give the researchers flexibility and attract more researchers to work in this arena.

Finally, future research should be done to improve the efficiency of the ML models in finding co-existing CP mechanisms while considering CP as a spectrum or continuum.

Appendix A

Literature Review

A.1 Search Strings

Web of Science Core Collection (TS=((chronic OR persistent OR recurrent) NEAR/3 Pain) AND TS=("Machine Learning" OR "Artificial* Intelligen*" OR "Deep Learning" OR (Unsupervised NEAR/2 Learning) OR "Artificial Neural Network" OR "Convolutional Neural Network" OR "Natural Language Processing" OR "Cluster Analysis" OR "Clustering"))

ACM Digital Library ((Chronic AND Pain) "chronic widespread pain" "persistent pain" "recurrent pain") AND ("Machine Learning" "Artificial Intelligence" "Deep Learning" "Unsupervised Learning" "Unsupervised Machine Learning" "Artificial Neural Network" "Convolutional Neural Network" "Natural Language Processing" "Cluster Analysis" "Cluster*")

IEEE Xplore (((chronic OR persistent OR recurrent) NEAR/3 Pain) AND ("Machine Learning" OR "Artificial* Intelligen*" OR "Deep Learning" OR (Unsupervised NEAR/2 Learning) OR "Artificial Neural Network" OR "Convolutional Neural Network" OR "Natural Language Processing" OR "Cluster Analysis" OR "Cluster*" "All Metadata":"Full Text & Metadata":))

A.2 Screening the Articles

The Full-text Screening (after Title and Abstract Screening) guidelines are provided here. A flowchart in Figure A.1 shows the series of steps that were involved.

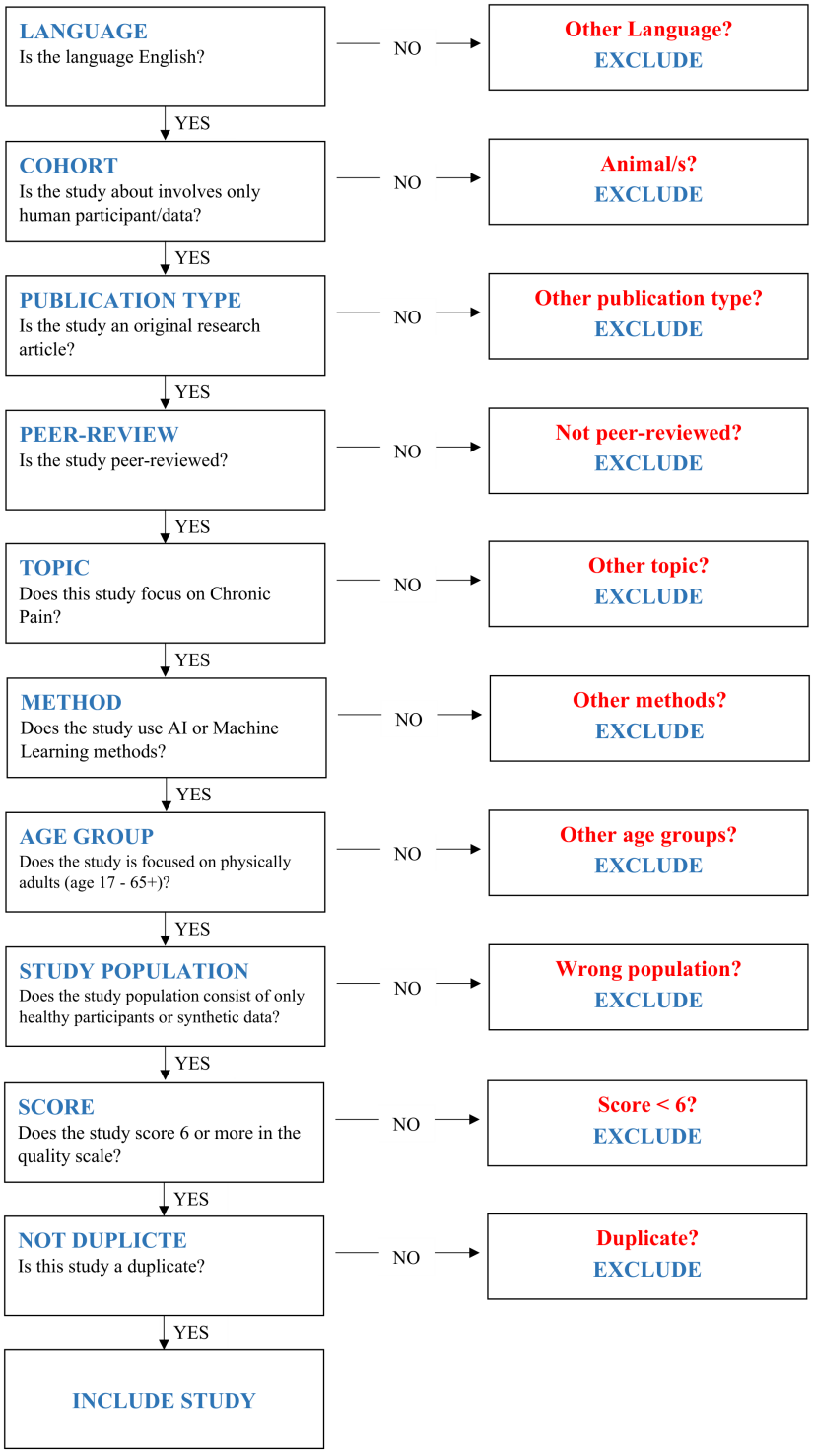


Figure A.1: Scoping review: Full-text screening

Appendix B

Datasets

B.1 Ethical Approval and Data Retention

B.1.1 Research Ethics Board Approval

An ethics application (MREB#: 5567) was submitted to McMaster University Research Ethics Board (MREB) for this research. The application was reviewed and cleared by the MREB on December 08, 2021, to ensure compliance with the Tri-Council Policy Statement and the McMaster Policies and Guidelines for Research Involving Human Participants.

B.1.2 Storage and Use

UHN de-identified the data that minimizes privacy and confidentiality risks. Nevertheless, we recognize that privacy and security are still a concern. As participants' privacy and confidentiality have been addressed by the collector of the data (UHN), we do not believe that there is a high risk associated with this data. The data is

stored on Dr. Doyle's password-protected server behind the McMaster University's server, secured by the McMaster firewall, and physically resides in a locked room at the university. This also addresses the issues regarding the storage and security of research data mentioned in the MREB [248] Data Storage & Security Guide. The research data will only be accessible to the research team comprised of the Faculty Supervisor (Dr. Doyle), and the Student Principal Investigator (Md Asif Khan) using passwords and controlled access to the server. The datasets were received from UHN on 22nd February 2022, and the plan is to retain the data till December 2022 (the expected graduation date for the Student Principal Investigator).

B.2 Comparison Between the Datasets

Table B.1: Comparison between TRICD and DADOSD datasets. This applies to the raw datasets. Here, 'Yes' indicates the feature is present, and 'No' indicates the feature is not available in that particular dataset.

Fields	**TRICD Dataset**	**DADOSD Dataset**
No. of Patients (rows)	201	738
No. of Variables (columns)	60	146
Sex	Yes	Yes
Age	Yes	Yes
Employment Status	Yes	No
Body Diagram (PSA)	Yes	No
Pain Persistence	No	Yes
Pain Time Gets Worse	No	Yes

Continued on the next page

Table B.1 continued from the previous page

Fields	TRICD Dataset	DADOSD Dataset
Pain Description	No	Yes
Pain Worse	No	Yes
Pain Better	No	Yes
Pain Improvement	No	Yes
Pain Affected by Stress	No	Yes
Pain Worsen Time	No	Yes
Diagnosis Reported by Patient	No	Yes
Brief Pain Inventory (BPI)	Yes	Yes
No. of Medications	Yes	Yes
More Information	No	Yes
Side Effects	No	Yes
Medications Stopped	No	Yes
Medications Type	No	Yes
No. of Opioids	Yes	No
Smoking	Yes	Yes
Alcohol	No	Yes
Recreational	No	Yes
Comorbidities	No	Yes
Co-existing Pain Conditions	Yes	Yes

Continued on the next page

Table B.1 continued from the previous page

Fields	TRICD Dataset	DADOSD Dataset
Approaches (Massage, Acupuncture, Chiropractic, Physio, Mental, osteopathy)	Yes	Yes
Workplace Safety and Insurance Board (WSIB) [257] Claim	Yes	Yes
Patient Stages of Change Questionnaire	Yes	No
Pain Patient Profile	Yes	No
Disposition	Yes	No
Mechanistic Classification	No	Yes

Note: This comparison table was initially provided by UHN and later modified by Md Asif Khan.

B.3 Reasons for Feature and Instance Removal

The reasons for features and subject removal are listed here. These steps were taken after careful analysis and consulting with SMEs. The rows were removed before the column/feature removal.

Reasons for instance (row) removal:

- Mechanistic classification labels missing (DADOSD data)
- More than 20% missing data points

- Duplicates

Reasons for feature removal:

- Unformatted and unstandardized text
- Unstandardized inputs
- Treatment and diagnosis
- Not relevant (e.g., WSIB claim number)
- More than 40% missing values
- Duplicates

B.4 Data Variables

DADOSD Dataset The table below contains the DADOSD dataset (cleaned) variables and their description.

Table B.2: DADOSD dataset variables and their description

Variables	Description
Subject_ID	Pseudo Subject ID
Age_Calculated	Calcualted from DOB and DOE (years)
Gender_Coded	Male - 0, Female- 1
Pain_Persistence_Coded	Comes and goes - 0, Always there - 1
Pain_Time	No - 0, Yes - 1
PF_Throbbing	No - 0, Yes - 1

Table B.2 continued from the previous page

Variables	**Description**
PF_Shooting	No - 0, Yes - 1
PF_Stabbing	No - 0, Yes - 1
PF_Sharp	No - 0, Yes - 1
PF_Cramping	No - 0, Yes - 1
PF_Gnawing	No - 0, Yes - 1
PF_Burning	No - 0, Yes - 1
PF_Aching	No - 0, Yes - 1
PF_Heavy	No - 0, Yes - 1
PF_Tender	No - 0, Yes - 1
PF_Splitting	No - 0, Yes - 1
PF_Tiring	No - 0, Yes - 1
PF_Sickening	No - 0, Yes - 1
PF_Fearful	No - 0, Yes - 1
PF_Punishing	No - 0, Yes - 1
Pain_Improvement_Coded	Getting worse - 0, Getting better - 1, Staying about the same - 2
Pain_Affected_by_Mood	No - 0, Yes - 1
Worst_Pain_Last_24hrs	Scale (last 24 hours): 0 - No Pain, 1, 2, 3, 4, 5, 6, 7 , 8, 9, 10 - Worst Pain Imaginable
Least_Pain_Last_24hrs	
Average_Pain	
Pain_Right_Now	

Table B.2 continued from the previous page

Variables	Description
Pain_Relief_Pct	Percentage of relief from treatment or medication provided 0% - No relief, 10% - 1, 20% - 2, 30% - 3, 40% - 4, 50% - 5, 60% - 6, 70% - 7, 80% - 8, 90% - 9, 100% - 10 (Complete relief)
General_Activity_Interference	Scale: 0 - Does not interfere, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10- Completely interferes
Mood_Interference	
Walking_Ability_Interference	
Normal_Work_Interference	
Relationship_Interference	
Sleep_Interference	
Enjoy_Life_Interference	
Concentration_Ability_Interference	
Apetite_Interference	
BPI_Score	BPI from Raw Data
Pain_Med_Qty_Coded	Total count of medicine
Pain_Med_SideEffects	No - 0, Yes - 1, Uncertain - 2
Med_More_Info_Needed	No - 0, Yes - 1
Nicotine_Smoking	No - 0, Yes - 1
Alcohol_Consumption_Weekly	

Table B.2 continued from the previous page

Variables	Description
_Categorized	0- None/quit, 1- Occasional, rarely, one/two glass/es a week, 2- More than 1.5 litres, 3- 6+ drinks
Medical_History_Surgery_Coded	None - 0, Others- 1
WSIB_Claim	No - 0, Yes - 1
Litigation	No - 0, Yes - 1
Mechanistic_Classes	Nociceptive - 0, Neuropathic - 1, Nociplastic - 2, Mixed - 3
Nociceptive	No - 0, Yes - 1
Neuropathic	No - 0, Yes - 1
Nociplastic	No - 0, Yes - 1

TRICD Dataset The table below contains the TRICD dataset (cleaned) variables and their description.

Table B.3: TRICD dataset variables and their description

Variables	Description
Subject_ID	Pseudo Subject ID
Sex_Coded	Male - 0, Female - 1
Age	Years
Employment_Status_Categorized	Employed - 0, Unemployed - 1, Retired - 2, Unknown - 3

Table B.3 continued from the previous page

Variables	Description
Body_Diagram	Body diagram value
BPI_Pain_Severity_Pct_Gen	Pain severity percentage
BPI_Relief_Pct_Gen	Pain relief percentage
BPI_Pain_Interference_Pct_Gen	Pain interference percentage
Med_Analgesics	Number of analgesics
Med_Opiods	Number of opioids
Smoke_Categorized	No - 0, Yes - 1, Unknown - 2
EtOH_week_Categorized	No - 0, Yes - 1, Unknown - 2
WSIB_Claim	No - 0, Yes - 1
Legal_Case_Active	No - 0, Yes - 1
PSCG_Age_1	From PSOCG questionnaire
PSCG_Mean_1	
PSCG_Age_2	
PSCG_Mean_2	
PSCG_Age_3	
PSCG_Mean_3	
PSCG_Age_4	
PSCG_Mean_4	

Table B.3 continued from the previous page

Variables	Description
Current_Stage_Coded	Unknown - 0, Precontemplation - 1, Precontemplation/Contemplation - 1.5, Contemplation - 2, Contemplation/Action - 2.5, Action - 3, Action/Maintenance - 3.5, Maintenance - 4
P3_DEP_T_Score	From P3 Questionnaire
P3_ANX_T_Score	From P3 Questionnaire
P3_SOM_T_Score	From P3 Questionnaire
Dpn_Interprofessional_Therapy _Binarized	No - 0, Yes - 1
Dpn_Groups_SM_Mindfulness	No - 0, Yes - 1
Dpn_Group_Exercise_Hydrotherapy	No - 0, Yes - 1
Dpn_Meds_Management	No - 0, Yes - 1
Dpn_Psych_Consult	No - 0, Yes - 1
Edu_Videos	No - 0, Yes - 1
Edu_Exercises	No - 0, Yes - 1

Appendix C

Data and Model

This chapter focuses on data, and ML model parameters, including details about their corresponding libraries used.

C.1 Data Analysis

C.2 Correlation Heatmap

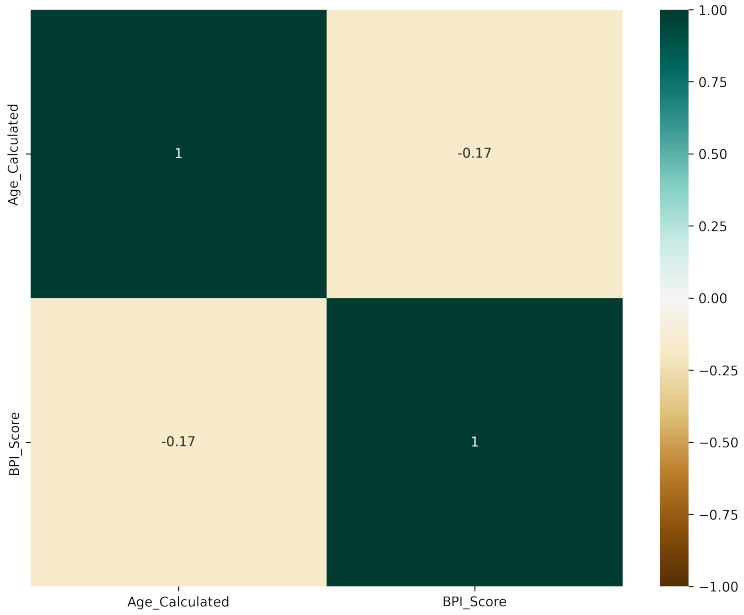


Figure C.1: Correlation heatmap of numerical variables of DADOSD dataset

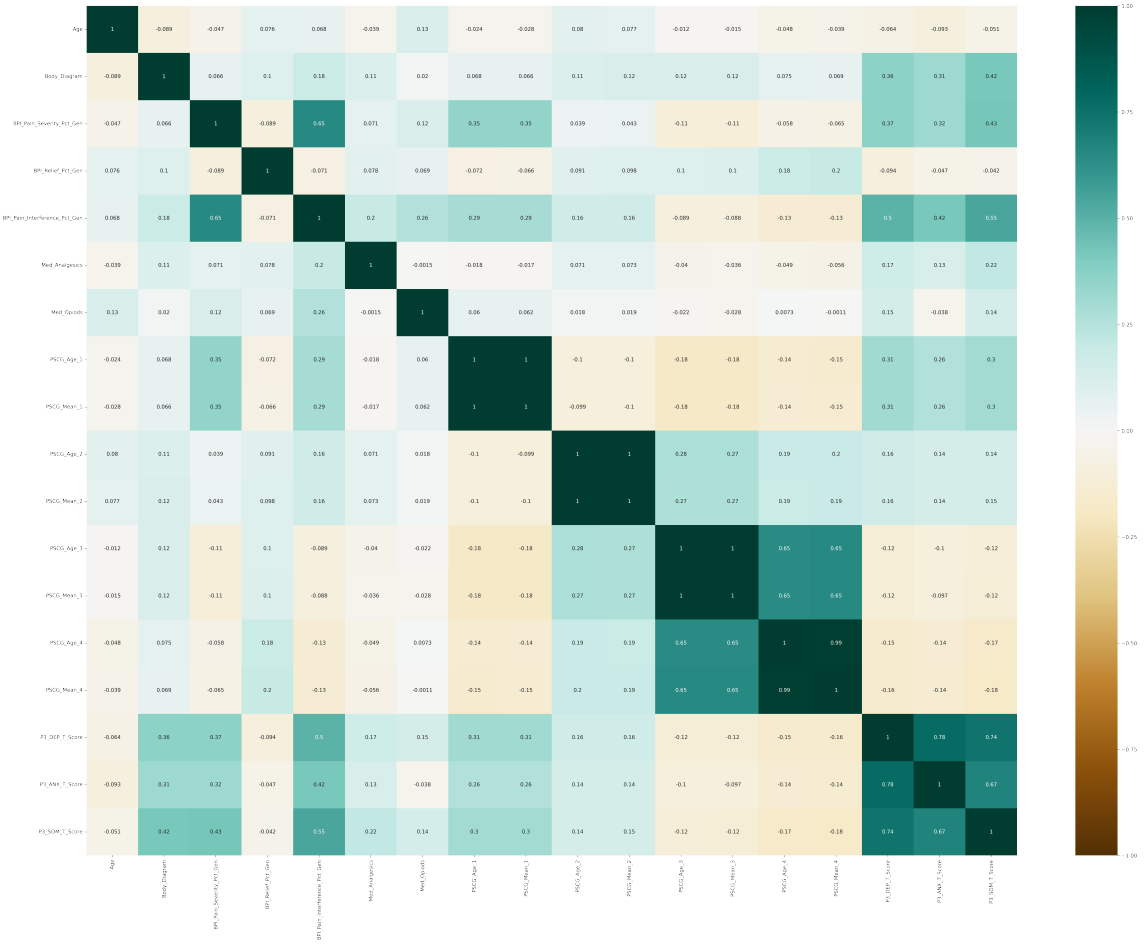


Figure C.2: Correlation heatmap of numerical variables of TRICD dataset

C.3 Model Hyper-parameters and Libraries

The ML model hyper-parameters and used libraries will be shared here. However, the Python version used in this work was 3.9.9. The random state was set to 1024 for all the models where applicable.

C.3.1 Visualization

C.3.1.1 UMAP

For the implementation of UMAP, *umap-learn* library (version 0.5.3) was used [258]. For the 2D and 3D scatterplots, n_components parameter was 2 and 3, respectively. Additionally, the other relevant parameters were changed accordingly. An example is provided below.

```
# numerical_features_fit
umap.UMAP(n_neighbors=35, n_components=2, metric='l2',
min_dist=0.3, n_epochs = 1000, random_state=seed,
transform_seed=seed, verbose = 1).fit(numerical.values)


# categorical_features_fit
umap.UMAP(n_neighbors=35, n_components=2, metric='dice',
min_dist=0.8, n_epochs = 1000, random_state=seed,
transform_seed=seed, verbose = 1).fit(categorical.values)
```

C.3.2 Unsupervised Learning (Regular Clustering)

C.3.2.1 k-prototypes

The implementation of the k-prototypes algorithm was done using *kmodes* (version: 0.12.2) library in Python [145]. The library can be found here.

DADOSD Dataset k-prototypes model parameters:

```
# cluster = range(2, 11)
```

```
n_clusters = cluster, init = 'Huang', gamma = None,
max_iter = 200, random_state = seed
```

TRICD Dataset k-prototypes model parameters:

```
# cluster = range(2, 11)
n_clusters = cluster, init = 'Huang', gamma = None,
max_iter = 200, random_state = seed
```

C.3.3 Semi-supervised Learning

Only applicable to the DADOSD dataset. The implementations of the Semi-supervised Learning algorithms were done using Python libraries by *scikit-learn* project (version 1.0.2) [259]. The libraries can be explored here: Self-Training, Grid Search, SVC, Label Propagation, Label Spreading.

C.3.3.1 SVC

Parameter Tuning The set of parameters selected for Grid Search to find the best fit is given below.

```
params = {
    'C': [0.1, 1, 10, 100, 1000],
    'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
    'degree': [3, 4, 5, 6, 7, 8],
    'kernel': ['poly', 'rbf'],
    'decision_function_shape': ['ovo', 'ovr']
}
```

SVC tuned parameters Fitting 5 folds for each of 600 candidates, totaling 3000 fits:

```
{'C': 1, 'decision_function_shape': 'ovo', 'degree': 6,
'gamma': 0.001, 'kernel': 'poly'}
```

Final Model Parameters The final set of parameters for the Self-training model and Semi-supervised SVC estimator model was:

```
{'base_estimator__C': 1,
 'base_estimator__break_ties': False,
 'base_estimator__cache_size': 200,
 'base_estimator__class_weight': 'balanced',
 'base_estimator__coef0': 0.0,
 'base_estimator__decision_function_shape': 'ovo',
 'base_estimator__degree': 6,
 'base_estimator__gamma': 0.001,
 'base_estimator__kernel': 'poly',
 'base_estimator__max_iter': -1,
 'base_estimator__probability': True,
 'base_estimator__random_state': 1024,
 'base_estimator__shrinking': True,
 'base_estimator__tol': 0.001,
 'base_estimator__verbose': False,
 'base_estimator': SVC(C=1, class_weight='balanced',
 decision_function_shape='ovo', degree=6,
```

```
gamma=0.001, kernel='poly', probability=True,
random_state=1024),
'criterion': 'k_best',
'k_best': 10,
'max_iter': 100,
'threshold': 0.75,
'verbose': True}
```

C.3.3.2 Label Propagation

Label Propagation model parameters:

```
kernel = 'knn', n_neighbors = 160, max_iter = 1000
```

C.3.3.3 Label Spreading

Label Spreading model parameters:

```
kernel = 'knn', n_neighbors = 50, alpha = 0.85,
max_iter = 1000, tol = 0.0001
```

C.3.4 Unsupervised Learning (AE and FCM)

C.3.4.1 Autoencoder

AE was implemented using *tensorflow* (version 2.5.0) [260]. The model summary for the DADOSD is given below:

```
-----------------------------------------------------------
```

```
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 44)]              0
_________________________________________________________________
encoder (Functional)         (None, 3)                 1336
_________________________________________________________________
model (Functional)           (None, 44)                1377
=================================================================
Total params: 2,713
Trainable params: 2,713
Non-trainable params: 0
_________________________________________________________________
```

The model summary for the TRICD is given below:

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, 28)]              0
_________________________________________________________________
encoder (Functional)         (None, 3)                 535
_________________________________________________________________
model (Functional)           (None, 28)                560
=================================================================
Total params: 1,095
Trainable params: 1,095
```

```
Non-trainable params: 0
_________________________________________________________________
```

C.3.4.2 FCM

FCM was implemented using *pyclustering* (version 0.10.1) and *fuzzy-c-means*(version 1.6.4) [105, 261].

FCM model parameters:

```
# initial centers
initial_centers = kmeans_plusplus_initializer(data,
amount_centers = 3, kmeans_plusplus_initializer.
FARTHEST_CENTER_CANDIDATE).initialize()

# Fuzzy C-Means algorithm
initial_centers
```

C.4 Finding Alignment with FCM Clusters

Table C.1: List of orders of the CP categories

Combinations	**Order of the CP Categories**		
A	Nociceptive	Neuropathic	Nociplastic
B	Nociceptive	Nociplastic	Neuropathic
C	Neuropathic	Nociceptive	Nociplastic
D	Neuropathic	Nociplastic	Nociceptive

E	Nociplastic	Nociceptive	Neuropathic
F	Nociplastic	Neuropathic	Nociceptive

Bibliography

[1] 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*, vol. 2, no. 4, pp. 230–243, 2017.

[2] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, p. 94, 2019.

[3] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[4] P. Mäntyselkä, E. Kumpusalo, R. Ahonen, A. Kumpusalo, J. Kauhanen, H. Vinamäki, P. Halonen, and J. Takala, "Pain as a reason to visit the doctor: a study in finnish primary health care," *Pain*, vol. 89, no. 2-3, pp. 175–180, 2001.

[5] D. S. Goldberg and S. J. McGee, "Pain as a global public health priority," *BMC public health*, vol. 11, no. 1, pp. 1–5, 2011.

[6] J. L. S. Sauver, D. O. Warner, B. P. Yawn, D. J. Jacobson, M. E. McGree, J. J. Pankratz, L. J. Melton III, V. L. Roger, J. O. Ebbert, and W. A. Rocca, "Why patients visit their doctors: Assessing the most prevalent conditions in a defined american population," *Mayo Clinic Proceedings*, vol. 88, pp. 56–67, Jan. 2013.

[7] D. S. Martin, "Pain management clinics: What to expect and how to find one." https://www.webmd.com/pain-management/pain-clinics-all-about, Sep 2019. Accessed on 2021-04-04.

[8] "Resources." https://uspainfoundation.org/resources/, May 2021. Accessed on 2021-05-13.

[9] J. Steglitz, J. Buscemi, and M. J. Ferguson, "The future of pain research, education, and treatment: a summary of the iom report "relieving pain in america: a blueprint for transforming prevention, care, education, and research"," *Translational behavioral medicine*, vol. 2, no. 1, pp. 6–8, 2012.

[10] F. Campbell, M. Hudspith, M. Choinière, H. El-Gabalaw, J. Laliberté, M. Sangster, J. Swidrovich, and L. Wilhelm, "Canadian pain task force report: October 2020." https://www.canada.ca/en/health-canada/corporate/about-health-canada/public-engagement/external-advisory-bodies/canadian-pain-task-force/report-2020.html, Oct 2020. Accessed on 2021-04-12.

[11] M. Dueñas, B. Ojeda, A. Salazar, J. A. Mico, and I. Failde, "A review of chronic pain impact on patients, their social environment and the health care system," *Journal of pain research*, vol. 9, p. 457, 2016.

[12] N. Ambardekar, "What is chronic pain management? symptoms and reasons to control chronic pain." https://www.webmd.com/pain-management/guide/understanding-pain-management-chronic-pain, May 2021. Accessed on 2021-05-15.

[13] R.-D. Treede, W. Rief, A. Barke, Q. Aziz, M. I. Bennett, R. Benoliel, M. Cohen, S. Evers, N. B. Finnerup, M. B. First, *et al.*, "Chronic pain as a symptom or a disease: the iasp classification of chronic pain for the international classification of diseases (icd-11)," *Pain*, vol. 160, no. 1, pp. 19–27, 2019.

[14] N. Ambardekar, "Pain management guide." `https://www.webmd.com/pain-management/guide/understanding-pain-management-chronic-pain#1`, May 2021. Accessed on 2021-05-15.

[15] S. P. Cohen, L. Vase, and W. M. Hooten, "Chronic pain: an update on burden, best practices, and new advances," *The Lancet*, vol. 397, no. 10289, pp. 2082–2097, 2021.

[16] R. L. Chimenti, L. A. Frey-Law, and K. A. Sluka, "A Mechanism-Based Approach to Physical Therapist Management of Pain," *Physical Therapy*, vol. 98, pp. 302–314, 04 2018.

[17] M. A. Shraim, H. Masse-Alarie, and P. W. Hodges, "Methods to discriminate between mechanism-based categories of pain experienced in the musculoskeletal system: a systematic review," *Pain*, vol. 162, no. 4, pp. 1007–1037, 2021.

[18] E. Kosek, D. Clauw, J. Nijs, R. Baron, I. Gilron, R. E. Harris, J.-A. Mico, A. S. Rice, and M. Sterling, "Chronic nociplastic pain affecting the musculoskeletal system: Clinical criteria and grading system," *Pain*, vol. 162, no. 11, pp. 2629–2634, 2021.

[19] M. A. Shraim, K. A. Sluka, M. Sterling, L. Arendt-Nielsen, C. Argoff, K. S.

Bagraith, R. Baron, H. Brisby, D. B. Carr, R. L. Chimenti, *et al.*, "Features and methods to discriminate between mechanism-based categories of pain experienced in the musculoskeletal system: a delphi expert consensus study," *Pain*, pp. 10–1097, 2022.

[20] J. Dahlhamer, J. Lucas, C. Zelaya, R. Nahin, S. Mackey, L. DeBar, R. Kerns, M. Von Korff, L. Porter, and C. Helmick, "Prevalence of chronic pain and high-impact chronic pain among adults—united states, 2016," *Morbidity and Mortality Weekly Report*, vol. 67, no. 36, p. 1001, 2018.

[21] A. Fayaz, P. Croft, R. Langford, L. Donaldson, and G. Jones, "Prevalence of chronic pain in the uk: a systematic review and meta-analysis of population studies," *BMJ open*, vol. 6, no. 6, p. e010364, 2016.

[22] É. B. d. M. Vieira, J. B. S. Garcia, A. A. M. d. Silva, R. L. T. M. Araújo, R. C. S. Jansen, and A. L. X. Bertrand, "Chronic pain, associated factors, and impact on daily life: are there differences between the sexes?," *Cadernos de saude publica*, vol. 28, pp. 1459–1467, 2012.

[23] T. Vos, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, R. S. Abdulkader, A. M. Abdulle, T. A. Abebo, S. F. Abera, *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet*, vol. 390, no. 10100, pp. 1211–1259, 2017.

[24] F. Campbell, M. Hudspith, M. Choinière, H. El-Gabalaw, J. Laliberté, M. Sangster, J. Swidrovich, and L. Wilhelm, "Chronic pain in canada:

Laying a foundation for action." https://www.canada.ca/content/dam/hc-sc/documents/corporate/about-health-canada/public-engagement/external-advisory-bodies/canadian-pain-task-force/report-2019/canadian-pain-task-force-June-2019-report-en.pdf, Jun 2019. Accessed on 2021-03-10.

[25] T. Vos, S. S. Lim, C. Abbafati, K. M. Abbas, M. Abbasi, M. Abbasifard, M. Abbasi-Kangevari, H. Abbastabar, F. Abd-Allah, A. Abdelalim, *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[26] S. Meints and R. Edwards, "Evaluating psychosocial contributions to chronic pain outcomes," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 87, pp. 168–182, 2018. Chronic Pain and Psychiatric Disorders.

[27] N. K. Tang and C. Crane, "Suicidality in chronic pain: a review of the prevalence, risk factors and psychological links," *Psychological medicine*, vol. 36, no. 5, pp. 575–586, 2006.

[28] D. Smith, R. Wilkie, O. Uthman, J. L. Jordan, and J. McBeth, "Chronic pain and mortality: a systematic review," *PloS one*, vol. 9, no. 6, p. e99048, 2014.

[29] W. Raffaeli and E. Arnaudo, "Pain as a disease: an overview," *Journal of pain research*, vol. 10, p. 2003, 2017.

[30] T. P. Jackson, V. S. Stabile, and K. K. McQueen, "The global burden of

chronic pain." https://pubs.asahq.org/monitor/article/78/6/24/3059/The-Global-Burden-Of-Chronic-Pain, Jun 2014. Accessed on 2021-04-04.

[31] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, 03 2020. baaa010.

[32] T. Lysaght, H. Y. Lim, V. Xafis, and K. Y. Ngiam, "Ai-assisted decision-making in healthcare," *Asian Bioethics Review*, vol. 11, no. 3, pp. 299–314, 2019.

[33] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. B. Abdallah, and A. Kronzer, "Predicting hospital readmission via cost-sensitive deep learning," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 6, pp. 1968–1978, 2018.

[34] T. Keerthika and K. Premalatha, "An effective feature selection for heart disease prediction with aid of hybrid kernel svm," *International Journal of Business Intelligence and Data Mining*, vol. 15, no. 3, pp. 306–326, 2019.

[35] G. Manogaran and D. Lopez, "Health data analytics using scalable logistic regression with stochastic gradient descent," *International Journal of Advanced Intelligence Paradigms*, vol. 10, no. 1-2, pp. 118–132, 2018.

[36] R. M. Sadek, S. A. Mohammed, A. R. K. Abunbehan, A. K. H. A. Ghattas, M. R. Badawi, M. N. Mortaja, B. S. Abu-Nasser, and S. S. Abu-Naser, "Parkinson's disease prediction using artificial neural network," *International Journal of Academic Health and Medical Research (IJAHMR)*, vol. 3, no. 1, 2019.

[37] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," *Advances in neural information processing systems*, vol. 29, 2016.

[38] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine learning for healthcare conference*, pp. 301–318, PMLR, 2016.

[39] J. Cho, K. Lee, E. Shin, G. Choy, and S. Do, "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?," *arXiv preprint arXiv:1511.06348*, 2015.

[40] F. van Wyk, A. Khojandi, R. Kamaleswaran, O. Akbilgic, S. Nemati, and R. L. Davis, "How much data should we collect? a case study in sepsis detection using deep learning," in *2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT)*, pp. 109–112, 2017.

[41] J. Shen, C. J. Zhang, B. Jiang, J. Chen, J. Song, Z. Liu, Z. He, S. Y. Wong, P.-H. Fang, W.-K. Ming, *et al.*, "Artificial intelligence versus clinicians in disease diagnosis: systematic review," *JMIR medical informatics*, vol. 7, no. 3, p. e10010, 2019.

[42] S. W. Chung, S. S. Han, J. W. Lee, K.-S. Oh, N. R. Kim, J. P. Yoon, J. Y. Kim, S. H. Moon, J. Kwon, H.-J. Lee, *et al.*, "Automated detection and classification of the proximal humerus fracture by using deep learning algorithm," *Acta orthopaedica*, vol. 89, no. 4, pp. 468–473, 2018.

[43] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and

S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[44] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. 1005 Gravenstein Highway North, Sebastopol, CA 95472: O'Reilly Media, Inc., 2 ed., 2019.

[45] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 5 2015.

[46] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[47] R. Xu and D. C. Wunsch, "Clustering algorithms in biomedical research: a review," *IEEE reviews in biomedical engineering*, vol. 3, pp. 120–154, 2010.

[48] M. R. Karim, O. Beyan, A. Zappa, I. G. Costa, D. Rebholz-Schuhmann, M. Cochez, and S. Decker, "Deep learning-based clustering approaches for bioinformatics," *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 393–415, 2021.

[49] B. Andreopoulos, A. An, X. Wang, and M. Schroeder, "A roadmap of clustering algorithms: finding a match for a biomedical application," *Briefings in bioinformatics*, vol. 10, no. 3, pp. 297–314, 2009.

[50] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[51] S. M. Gaynor, A. Bortsov, E. Bair, R. B. Fillingim, J. D. Greenspan, R. Ohrbach, L. Diatchenko, A. Nackley, I. E. Tchivileva, W. Whitehead, *et al.*, "Phenotypic

profile clustering pragmatically identifies diagnostically and mechanistically informative subgroups of chronic pain patients," *Pain*, vol. 162, no. 5, pp. 1528–1538, 2021.

[52] E. Bair, S. Gaynor, G. D. Slade, R. Ohrbach, R. B. Fillingim, J. D. Greenspan, R. Dubner, S. B. Smith, L. Diatchenko, and W. Maixner, "Identification of clusters of individuals relevant to temporomandibular disorders and other chronic pain conditions: the oppera study," *Pain*, vol. 157, no. 6, p. 1266, 2016.

[53] S. H. Ralston, I. D. Penman, M. W. Strachann, and R. P. Hobson, *Davidson's Principles and Practice of Medicine*. London, England: Elsevier Health Sciences, 23 ed., 2018.

[54] D. Vardeh, R. J. Mannion, and C. J. Woolf, "Toward a mechanism-based approach to pain diagnosis," *The Journal of Pain*, vol. 17, no. 9, pp. T50–T69, 2016.

[55] M. D. DiBonaventura, A. Sadosky, K. Concialdi, M. Hopps, I. Kudel, B. Parsons, J. C. Cappelleri, P. Hlavacek, A. H. Alexander, B. R. Stacey, *et al.*, "The prevalence of probable neuropathic pain in the us: results from a multimodal general-population health survey," *Journal of pain research*, vol. 10, p. 2525, 2017.

[56] N. B. Finnerup, S. Haroutounian, P. Kamerman, R. Baron, D. L. Bennett, D. Bouhassira, G. Cruccu, R. Freeman, P. Hansson, T. Nurmikko, *et al.*, "Neuropathic pain: an updated grading system for research and clinical practice," *Pain*, vol. 157, no. 8, p. 1599, 2016.

[57] S. P. Cohen and J. Mao, "Neuropathic pain: mechanisms and their clinical implications," *Bmj*, vol. 348, 2014.

[58] M.-A. Fitzcharles, S. P. Cohen, D. J. Clauw, G. Littlejohn, C. Usui, and W. Häuser, "Nociplastic pain: towards an understanding of prevalent pain conditions," *The Lancet*, vol. 397, no. 10289, pp. 2098–2110, 2021.

[59] C. J. Woolf, "Central sensitization: implications for the diagnosis and treatment of pain," *pain*, vol. 152, no. 3, pp. S2–S15, 2011.

[60] R. L. Chimenti, L. A. Frey-Law, and K. A. Sluka, "A mechanism-based approach to physical therapist management of pain," *Physical therapy*, vol. 98, no. 5, pp. 302–314, 2018.

[61] P. S. T. Chong and D. P. Cros, "Technology literature review: quantitative sensory testing," *Muscle & Nerve: Official Journal of the American Association of Electrodiagnostic Medicine*, vol. 29, no. 5, pp. 734–747, 2004.

[62] C. J. Woolf and M. B. Max, "Mechanism-based pain diagnosis: issues for analgesic drug development," *The Journal of the American Society of Anesthesiologists*, vol. 95, no. 1, pp. 241–249, 2001.

[63] R. Freynhagen, H. A. Parada, C. A. Calderon-Ospina, J. Chen, D. Rakhmawati Emril, F. J. Fernández-Villacorta, H. Franco, K.-Y. Ho, A. Lara-Solares, C. C.-F. Li, *et al.*, "Current understanding of the mixed pain concept: a brief narrative review," *Current medical research and opinion*, vol. 35, no. 6, pp. 1011–1018, 2019.

[64] P. J. Ibor, I. Sánchez-Magro, J. Villoria, A. Leal, and A. Esquivias, "Mixed pain can be discerned in the primary care and orthopedics settings in spain," *The Clinical Journal of Pain*, vol. 33, no. 12, pp. 1100–1108, 2017.

[65] M. A. Shraim, H. Massé-Alarie, L. M. Hall, and P. W. Hodges, "Systematic review and synthesis of mechanism-based classification systems for pain experienced in the musculoskeletal system," *The Clinical journal of pain*, vol. 36, no. 10, pp. 793–812, 2020.

[66] D. A. Seminowicz, T. H. Wideman, L. Naso, Z. Hatami-Khoroushahi, S. Fallatah, M. A. Ware, P. Jarzem, M. C. Bushnell, Y. Shir, J. A. Ouellet, *et al.*, "Effective treatment of chronic low back pain in humans reverses abnormal brain anatomy and function," *Journal of Neuroscience*, vol. 31, no. 20, pp. 7540–7550, 2011.

[67] R. J. Gatchel, D. D. McGeary, C. A. McGeary, and B. Lippe, "Interdisciplinary chronic pain management: past, present, and future.," *American psychologist*, vol. 69, no. 2, p. 119, 2014.

[68] F. Campbell, M. Hudspith, M. Choinière, H. El-Gabalawy, J. Laliberté, M. Sangster, J. Swidrovich, and L. Wilhelm, "Canadian pain task force report: March 2021." `https://www.canada.ca/en/health-canada/corporate/about-health-canada/public-engagement/external-advisory-bodies/canadian-pain-task-force/report-2021.html`, March 2021. Accessed on 2021-04-12.

[69] U. D. of Health, H. Services, *et al.*, "Pain management best practices inter-agency task force report: Updates, gaps, inconsistencies, and recommendations."

`https://www.hhs.gov/opioids/prevention/pain-management-options/index.html`, 2019. Accessed on 2021-03-10.

[70] T. Miettinen, P. Mäntyselkä, N. Hagelberg, S. Mustola, E. Kalso, and J. Lötsch, "Machine learning suggests sleep as a core factor in chronic pain," *Pain*, vol. 162, no. 1, pp. 109–123, 2021.

[71] J. Lötsch and S. Malkusch, "Interpretation of cluster structures in pain-related phenotype data using explainable artificial intelligence (xai)," *European Journal of Pain*, vol. 25, no. 2, pp. 442–465, 2021.

[72] E. Bäckryd, E. B. Persson, A. I. Larsson, M. R. Fischer, and B. Gerdle, "Chronic pain patients can be classified into four groups: Clustering-based discriminant analysis of psychometric data from 4665 patients referred to a multidisciplinary pain centre (a sqrp study)," *PLoS One*, vol. 13, no. 2, p. e0192623, 2018.

[73] I. Axén, L. Bodin, G. Bergström, L. Halasz, F. Lange, P. W. Lövgren, A. Rosenbaum, C. Leboeuf-Yde, and I. Jensen, "Clustering patients on the basis of their individual course of low back pain over a six month period," *BMC musculoskeletal disorders*, vol. 12, no. 1, pp. 1–10, 2011.

[74] S. D. Tagliaferri, T. Wilkin, M. Angelova, B. M. Fitzgibbon, P. J. Owen, C. T. Miller, and D. L. Belavy, "Chronic back pain sub-grouped via psychosocial, brain and physical factors using machine learning," *Scientific reports*, vol. 12, no. 1, pp. 1–15, 2022.

[75] R. Chimenti, "A mechanistic approach to pain management: Applying the biopsychosocial model to physical therapy - international association for the study

of pain (iasp)." https://www.iasp-pain.org/publications/relief-news/article/mechanistic-pain-management-biopsychosocial-model/, May 2018. Accessed on 2022-06-02.

[76] M. A. Khan, R. Koh, S. Hassan, T. Liu, V. Tucci, D. Kumbhare, and T. E. Doyle, "Star-ml: A rapid screening tool for assessing reporting of machine learning in research," in *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 336–341, Sep 2022.

[77] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, L. Weeks, *et al.*, "Prisma extension for scoping reviews (prisma-scr): checklist and explanation," *Annals of internal medicine*, vol. 169, no. 7, pp. 467–473, 2018.

[78] S. García, J. Luengo, and F. Herrera, *Data preprocessing in data mining*, vol. 72. Springer, 2015.

[79] P. S. Mann, *Introductory statistics*. John Wiley & Sons, 2007.

[80] T. R. Vetter, "Descriptive statistics: Reporting the answers to the 5 basic questions of who, what, why, when, where, and a sixth, so what?," *Anesthesia & Analgesia*, vol. 125, no. 5, pp. 1797–1802, 2017.

[81] D. Howitt and D. Cramer, *Introduction to statistics in psychology*. Pearson education, 2007.

[82] N. Robbins, "A histogram is not a bar chart." https://www.forbes.com/sites/naomirobbins/2012/01/04/a-histogram-is-not-a-bar-chart/?sh=7998ff3c6d77. Accessed on 2022-03-22.

[83] M. E. Magnello, "Karl pearson and the origins of modern statistics: An elastician becomes a statistician," *The New Zealand Journal for the History and Philosophy of Science and Technology*, vol. 1, 2005.

[84] "pandas.dataframe.corr — pandas 1.5.1 documentation." `https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html`. Accessed on 2022-04-02.

[85] "Correlation coefficient: Simple definition, formula, easy calculation steps." `https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/`. Accessed on 2022-04-02.

[86] C. C. Aggarwal *et al.*, *Data mining: the textbook*, vol. 1. Springer, 2015.

[87] A. M. Committee, "Robust statistics–how not to reject outliers. part 1. basic concepts," *Analyst*, vol. 114, pp. 1693–1697, 1989.

[88] J. Yang, S. Rahardja, and P. Fränti, "Outlier detection: How to threshold outlier scores?," in *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, AIIPCC '19, (New York, NY, USA), Association for Computing Machinery, 2019.

[89] H. Perez and J. H. M. Tah, "Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-sne," *Mathematics*, vol. 8, no. 5, 2020.

[90] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer, 2005.

[91] "seaborn.boxplot — seaborn 0.12.1 documentation." `https://seaborn.pydata.org/generated/seaborn.boxplot.html`. Accessed on 2022/04/21.

[92] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine*, vol. 4, no. 1, 2016.

[93] A. Kumar, "Python - replace missing values with mean, median & mode - data analytics." `https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/`, October 2021. Accessed on 2022-03-25.

[94] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[95] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[96] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 1990.

[97] G. Preud'homme, K. Duarte, K. Dalleau, C. Lacomblez, E. Bresso, M. Smaïl-Tabbone, M. Couceiro, M.-D. Devignes, M. Kobayashi, O. Huttin, *et al.*, "Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021.

[98] G. Szepannek, "clustmixtype: User-friendly clustering of mixed-type data in r.," *R J.*, vol. 10, no. 2, p. 200, 2018.

[99] L. A. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, no. 3, pp. 338–353, 1965.

[100] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[101] L. A. Zadeh, "Information and control," *Fuzzy sets*, vol. 8, no. 3, pp. 338–353, 1965.

[102] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[103] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons, 1999.

[104] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," Technical Report 2006-13, Stanford InfoLab, June 2006.

[105] "pyclustering: pyclustering.cluster.fcm.fcm class reference." `https://pyclustering.github.io/docs/0.10.1/html/d2/d6a/classpyclustering_1_1cluster_1_1fcm_1_1fcm.html#a401ef1e505a13488ea1515d1c570f6b2`. Accessed on 2022-07-28.

[106] X. Zhu and A. B. Goldberg, *Introduction to semi-supervised learning*, vol. 3. Morgan & Claypool Publishers, 2009.

[107] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised

methods," *33rd Annual Meeting on Association for Computational Linguistics*, p. 189–196, 1995.

[108] "1.14. semi-supervised learning — scikit-learn 1.1.3 documentation." `https://scikit-learn.org/stable/modules/semi_supervised.html`. Accessed on 2022-07-02.

[109] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[110] W. S. Noble, "What is a support vector machine?," *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[111] A. Shmilovici, *Support Vector Machines*, pp. 231–247. Boston, MA: Springer US, 2010.

[112] "1.4. support vector machines — scikit-learn 1.1.3 documentation." `https://scikit-learn.org/stable/modules/svm.html#svm`. Accessed on 2022-04-02.

[113] D. Meyer and F. Wien, "Support vector machines," *The Interface to libsvm in package e1071*, vol. 28, p. 20, 2015.

[114] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," *CMU-CALD-02-107*, June 2002.

[115] "sklearn.semi_supervised.labelpropagation — scikit-learn 1.1.3 documentation." `https://scikit-learn.org/stable/modules/generated/sklearn.`

semi_supervised.LabelPropagation.html#sklearn.semi_supervised.LabelPropagation. Accessed on 2022-06-21.

[116] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[117] "sklearn.semi_supervised.labelspreading — scikit-learn 1.1.3 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.semi_supervised.LabelSpreading.html#sklearn.semi_supervised.LabelSpreading. Accessed on 2022-06-21.

[118] "1.14. semi-supervised learning — scikit-learn 1.1.3 documentation." https://scikit-learn.org/stable/modules/semi_supervised.html#label-propagation. Accessed on 2022-06-21.

[119] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[120] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, eds., *Feature Extraction*. Princeton, NJ, USA: Springer-Verlag Berlin Heidelberg, 1 ed., 2006.

[121] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Medical Research Methodology*, vol. 19, no. 1, p. 64, 2019.

[122] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, and C. S.

Haley, "Application of high-dimensional feature selection: evaluation for genomic prediction in man," *Scientific reports*, vol. 5, p. 10312, 2015.

[123] A. Coenen and A. Pearce, "Understanding umap." `https://pair-code.github.io/understanding-umap/`. Accessed on 2022-06-12.

[124] N. Tavakoli, S. Siami-Namini, M. Adl Khanghah, F. Mirza Soltani, and A. Siami Namin, "An autoencoder-based deep learning approach for clustering time series data," *SN Applied Sciences*, vol. 2, no. 5, pp. 1–25, 2020.

[125] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[126] J. Brownlee, "How to reduce generalization error with activity regularization in keras - machinelearningmastery.com." `https://machinelearningmastery.com/how-to-reduce-generalization-error-in-deep-neural-networks-with-activity-regu` August 2020. Accessed on 2022-08-25.

[127] "tf.keras.callbacks.earlystopping — tensorflow v2.11.0." `https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping`. Accessed on 2022-07-28.

[128] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation.," in *Ijcai*, pp. 1753–1759, 2017.

[129] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *arXiv preprint arXiv:2003.05991*, 2020.

[130] R. Bellman and R. Kalaba, "A mathematical theory of adaptive control processes," *Proceedings of the National Academy of Sciences*, vol. 45, no. 8, pp. 1288–1290, 1959.

[131] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?," in *International conference on database theory*, pp. 217–235, Springer, 1999.

[132] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, pp. 478–487, PMLR, 2016.

[133] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[134] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[135] M. J. Warrens and H. van der Hoef, "Understanding the adjusted rand index and other partition comparison indices based on counting object pairs," *Journal of Classification*, pp. 1–23, 2022.

[136] "2.3. clustering — scikit-learn 1.1.3 documentation." `https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation`. Accessed on 2022-05-12.

[137] S. Wagner and D. Wagner, *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe, 2007.

[138] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, p. 2837–2854, Dec 2010.

[139] S. Romano, N. X. Vinh, J. Bailey, and K. Verspoor, "Adjusting for chance clustering comparison measures," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 4635–4666, 2016.

[140] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, (Prague, Czech Republic), pp. 410–420, Association for Computational Linguistics, June 2007.

[141] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[142] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[143] C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019.

[144] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, vol. 336, p. 012017, IOP Publishing, 2018.

[145] N. J. de Vos, "kmodes categorical clustering library." https://github.com/nicodv/kmodes, 2015–2021.

[146] "Developer interface — kprototypes 0.1.2 documentation." https://kprototypes.readthedocs.io/en/latest/api.html#main-interface. Accessed on 2022-05-12.

[147] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 747–748, IEEE, 2020.

[148] "3.2. tuning the hyper-parameters of an estimator — scikit-learn 1.1.3 documentation." https://scikit-learn.org/stable/modules/grid_search.html. Accessed on 2022-06-12.

[149] B. H. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pp. 1–8, 2019.

[150] "3.3. metrics and scoring: quantifying the quality of predictions — scikit-learn 1.1.3 documentation." https://scikit-learn.org/stable/modules/model_evaluation.html. Accessed on 2022-05-12.

[151] "sklearn.metrics.balanced_accuracy_score — scikit-learn 1.1.3 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html#sklearn.metrics.balanced_accuracy_score. Accessed on 2022-05-12.

[152] M. Heydarian, T. E. Doyle, and R. Samavi, "Mlcm: Multi-label confusion matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022.

[153] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[154] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[155] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[156] N. Spolaôr, E. A. Cherman, J. Metz, M. C. Monard, *et al.*, "A systematic review on experimental multi-label learning," *Institute of Mathematics and Computational Sciences*, 2013.

[157] M. S. Sorower, "A literature survey on algorithms for multi-label learning," *Oregon State University, Corvallis*, vol. 18, no. 1, p. 25, 2010.

[158] "Explainable artificial intelligence." `https://towardsdatascience.com/explainable-artificial-intelligence-14944563cc79`, Apr 2020. Accessed on 2021-04-28.

[159] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model

predictions," in *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.

[160] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[161] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89, IEEE, 2018.

[162] J. Dersh, P. B. Polatin, and R. J. Gatchel, "Chronic pain and psychopathology: research findings and theoretical considerations," *Psychosomatic medicine*, vol. 64, no. 5, pp. 773–786, 2002.

[163] L. A. McWilliams, B. J. Cox, and M. W. Enns, "Mood and anxiety disorders associated with chronic pain: an examination in a nationally representative sample," *Pain*, vol. 106, no. 1-2, pp. 127–133, 2003.

[164] L. R. Miller and A. Cano, "Comorbid chronic pain and depression: who is at risk?," *The journal of pain*, vol. 10, no. 6, pp. 619–627, 2009.

[165] S. Bruehl, R. Ohrbach, S. Sharma, E. Widerstrom-Noga, R. H. Dworkin, R. B. Fillingim, and D. C. Turk, "Approaches to demonstrating the reliability and validity of core diagnostic criteria for chronic pain," *The journal of pain*, vol. 17, no. 9, pp. T118–T131, 2016.

[166] R. H. Dworkin, S. Bruehl, R. B. Fillingim, J. D. Loeser, G. W. Terman, and D. C. Turk, "Multidimensional diagnostic criteria for chronic pain: introduction to the acttion–american pain society pain taxonomy (aapt)," *The Journal of Pain*, vol. 17, no. 9, pp. T1–T9, 2016.

[167] L. Diatchenko, A. G. Nackley, G. D. Slade, R. B. Fillingim, and W. Maixner, "Idiopathic pain disorders–pathways of vulnerability," *Pain*, vol. 123, no. 3, pp. 226–230, 2006.

[168] K. M. Smart, C. Blake, A. Staines, and C. Doody, "Clinical indicators of 'nociceptive','peripheral neuropathic'and 'central'mechanisms of musculoskeletal pain. a delphi survey of expert clinicians," *Manual therapy*, vol. 15, no. 1, pp. 80–87, 2010.

[169] K. M. Smart, C. Blake, A. Staines, M. Thacker, and C. Doody, "Mechanisms-based classifications of musculoskeletal pain: part 1 of 3: symptoms and signs of central sensitisation in patients with low back (±leg) pain," *Manual therapy*, vol. 17, no. 4, pp. 336–344, 2012.

[170] K. M. Smart, C. Blake, A. Staines, M. Thacker, and C. Doody, "Mechanisms-based classifications of musculoskeletal pain: part 2 of 3: symptoms and signs of peripheral neuropathic pain in patients with low back (±leg) pain," *Manual therapy*, vol. 17, no. 4, pp. 345–351, 2012.

[171] K. M. Smart, C. Blake, A. Staines, M. Thacker, and C. Doody, "Mechanisms-based classifications of musculoskeletal pain: part 3 of 3: symptoms and signs of nociceptive pain in patients with low back (±leg) pain," *Manual therapy*, vol. 17, no. 4, pp. 352–357, 2012.

[172] J. Nijs, A. Apeldoorn, H. Hallegraeff, J. Clark, R. Smeets, A. Malfliet, E. Lluch Girbes, M. De Kooning, and K. Ickmans, "Low back pain: guidelines for the clinical classification of predominant neuropathic, nociceptive, or central sensitization pain," *Pain physician*, vol. 18, no. 3, pp. E333–E345, 2015.

[173] M. C. Kolski, A. O'Connor, K. Van Der Laan, J. Lee, A. J. Kozlowski, and A. Deutsch, "Validation of a pain mechanism classification system (pmcs) in physical therapy practice," *Journal of Manual & Manipulative Therapy*, vol. 24, no. 4, pp. 192–199, 2016.

[174] M. J. Grant and A. Booth, "A typology of reviews: an analysis of 14 review types and associated methodologies," *Health information & libraries journal*, vol. 26, no. 2, pp. 91–108, 2009.

[175] Z. Munn, M. D. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris, "Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach," *BMC medical research methodology*, vol. 18, no. 1, pp. 1–7, 2018.

[176] A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, "The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration," *BMJ*, vol. 339, 2009.

[177] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, *et al.*, "The

prisma 2020 statement: an updated guideline for reporting systematic reviews," *Systematic reviews*, vol. 10, no. 1, pp. 1–11, 2021.

[178] D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman, "Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials.," *BMJ (Clinical research ed.)*, vol. 340, 2010.

[179] P. F. Whiting, A. W. Rutjes, M. E. Westwood, S. Mallett, J. J. Deeks, J. B. Reitsma, M. M. Leeflang, J. A. Sterne, and P. M. Bossuyt, "Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies," *Annals of Internal Medicine*, vol. 155, pp. 529–536, 2011.

[180] P. Whiting, A. W. Rutjes, J. B. Reitsma, P. M. Bossuyt, and J. Kleijnen, "The development of quadas: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews," *BMC Medical Research Methodology*, vol. 3, pp. 1–13, 11 2003.

[181] K. G. Moons, J. A. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D. G. Altman, J. B. Reitsma, and G. S. Collins, "Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The charms checklist," *PLOS Medicine*, vol. 11, p. e1001744, 2014.

[182] H. Arksey and L. O'Malley, "Scoping studies: towards a methodological framework," *International journal of social research methodology*, vol. 8, no. 1, pp. 19–32, 2005.

[183] P. Ayala, "Screening for articles - knowledge syntheses: Systematic &

scoping reviews, and other review types - research guides at university of toronto." https://guides.library.utoronto.ca/c.php?g=713309&p=5104946%E2%80%8B, 2020. Accessed on 2022/01/10.

[184] V. H. Innovation, "Covidence." https://app.covidence.org/. Accessed on 10/03/2022.

[185] D. Badger, J. Nursten, P. Williams, and M. Woodward, "Should all literature reviews be systematic?," *Evaluation & Research in Education*, vol. 14, no. 3-4, pp. 220–230, 2000.

[186] S. C. Almeida, S. Z. George, R. D. Leite, A. S. Oliveira, and T. C. Chaves, "Cluster subgroups based on overall pressure pain sensitivity and psychosocial factors in chronic musculoskeletal pain: differences in clinical outcomes," *Physiotherapy theory and practice*, 2018.

[187] S. J. Fodeh, D. Finch, L. Bouayad, S. Luther, R. D. Kerns, and C. Brandt, "Classifying clinical notes with pain assessment.," *Studies in Health Technology and Informatics*, vol. 245, pp. 1261–1261, 2017.

[188] J.-H. Kang, H.-S. Chen, S.-C. Chen, and F.-S. Jaw, "Disability in patients with chronic neck pain: heart rate variability analysis and cluster analysis," *The Clinical journal of pain*, vol. 28, no. 9, pp. 797–803, 2012.

[189] N. Pancino, C. Graziani, V. Lachi, M. L. Sampoli, E. Ștefănescu, M. Bianchini, and G. M. Dimitri, "A mixed statistical and machine learning approach for the analysis of multimodal trail making test data," *Mathematics*, vol. 9, no. 24, p. 3159, 2021.

[190] A. Rogachov, J. C. Cheng, K. S. Hemington, R. L. Bosma, J. A. Kim, N. R. Osborne, R. D. Inman, and K. D. Davis, "Abnormal low-frequency oscillations reflect trait-like pain ratings in chronic pain patients revealed through a machine learning approach," *Journal of Neuroscience*, vol. 38, no. 33, pp. 7293–7302, 2018.

[191] A. N. Santana, C. N. de Santana, and P. Montoya, "Chronic pain diagnosis using machine learning, questionnaires, and qst: A sensitivity experiment," *Diagnostics*, vol. 10, no. 11, p. 958, 2020.

[192] A. N. Santana, I. Cifre, C. N. De Santana, and P. Montoya, "Using deep learning and resting-state fmri to classify chronic pain conditions," *Frontiers in neuroscience*, vol. 13, p. 1313, 2019.

[193] O. Alge, S. R. Soroushmehr, J. Gryak, A. Kratz, and K. Najarian, "Predicting poor sleep quality in fibromyalgia with wrist sensors," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 4290–4293, IEEE, 2020.

[194] L. A. Antonucci, A. Taurino, D. Laera, P. Taurisano, J. Losole, S. Lutricuso, C. Abbatantuono, M. Giglio, M. F. De Caro, G. Varrassi, *et al.*, "An ensemble of psychological and physical health indices discriminates between individuals with chronic pain and healthy controls with high reliability: a machine learning study," *Pain and Therapy*, vol. 9, no. 2, pp. 601–614, 2020.

[195] E. Bagarinao, K. A. Johnson, K. T. Martucci, E. Ichesco, M. A. Farmer, J. Labus, T. J. Ness, R. Harris, G. Deutsch, A. V. Apkarian, *et al.*, "Preliminary structural

mri based brain classification of chronic pelvic pain: A mapp network study," *Pain®*, vol. 155, no. 12, pp. 2502–2509, 2014.

[196] E. Bair, S. Gaynor, G. D. Slade, R. Ohrbach, R. B. Fillingim, J. D. Greenspan, R. Dubner, S. B. Smith, L. Diatchenko, and W. Maixner, "Identification of clusters of individuals relevant to temporomandibular disorders and other chronic pain conditions: the oppera study," *Pain*, vol. 157, no. 6, p. 1266, 2016.

[197] J. Barroso, A. D. Vigotsky, P. Branco, A. M. Reis, T. J. Schnitzer, V. Galhardo, and A. V. Apkarian, "Brain grey matter abnormalities in osteoarthritis pain: a cross-sectional evaluation," *Pain*, vol. 161, no. 9, p. 2167, 2020.

[198] M. Behr, M. Noseworthy, and D. Kumbhare, "Feasibility of a support vector machine classifier for myofascial pain syndrome: diagnostic case-control study," *Journal of ultrasound in medicine*, vol. 38, no. 8, pp. 2119–2132, 2019.

[199] M. Behr, S. Saiel, V. Evans, and D. Kumbhare, "Machine learning diagnostic modeling for classifying fibromyalgia using b-mode ultrasound images," *Ultrasonic Imaging*, vol. 42, no. 3, pp. 135–147, 2020.

[200] J. Bianchi, A. C. de Oliveira Ruellas, J. R. Gonçalves, B. Paniagua, J. C. Prieto, M. Styner, T. Li, H. Zhu, J. Sugai, W. Giannobile, *et al.*, "Osteoarthritis of the temporomandibular joint can be diagnosed earlier using biomarkers and machine learning," *Scientific reports*, vol. 10, no. 1, pp. 1–14, 2020.

[201] D. Callan, L. Mills, C. Nott, R. England, and S. England, "A tool for classifying individuals with chronic back pain: using multivariate pattern analysis with

functional magnetic resonance imaging data," *PloS one*, vol. 9, no. 6, p. e98007, 2014.

[202] M. Caza-Szoka, D. Massicotte, F. Nougarou, and M. Descarreaux, "Surrogate analysis of fractal dimensions from semg sensor array as a predictor of chronic low back pain," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6409–6412, IEEE, 2016.

[203] M. Caza-Szoka, D. Massicotte, F. Nougarou, and M. Descarreaux, "Surrogate analysis of fractal dimensions from semg sensor array as a predictor of chronic low back pain," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6409–6412, IEEE, 2016.

[204] J. C. Cheng, A. Rogachov, K. S. Hemington, A. Kucyi, R. L. Bosma, M. A. Lindquist, R. D. Inman, and K. D. Davis, "Multivariate machine learning distinguishes cross-network dynamic functional connectivity patterns in state and trait neuropathic pain," *Pain*, vol. 159, no. 9, pp. 1764–1776, 2018.

[205] S. M. Gaynor, A. Bortsov, E. Bair, R. B. Fillingim, J. D. Greenspan, R. Ohrbach, L. Diatchenko, A. Nackley, I. E. Tchivileva, W. Whitehead, *et al.*, "Phenotypic profile clustering pragmatically identifies diagnostically and mechanistically informative subgroups of chronic pain patients," *Pain*, vol. 162, no. 5, pp. 1528–1538, 2021.

[206] J. Gudin, S. Mavroudi, A. Korfiati, K. Theofilatos, D. Dietze, and P. Hurwitz, "Reducing opioid prescriptions by identifying responders on topical analgesic

treatment using an individualized medicine and predictive analytics approach," *Journal of pain research*, vol. 13, p. 1255, 2020.

[207] S. E. Harte, E. Ichesco, J. P. Hampson, S. J. Peltier, T. Schmidt-Wilcke, D. J. Clauw, and R. E. Harris, "Pharmacologic attenuation of cross-modal sensory augmentation within the chronic pain insula," *Pain*, vol. 157, no. 9, p. 1933, 2016.

[208] K. F. Holton, A. E. Kirkland, M. Baron, S. S. Ramachandra, M. T. Langan, E. T. Brandley, and J. N. Baraniuk, "The low glutamate diet effectively improves pain and other symptoms of gulf war illness," *Nutrients*, vol. 12, no. 9, p. 2593, 2020.

[209] E. Ichesco, S. J. Peltier, I. Mawla, D. E. Harper, L. Pauer, S. E. Harte, D. J. Clauw, and R. E. Harris, "Prediction of differential pharmacologic response in chronic pain using functional neuroimaging biomarkers and a support vector machine algorithm: An exploratory study," *Arthritis & Rheumatology*, vol. 73, no. 11, pp. 2127–2137, 2021.

[210] D. Jiménez-Grande, S. F. Atashzar, E. Martinez-Valdes, A. M. De Nunzio, and D. Falla, "Kinematic biomarkers of chronic neck pain measured during gait: A data-driven classification approach," *Journal of Biomechanics*, vol. 118, p. 110190, 2021.

[211] D. Jiménez-Grande, S. F. Atashzar, E. Martinez-Valdes, and D. Falla, "Muscle network topology analysis for the classification of chronic neck pain based on emg biomarkers extracted during walking," *Plos one*, vol. 16, no. 6, p. e0252657, 2021.

[212] B. Lamichhane, D. Jayasekera, R. Jakes, W. Z. Ray, E. C. Leuthardt, and A. H. Hawasli, "Functional disruptions of the brain in low back pain: A potential imaging biomarker of functional disability," *Frontiers in Neurology*, vol. 12, 2021.

[213] B. Lamichhane, D. Jayasekera, R. Jakes, M. F. Glasser, J. Zhang, C. Yang, D. Grimes, T. L. Frank, W. Z. Ray, E. C. Leuthardt, *et al.*, "Multi-modal biomarkers of low back pain: A machine learning approach," *NeuroImage: Clinical*, vol. 29, p. 102530, 2021.

[214] B. Larsson, B. Gerdle, L. Bernfort, L.-Å. Levin, and E. Dragioti, "Distinctive subgroups derived by cluster analysis based on pain and psychological symptoms in swedish older adults with chronic pain–a population study (pains65+)," *BMC geriatrics*, vol. 17, no. 1, pp. 1–11, 2017.

[215] J. Lee, I. Mawla, J. Kim, M. L. Loggia, A. Ortiz, C. Jung, S.-T. Chan, J. Gerber, V. J. Schmithorst, R. R. Edwards, *et al.*, "Machine learning-based prediction of clinical pain using multimodal neuroimaging and autonomic metrics," *Pain*, vol. 160, no. 3, p. 550, 2019.

[216] J. Levitt, M. M. Edhi, R. V. Thorpe, J. W. Leung, M. Michishita, S. Koyama, S. Yoshikawa, K. A. Scarfo, A. G. Carayannopoulos, W. Gu, *et al.*, "Pain phenotypes classified by machine learning using electroencephalography features," *NeuroImage*, vol. 223, p. 117256, 2020.

[217] H. Mano, G. Kotecha, K. Leibnitz, T. Matsubara, C. Sprenger, A. Nakae, N. Shenker, M. Shibata, V. Voon, W. Yoshida, *et al.*, "Classification and characterisation of brain network changes in chronic back pain: A multicenter study," *Wellcome open research*, vol. 3, 2018.

[218] C. P. Mao, F. R. Chen, J. H. Huo, L. Zhang, G. R. Zhang, B. Zhang, and X. Q. Zhou, "Altered resting-state functional connectivity and effective connectivity of the habenula in irritable bowel syndrome: A cross-sectional and machine learning study," *Human brain mapping*, vol. 41, no. 13, pp. 3655–3666, 2020.

[219] T. Miettinen, P. M¨antyselk¨a, N. Hagelberg, S. Mustola, E. Kalso, and J. L¨otsch, "Machine learning suggests sleep as a core factor in chronic pain," *Pain*, vol. 162, no. 1, pp. 109–123, 2021.

[220] A. Minerbi, E. Gonzalez, N. J. Brereton, A. Anjarkouchian, K. Dewar, M.-A. Fitzcharles, S. Chevalier, and Y. Shir, "Altered microbiome composition in individuals with fibromyalgia," *Pain*, vol. 160, no. 11, pp. 2589–2602, 2019.

[221] J. Mo, J. Zhang, W. Hu, F. Luo, and K. Zhang, "Whole-brain morphological alterations associated with trigeminal neuralgia," *The journal of headache and pain*, vol. 22, no. 1, pp. 1–10, 2021.

[222] A. G. Morales, J. J. Lee, F. Caliva, C. Iriondo, F. Liu, S. Majumdar, and V. Pedoia, "Uncovering associations between data-driven learned qmri biomarkers and chronic pain," *Scientific reports*, vol. 11, no. 1, pp. 1–14, 2021.

[223] S. S. Olesen, C. Graversen, S. A. Bouwense, O. H. Wilder-Smith, H. van Goor, and A. M. Drewes, "Is timing of medical therapy related to outcome in painful chronic pancreatitis?," *Pancreas*, vol. 45, no. 3, pp. 381–387, 2016.

[224] O. Ozkan, M. Yildiz, E. Arslan, S. Yildiz, S. Bilgin, S. Akkus, H. R. Koyuncuoglu, and E. Koklukaya, "A study on the effects of sympathetic skin response

parameters in diagnosis of fibromyalgia using artificial neural networks," *Journal of medical systems*, vol. 40, no. 3, pp. 1–9, 2016.

[225] R. Pinedo-Villanueva, S. Khalid, V. Wylde, R. Gooberman-Hill, A. Soni, and A. Judge, "Identifying individuals with chronic pain after knee replacement: a population-cohort, cluster-analysis of oxford knee scores in 128,145 patients from the english national health service," *BMC musculoskeletal disorders*, vol. 19, no. 1, pp. 1–9, 2018.

[226] A. Richter, J. Truthmann, J.-F. Chenot, and C. O. Schmidt, "Predicting physician consultations for low back pain using claims data and population-based cohort data—an interpretable machine learning approach," *International journal of environmental research and public health*, vol. 18, no. 22, p. 12013, 2021.

[227] M. A. Russo, P. Georgius, A. S. Pires, B. Heng, M. Allwright, B. Guennewig, D. M. Santarelli, D. Bailey, N. T. Fiore, V. X. Tan, *et al.*, "Novel immune biomarkers in complex regional pain syndrome," *Journal of Neuroimmunology*, vol. 347, p. 577330, 2020.

[228] W. Shen, Y. Tu, R. L. Gollub, A. Ortiz, V. Napadow, S. Yu, G. Wilson, J. Park, C. Lang, M. Jung, *et al.*, "Visual network alterations in brain functional connectivity in chronic low back pain: A resting state functional connectivity and machine learning study," *NeuroImage: Clinical*, vol. 22, p. 101775, 2019.

[229] J.-G. Shim, K.-H. Ryu, E.-A. Cho, J. H. Ahn, H. K. Kim, Y.-J. Lee, and S. H. Lee, "Machine learning approaches to predict chronic lower back pain in people aged over 50 years," *Medicina*, vol. 57, no. 11, p. 1230, 2021.

[230] S. T. Dinh, M. M. Nickel, L. Tiemann, E. S. May, H. Heitmann, V. D. Hohn, G. Edenharter, D. Utpadel-Fischler, T. R. T¨olle, P. Sauseng, *et al.*, "Brain dysfunction in chronic pain patients assessed by resting-state electroencephalography," *Pain*, vol. 160, no. 12, p. 2751, 2019.

[231] K. Thieme, D. C. Turk, R. H. Gracely, W. Maixner, and H. Flor, "The relationship among psychological and psychophysiological characteristics of fibromyalgia patients," *The Journal of Pain*, vol. 16, no. 2, pp. 186–196, 2015.

[232] Y. Tu, F. Zeng, L. Lan, Z. Li, N. Maleki, B. Liu, J. Chen, C. Wang, J. Park, C. Lang, *et al.*, "An fmri-based neural marker for migraine without aura," *Neurology*, vol. 94, no. 7, pp. e741–e751, 2020.

[233] Y. Tu, A. Ortiz, R. L. Gollub, J. Cao, J. Gerber, C. Lang, J. Park, G. Wilson, W. Shen, S.-T. Chan, *et al.*, "Multivariate resting-state functional connectivity predicts responses to real and sham acupuncture treatment in chronic low back pain," *NeuroImage: Clinical*, vol. 23, p. 101885, 2019.

[234] K. Tuechler, E. Fehrmann, T. Kienbacher, P. Mair, L. Fischer-Grote, and G. Ebenbichler, "Mapping patient reported outcome measures for low back pain to the international classification of functioning, disability and health using random forests.," *European Journal of Physical and Rehabilitation Medicine*, vol. 56, no. 3, pp. 286–296, 2020.

[235] B. You, H. Wen, and T. Jackson, "Identifying resting state differences salient for resilience to chronic pain based on machine learning multivariate pattern analysis," *Psychophysiology*, vol. 58, no. 12, p. e13921, 2021.

[236] J. Zhong, D. Q. Chen, P. S.-P. Hung, D. J. Hayes, K. E. Liang, K. D. Davis, and M. Hodaie, "Multivariate pattern classification of brain white matter connectivity predicts classic trigeminal neuralgia," *Pain*, vol. 159, no. 10, pp. 2076–2087, 2018.

[237] J. Zhou, P. F. Damasceno, R. Chachad, J. R. Cheung, A. Ballatori, J. C. Lotz, A. A. Lazar, T. M. Link, A. J. Fields, and R. Krug, "Automatic vertebral body segmentation based on deep learning of dixon images for bone marrow fat fraction quantification," *Frontiers in endocrinology*, vol. 11, p. 612, 2020.

[238] J. Lin, L. Mou, Q. Yan, S. Ma, X. Yue, S. Zhou, Z. Lin, J. Zhang, J. Liu, and Y. Zhao, "Automated segmentation of trigeminal nerve and cerebrovasculature in mr-angiography images by deep learning," *Frontiers in Neuroscience*, p. 1684, 2021.

[239] J. L¨otsch, L. Alfredsson, and J. Lampa, "Machine-learning–based knowledge discovery in rheumatoid arthritis–related registry data to identify predictors of persistent pain," *Pain*, vol. 161, no. 1, pp. 114–126, 2020.

[240] A. Ounajim, M. Billot, L. Goudman, P.-Y. Louis, Y. Slaoui, M. Roulaud, B. Bouche, P. Page, B. Lorgeoux, S. Baron, *et al.*, "Machine learning algorithms provide greater prediction of response to scs than lead screening trial: a predictive ai-based multicenter study," *Journal of clinical medicine*, vol. 10, no. 20, p. 4764, 2021.

[241] P. S.-P. Hung, J. Y. Zhang, A. Noorani, M. R. Walker, M. Huang, J. W. Zhang, N. Laperriere, F. Rudzicz, and M. Hodaie, "Differential expression of a brain aging biomarker across discrete chronic pain disorders," *Pain*, pp. 10–1097, 2022.

[242] L. Frostholm, C. Hornemann, E. Ørnbøl, P. Fink, and M. Mehlsen, "Using illness perceptions to cluster chronic pain patients," *The Clinical Journal of Pain*, vol. 34, no. 11, pp. 991–999, 2018.

[243] Q. A. Rahman, T. Janmohamed, M. Pirbaglou, H. Clarke, P. Ritvo, J. M. Heffernan, and J. Katz, "Defining and predicting pain volatility in users of the manage my pain app: analysis using data mining and machine learning methods," *Journal of medical Internet research*, vol. 20, no. 11, p. e12001, 2018.

[244] Q. A. Rahman, T. Janmohamed, H. Clarke, P. Ritvo, J. Heffernan, and J. Katz, "Interpretability and class imbalance in prediction models for pain volatility in manage my pain app users: analysis using feature selection and majority voting methods," *JMIR medical informatics*, vol. 7, no. 4, p. e15601, 2019.

[245] C. Wang, T. A. Olugbade, A. Mathur, A. C. D. C. Williams, N. D. Lane, and N. Bianchi-Berthouze, "Chronic pain protective behavior detection with deep learning," *ACM Transactions on Computing for Healthcare*, vol. 2, no. 3, pp. 1–24, 2021.

[246] "University health network." `https://www.uhn.ca/`. Accessed on 2022-04-05.

[247] "Dados platform - dados electronic data capture platform." `https://www.dadosproject.com/`. Accessed on 2022-11-28.

[248] "Mcmaster research ethics board (mreb)." `https://research.mcmaster.ca/ethics/mcmaster-research-ethics-board-mreb/`. Accessed on 2021-08-25.

[249] C. Cleeland *et al.*, "Measurement of pain by subjective report," *Advances in pain research and therapy*, vol. 12, pp. 391–403, 1989.

[250] "Brief pain inventory - short form - physiopedia." https://www.physio-pedia.com/Brief_Pain_Inventory_-_Short_Form. Accessed on 2022-11-28.

[251] R. D. Kerns, R. Rosenberg, R. N. Jamison, M. A. Caudill, and J. Haythornthwaite, "Readiness to adopt a self-management approach to chronic pain: the pain stages of change questionnaire (psocq)," *Pain*, vol. 72, no. 1-2, pp. 227–234, 1997.

[252] D. C. Tollison and J. C. Langley, "Pain patient profile manual," *Minneapolis: National Computer Services*, 1995.

[253] G. van Rossum and J. de Boer, "Interactively testing remote servers using the python programming language," *CWI quarterly*, vol. 4, no. 4, pp. 283–303, 1991.

[254] "sklearn.preprocessing.labelencoder — scikit-learn 1.1.3 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html. Accessed on 2022/04/21.

[255] C. del Coso, D. Fustes, C. Dafonte, F. J. Nóvoa, J. M. Rodríguez-Pedreira, and B. Arcay, "Mixing numerical and categorical data in a self-organizing map by means of frequency neurons," *Appl. Soft Comput.*, vol. 36, p. 246–254, nov 2015.

[256] S. Tsang, C. F. Royse, A. S. Terkawi, *et al.*, "Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine," *Saudi journal of anaesthesia*, vol. 11, no. 5, p. 80, 2017.

[257] "Home — wsib." https://www.wsib.ca/en. Accessed on 2022-07-28.

[258] "umap-learn · pypi." https://pypi.org/project/umap-learn/. Accessed on 2022-05-28.

[259] "1.14. semi-supervised learning — scikit-learn 1.2.0 documentation." https://scikit-learn.org/stable/modules/semi_supervised.html. Accessed on 2022-04-28.

[260] "tensorflow · pypi." https://pypi.org/project/tensorflow/. Accessed on 2022-07-12.

[261] M. L. D. Dias, "fuzzy-c-means: An implementation of fuzzy c-means clustering algorithm.," May 2019.