A Personalized, Uncertainty-Aware, Trustworthy Algorithm for Effective Pain Assessment using Biosignals

XINWEI JI

Doctor of Philosophy

Supervisor: Professor Albert Y. Zomaya
Associate Supervisor: Doctor Wei Li

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

1 July 2024

Authorship Attribution Statement

Chapter 3 of this thesis is a data collection study. I initiated the study, designed the experiment, conducted the analysis and experiments, and this study will be made public when I decide to make the data public.

Chapter 4 of this thesis is published as [1]. I initiated the study, designed the algorithms, conducted the analysis, and wrote the drafts of the manuscript.

Chapter 5 of this thesis is published as [2]. I initiated the study, designed the algorithms, conducted the analysis and experiments, and wrote the drafts of the manuscript.

Chapter 6 of this thesis is published as [3]. I designed the algorithms, conducted the analysis and experiments, and wrote the drafts of the manuscript.

I certify that the aforementioned authorship attribution statements are correct and I have received permission from the other authors to include the published materials.

Student Name: Xinwei Ji
Signature:
Date:

As the supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name: Albert Y. Zomaya
Signature:
Date:

Publications

[1] Ji, X., Shi, J., Li, W., and Zomaya, A. Y. (2024). Pain Personalization via Test Time Adaptation. In *Thirty-eighth Annual Conference on Neural Information Processing Systems*. Ready for Submission.

[2] Ji, X., Chang, X., Li, W., and Zomaya, A. Y. (2024, March). Unraveling Pain Levels: A Data-Uncertainty Guided Approach for Effective Pain Assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 20, pp. 22167-22175).

[3] Ji, X., Zhao, T., Li, W., and Zomaya, A. (2023, March). Automatic Pain Assessment with Ultra-short Electrodermal Activity Signal. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing* (pp. 618-625).

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Name: Xinwei Ji

Signature: **Date:**

Student Plagiarism: Compliance Statement

I certify that:

I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure;

I understand that failure to comply with the Student Plagiarism: Coursework Policy and Procedure can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the University of Sydney By-Law 1999 (as amended);

This Work is substantially my own, and to the extent that any part of this Work is not my own I have indicated that it is not my own by Acknowledging the Source of that part or those parts of the Work.

Name: Xinwei Ji

Signature: **Date:**

Abstract

Automatic pain assessment algorithms are used to improve pain assessment and assist subsequent pain treatment and management for patients without healthcare provider supervision. This thesis proposes a new pain assessment framework called "A Personalized, Uncertainty-Aware, Trustworthy Algorithm for Effective Pain Assessment using Biosignals." The framework takes into account the uncertainty of the data itself and the strong subjectivity of the pain experience, utilizing heart rate variability analysis to assess data uncertainty and test time adaptation to deal with distribution drift. It considers that pain data is imperfect, that there are data-label inconsistencies, and that the personalization of pain prediction algorithms is important. Our aim is to create complete frameworks for automated pain assessment that reduce the complexity of algorithms while predicting well. We collected experimental pain data and data from real pain patients, including post-surgical patients and women in labor. Through experiments and analyses, the framework outperforms state-of-the-art methods.

Acknowledgements

I would like to thank my supervisors, Professor Albert Y. Zomaya and Dr Wei Li, for their guidance, support and encouragement throughout my PhD studies. They created an extraordinary academic environment for my research and provided me with all the necessary support along the way. Without them, today's research would not have been possible. I would also like to thank Dr Hongyi Zhu for helping me during my academic research. I thank my mother for giving me the freedom to make my own life choices and for supporting my choices. I would also like to thank Dr Wei Li's wife, Esther Sha Li, for helping me when I was most confused and helpless. I would like to thank my research partners - Tianming Zhao, Xiaomin Chang, Yunchuan Shi, Chunqiu Xia, thank you for your support in making the lab work run smoothly and for the good times we had together. Last but not least, I would like to thank my ex-girlfriend, Yuanhui Zhang, for being there for me during the toughest times, which led to the completion of the most important paper of my PhD.

This research reported in this thesis was supported by the award of a Research Training Program scholarship to the PhD Candidate.

Contents

List of Figures

CHAPTER 1

Introduction

Automated pain assessment relies on algorithms for pain level prediction using the patient's own data, a process that does not require subjective patient involvement. The main application scenarios include the post-surgical coma phase or in the ICU ward. This paper focuses on building a framework for automatic pain assessment using the patient's physiological signals as input to the algorithm. In current clinics, subjective-based automated reporting is still the main method of pain assessment, but this method is very time-consuming and laborious and is not feasible for patients with speech dysfunction. The objective automatic pain assessment algorithm was developed with the goal of serving as an assistant tool to close the loop of pain management and treatment. This step is essential because of the patient's need for pain management and limited medical resources.

There are three main challenges in the development of automated pain assessment algorithms, the first being pain data, which is often more difficult to publicize due to its experimental specificity and the private nature of patient data, which has resulted in a very limited publicly available pain dataset. The second is the uncertainty of the data, whether it is experimental pain data or real pain patient data, whether the data collected is the patient's facial expression or physiological signals or a combination of the two, the data itself has a certain degree of noise, and there is a certain degree of error in the correspondence between the data and the label. The third one is that pain is highly variable between individuals due to the complexity of pain itself and the fact that personality judgments of different pains are strongly subjective.

A number of studies have attempted different approaches to address the three previously mentioned problems, focusing mainly on the analysis of data from real pain patients and

the application of tree-based machine learning algorithms. These approaches are rather fragmented and do not constitute an overall framework. This motivates the thesis, which develops a new framework called *A Personalized, Uncertainty-Aware, Trustworthy Algorithm for Effective Pain Assessment using Biosignals*.

The inputs to the proposed pain assessment framework are biosignals due to their relative objectivity compared to facial expressions. Considering the natural ordering carried by pain labels, the algorithm designs an ordered regression neural network as an inference model by converting the ordered pain level prediction problem into a series of binary classification problems. Within this framework, we substitute methods for data uncertainty assessment and domain generalization with the aim of developing a personalized pain assessment algorithm as well as taking data uncertainty into account. In addition to this, a hierarchical pain monitoring system is proposed with the aim of reducing the computational power requirements of medical hardware. This thesis first describes current research advances and methods in pain assessment algorithms, focusing mainly on the study of pain algorithms based on physiological signals. Chapter 3 focuses on the data collection study; in this chapter, we are doing three different types of pain data collection by collaborating with healthcare companies and hospitals, which are thermal stimulation pain data, maternal pain data, and postoperative patient pain data. The technical chapters make the following contributions.

In Chapter 4, we introduce a personalized pain intensity inference model for automated pain assessment in clinical settings. This method distinguishes itself in the domain of pain personalization by facilitating both online and passive assessments of pain, demonstrating efficacy comparable to approaches utilizing source data. Moreover, we have integrated a novel test-time adaptation strategy aimed at minimizing the predictive entropy within ordinal regression neural networks, thereby enhancing the precision and reliability of assessments. Our algorithms were developed across two datasets: the public BioVid and the private Apon, achieving state-of-the-art results on the public dataset. Further validation of our approach's effectiveness was conducted through its application on real patients, affirming the potential of our methodology for practical medical scenarios.

In Chapter 5, we introduce DUG-CORAL, an innovative approach for automatic pain assessment that marries the concept of pain data uncertainty with the CORAL ordinal regression algorithm. This method marks the first exploration into utilizing heart-related indicators and derived heart rate variability (HRV) features for estimating individual data uncertainty, subsequently integrating this uncertainty into the training process of an electrodermal activity (EDA)-based pain prediction model. To tackle the inherent ambiguity in pain labeling, we implement a task-importance weighting strategy during the loss calculation phase, aiming to refine the model's accuracy. Our extensive experiments conducted on the pain datasets, BioVid (public) and Apon (private), showcase our method's superior performance over traditional approaches. Additionally, we extend our validation to real pain patients, further confirming the robustness and effectiveness of our approach in practical clinical scenarios

In Chapter 6, we present a novel machine-learning framework designed to identify pain and assess pain intensity utilizing ultra-short segments of physiological signals, approximately 5 seconds in length. This framework demonstrates superior performance across various metrics, outpacing the current state-of-the-art solutions. Our research further delves into the potential of exclusively using Electrodermal Activity (EDA) and its derived features for the purpose of automatic pain assessment, pioneering a simplified yet effective approach. In addition to theoretical development, we have designed and executed experiments involving heat-induced pain stimuli to empirically validate the feasibility and accuracy of automatic pain assessment. Our methodologies are rigorously compared with leading-edge solutions using both publicly available data and data collected by us. An extensive ablation study is conducted to underline the effectiveness of our design, providing comprehensive insights into the contributions of different components of our framework to its overall performance.

Many of the chapters in this thesis use a large number of similar symbols and notations. However, due to unavoidable differences in problem settings, each chapter assumes a separate set of symbols to avoid reader confusion. The second chapter has to do with pain data collection and is the basis for algorithm development to be carried out later. The third and fourth chapters are devoted to two key issues in pain algorithm development, one being

FIGURE 1.1: Thesis Outlines and relationship between chapters

uncertainty and the other subjectivity. The fifth chapter is a study of the efficiency and usability of the algorithm for applications in realistic scenarios.

Pain has always been one of the main reasons for patients to seek medical help, and since 2013, when the first publicly available pain dataset, BioVid, was introduced, the development of pain prediction algorithms has attracted the attention of both computer scientists and healthcare practitioners, not only in terms of pain rating scores, but also in a number of areas, such as the assessment of pain sensitivity, the assessment of the level of exposure to injuries, the assessment of the pain perception state, and so on, with a number of achievements. The increasing power of learning models has led to higher predictive capabilities. These models are becoming accurate and lightweight, and AI predictive modeling can go a long way in helping physicians in decision-making. The authors believe that AI-assisted pain assessment and management must be the right direction for the future.

CHAPTER 2

Current Research Advances and Methods in Pain Assessment

Pain management has always been a major problem in the field of clinical medicine and the main factor for many patients to seek medical services. More than 40% of postoperative patients' pain has not been fully controlled. For these patients, the pain has a serious impact on their life and psychology. During the outbreak of the Wuhan epidemic in 2019, the closure of the pain management department caused great challenges to the pain management of patients with severe chronic pain. Many patients with chronic pain could not seek medical help face-to-face during the isolation period, which led to the inability of patients to get timely and effective pain management.

Pain assessment is the basis of pain management. At present, there are two main methods of clinical pain assessment: self-report and doctor's observation. However, these two methods have obvious defects respectively. In the real medical scene, many patients can't give autonomous reports with clear cognition when they feel pain. It has been found that the doctor's observation method usually underestimates the patient's real pain level, and the objective observation method is always labor-intensive and impractical in pain monitoring. Therefore, the automatic (intelligent) pain assessment tool is an essential component in the future pain management system.

In the past ten years, with the development of machine learning and artificial intelligence, automatic pain assessment tools based on facial expressions, physiological signals, and behavior seem to be feasible solutions for patients with communication difficulties. At present, the pain assessment method for dementia patients is still to use objective observation assessment tools. In 2021, research by the University of California in the United States showed

that the change in skin conductivity was an important index that reflected pain. However, in their research, the pain of patients was artificially aroused, not naturally caused by patients.

The International Association for the Study of Pain (IASP) conceptualizes pain as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage." This definition underscores the multifaceted nature of pain, delineating at least three integral dimensions: unpleasantness, sensory experiences, and emotional responses. Each dimension contributes equally to the patient's pain experience, rendering evaluations that focus on a single aspect both incomplete and insufficient. In the future, the main obstacle to applying AI to automated pain assessment remains pain datasets; as it stands, the majority of publicly available pain datasets are from experimental pain rather than real patients with pain, and even when real pain data collection has been performed in some studies, the data are not publicly available. Second, pain data tend to be labeled with more homogeneous metrics rather than a composite score of pain concepts, which leads to inconsistency in pain data and labeling.

2.1 Patients Self-Reporting

The pain rating scale measures how badly a patient suffers in pain. The measure often comes from a patient's self-reporting. Some common measurements are the Verbal Rating Scale (VRS), Visual Analogue Scale (VAS), and Numerical Rating Scale (NRS). With VRS, one reports pain intensity by choosing the best adjective to reflect the feeling. With VAS, the pain intensity is decided by a point on a 100mm line. With NRS, one reports pain intensity by deciding an interval level on an 11, 21, or 101-point box scale where two endpoints represent no pain and worst pain (Williamson and Hoggart, 2005).

In clinical settings, the Visual Analog Scale (VAS) is prevalently utilized for pain assessment, requiring patients to indicate their pain level along a continuum between two endpoints. The distance from the starting point to the mark indicates the patient's self-reported pain intensity. Despite its widespread use, the VAS method has notable limitations. Firstly, it fails to fully capture the ordinal nature of pain as an emotional experience, potentially preventing

patients from indicating extreme pain levels that surpass their tolerance threshold. Secondly, the subjective nature of pain assessment remains a challenge. Lastly, the assumption that individuals in severe pain can accurately gauge and communicate their pain intensity is problematic.

Pain scoring inherently represents a latent, continuous variable, yet research frequently employs discretized scales—ranging from "slight" to "severe" or from "no pain" to "as painful as possible"—to quantify the pain intensity. These methodologies often rely directly on VAS scores for quantification, overlooking the nuanced complexities of pain experiences. Given these considerations, there is a substantial global demand for objective pain assessment. The burgeoning field of AI-assisted pain management offers considerable promise, heralding advancements in the precision and understanding of pain evaluation and treatment strategies.

2.2 Objective Questionnaire

Most observational scales consider facial expressions, vocalizations, and body language, while some scales also include vital signs. Assessing and comparing the validity of various scales is difficult because studies vary widely in terms of design, methodology, subjects, and conceptualization of pain phenomena. Significant prior training and experience is required to reliably apply a scale. Even if trained medical personnel are able to record pain intensity by observation several times a day, such frequent measurements may be minimized under financial pressure unless cost savings are shown to be possible. As a result, relevant pain events may be missed, or human observers may detect changes later(Werner *et al.*, 2019).

2.3 Electrodermal Activity

Electrodermal Activity (EDA), a.k.a. Galvanic Skin Response or Skin Conductivity Level (SCL), is the sympathetic activity from eccrine sweat glands on the skin. The EDA signal is sensitive to psychological processes (Dawson *et al.*, 2017). EDA has been used in multiple psychology studies, for example, stress evaluation (Jiang and Chen, 2021) and emotion

classification (Shukla *et al.*, 2019). EDA is also easy to collect (Banganho *et al.*, 2021). Many commercial wearable sensors, such as Emptica E4 wristband (Ollander *et al.*, 2016), can continuously monitor skin conductivity changes.

2.4 Pain Assessment with EDA Signal

As a sensitive pain modality, EDA-based algorithms have been developed for automatic pain assessment. Existing methods fall into three categories: (1) statistical methods, (2) machine learning with hand-crafted features, and (3) deep learning models that extract features and build models in a black-box manner.

The work (Treister *et al.*, 2012) defines four levels in pain intensity, and the pain intensity increases along with the level number. While the patients are tested in thermal stimuli, their body signals, photoplethysmography, SCL, and electrocardiogram (ECG), are collected. The data is then analyzed using the Friedman test and post-hoc Wilcoxon test. As a finding, a linear model of SCL variables can separate no pain from others ($p < 0.001$). It also predicts the pain levels ($p < 0.001$ to 0.02). In addition, the work (Sugimine *et al.*, 2020) applies the one-way analysis of variance with post-hoc Student–Newman–Keuls test on EDA signal. The result shows that normalized SCL tells non-physical pain stimuli, including noisy auditory and visual stimuli, which makes one think of the pain and physical pain stimuli apart.

The work (Chu *et al.*, 2017) uses various basic statistical features (maximum, minimum, median, and others) to represent 30-second ECG signals collected from the middle of electrical stimuli of 1 minute. The authors use multiple machine learning algorithms on the concatenated feature sets and observe that the models can predict pain states induced by four electrical stimulus levels. The work (Susam *et al.*, 2021) develops an SVM-based classification schema, which uses the fusion of EDA signal and video data to detect clinically significant pain from clinically nonsignificant pain in children.

A variety of multi-modal pain assessment algorithms have been proposed in recent years. The work (Thiam *et al.*, 2021) proposes novel multi-modal deep learning approaches based

on bio-signals (EDA, electromyogram (EMG), ECG, and respiratory rate) and evaluates the models on BioVid with an accuracy of 35.44% and a mean absolute error of 0.97 on five-level pain intensity prediction task. The work (Thiam *et al.*, 2019) uses a convolutional neural network to achieve end-to-end learning schema in developing pain assessment algorithms. The experimental result presents an accuracy of 36.54% on a five-class classification task using the fusion of EDA, EMG, and ECG.

The success of these works opens the door for using EDA signals only for pain assessment. However, all these works either use EDA and other physiological signals as inputs or a long EDA signal for the model. We argue that these solutions still have room to improve on the efficiency and medical device dependencies.

2.5 Automatic Pain Assessment Using Biosignals

Bio-signals have shown significant potential for pain assessment in numerous studies, with skin conductance response and heart rate/heart rate variability being the most characterized autonomic responses (Chae *et al.*, 2022). Differential characteristics of EDA have been identified as sensitive indices for classifying experimental pain stimulation induced by electric pulse (Kong *et al.*, 2021), heating (Pouromran *et al.*, 2021), and cold pressor test (Winslow *et al.*, 2022). Furthermore, normalized skin conductance level can distinguish physical pain stimuli from other sympathetic stimuli (Sugimine *et al.*, 2020). ECG data has achieved an 81.9% F1 score for acute pain classification in laboratory and clinical settings (Winslow *et al.*, 2022). Studies have also validated the use of electrodermal activity for postoperative pain assessment in real patients (Aqajari *et al.*, 2021), and the effectiveness of pain assessment using EDA and video on children following laparoscopic appendectomy (Susam *et al.*, 2021). These studies underscore the potential of pain assessment tools in real medical scenarios. However, the transition of these learning methods from experimental to clinical settings is yet to be achieved, and their datasets are not publicly available due to privacy concerns.

2.6 Conclusion

In this section, we present prevalent clinical pain assessment methods that have proven their effectiveness based on subjective pain reports. However, they also have practical limitations, such as the requirement for verbal communication from the patient and the need for human intervention by healthcare workers, among others. We present the progress in the development of EDA signals and current algorithms for automated pain assessment using EDA, demonstrating their research results. These findings demonstrate the great potential of the EDA as an objective automated pain assessment metric, but to verify its true validity, pain researchers still need to bring it into the clinic, a process that requires the combined efforts of patients, physicians, and researchers. Research on objective pain markers other than EDA (e.g., facial expressions and other physiological signals) has also yielded promising results. The fusion of this range of objective indicators may be able to eliminate to some extent the effects of non-specificity of a single indicator, which will lay the foundation for future automated pain applications that can be practically applied to real pain patients.

CHAPTER 3

Data collection study

3.1 Experimental Pain Data Collection Study

We conducted a data collection study of an experimental electrical stimulation-induced pain model in collaboration with Alpine Medical. In this study, we recruited a total of 60 volunteers, and excluding invalid data, the total sample size totaled 1363 observations. Among them, 519 observations were mild pain, 517 observations were moderate pain, and 312 observations were severe pain. The experiment is divided into 4 phases: experiment preparation phase, pre-laboratory phase, experiment phase and data organization phase.

3.1.1 Experiment Background

Pain assessment is an essential step in pain management, and currently there are two main methods of clinical pain assessment: autonomous report and physician observation; however, the current methods of pain assessment have obvious flaws respectively. In real medical scenarios, many patients are unable to give cognitively clear autonomous reports when they feel pain. Physician observation has been found to underestimate the patient's true pain and is always labor-intensive, making it impractical for continuous pain monitoring. Therefore, automated pain assessment tools are an essential component of future pain management systems.

3.1.2 Experimental Purpose

- Exploring the efficacy of electrodermal activity (EDA) as an important metric in automated pain assessment
- Exploring the Optimal Time Window for EDA Signals to Assess Pain Levels in Thermal Stimulation Simulated Pain

3.1.3 Experimental Apparatus

- Thermal stimulation equipment/electrical stimulation equipment
- Data acquisition equipment (EDA sensor, data acquisition software)

3.1.4 Experimental Steps

The experiment is divided into four stages: experiment preparation stage, pre-experiment stage, experiment stage and data organization stage.

3.1.5 Experimental Preparation Stage

- experimenter introduces the background, purpose, process and results of the experiment to the volunteers, and answers the questions raised by the volunteers and records them.
- Volunteers fill in the personal information form and volunteer consent form.
- Commissioning the experimental instruments

3.1.6 Pre-experiment Stage

Objective: To find out the pain threshold (mild, moderate and severe) of each volunteer for thermal stimulation. Experimental parameters:

- The hot electrode is fixed on the skin of the forearm.
- the initial temperature is 32°C (baseline)

- The temperature rises and falls at 8°C per second.
- The thermal stimulation time is 5 seconds. When the temperature is reached, the calculation begins, excluding the time for temperature rise and fall.
- The stimulation interval is 30 seconds.
- The maximum temperature is set to 50°C
- Measurement of pain assessment: VAS

3.1.7 Pre-experiment Process

After everything is ready, the experimenter can start the experiment, adjust the temperature to 32°C for 30 seconds to ensure that the patient has no pain at this temperature, and the first pain stimulation is performed after 30 seconds, so that the experimenter needs to record the VAS score of the patient after each pain stimulation. Pre-experiment will be divided into two groups: step-by-step group and cross-over group. Each time, the step-by-step type group takes precedence over the cross-repetition type. The initial temperature of step-by-step type is 46°C, the temperature gradually increases, and each stimulation increases by 0.5°C. When the temperature reaches 50°C, the experimental group is terminated. Based on the VAS score given by the volunteers with the initial temperature of 46°C in the step-by-step group, the selection of temperature in the cross-iteration group can be alternated in different temperature ranges, and the thermal stimulation temperature of each time can be limited within the range, with no fixed characteristic temperature. In the repeated group, the continuous stimulation temperature will alternate continuously to ensure that there will be no pain memory effect.

3.1.7.1 Analysis on Pre-experiment Data

The total number of participants in the pre-experiment was 60. The purpose of the pre-experiment was to find out each subject's temperature threshold for mild, moderate and severe pain. All subjects completed 2 groups (progressive and crossover) of pain assessment experiments. In order to select subjects with more reasonable pain assessment individuals, we depicted the mean VAS scores at each temperature based on the pre-experimental data of each subject. On this basis, a reasonable temperature threshold was manually manually

labeled reasonable temperature thresholds and excluded subjects with significant errors in pain scores under incremental thermal stimulation conditions, of which 28 Subject 28, whose data showed over-sensitivity to temperature, and subject 38, whose data appeared to be significantly abnormal, were excluded.

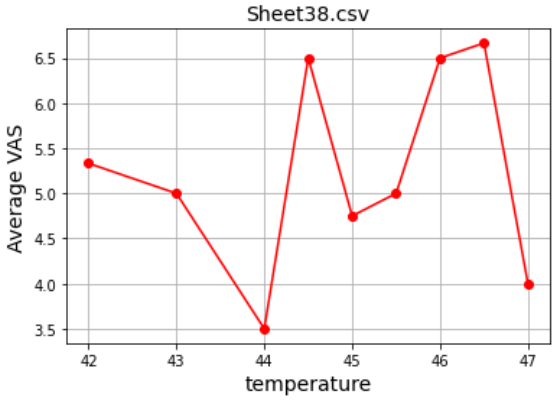


FIGURE 3.1: Distribution of VAS score (Subject 38)

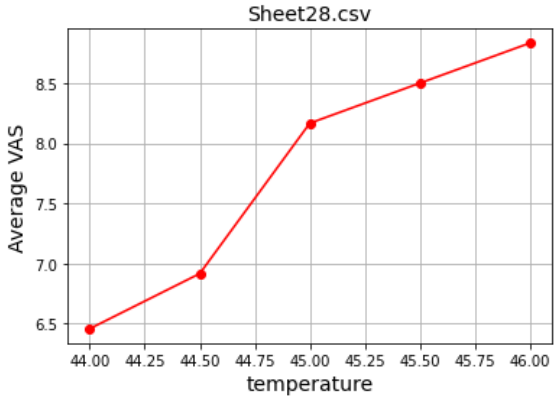


FIGURE 3.2: Distribution of VAS score (Subject 28)

Participants in the formal experiment: 58, subjects will have 10 pain stimuli at each pain threshold in the formal experiment.

Mild, moderate, severe	Mild, Moderate	moderate, severe	Mild, severe
34	9	3	2

Distribution of subjects' bmi (body mass index) As can be seen from the graph, most of the subjects showed normal levels of body mass index and very few subjects had bad indices such as overweight and obesity.

Distribution of pain sensitivity According to the results of the pain sensitivity questionnaire, the pain sensitivity of the subject population basically satisfied the normal distribution and met the experimental conditions.

Pain Catastrophizing Scale The overall distribution was left-skewed according to the Pain Catastrophizing Scale statistics, and the exact experimental impact remains to be further investigated.

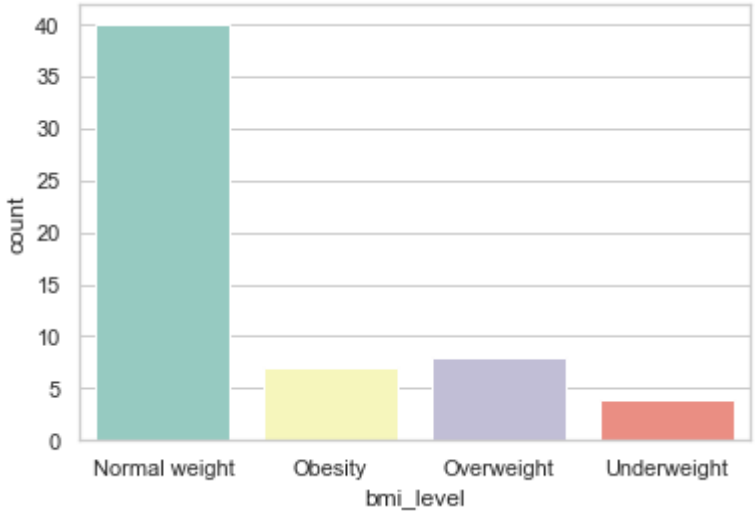


FIGURE 3.3: Distribution of subjects' bmi (body mass index)

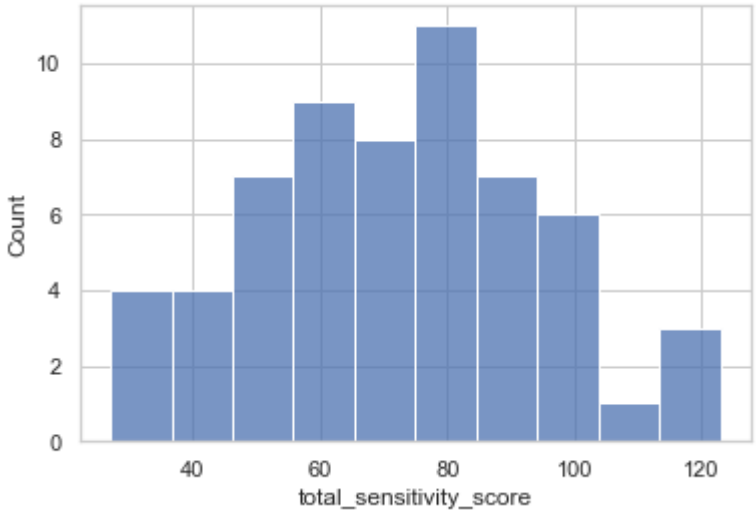


FIGURE 3.4: Distribution of pain sensitivity

3.1.8 Experiment Stage

By calculating the average value of volunteers' scores of different pain intensity in the pre-experiment, the average value of temperature corresponding to the pre-experiment is calculated here. This temperature value is used as the simulated temperature value of volunteers with different pain levels in the formal experiment.

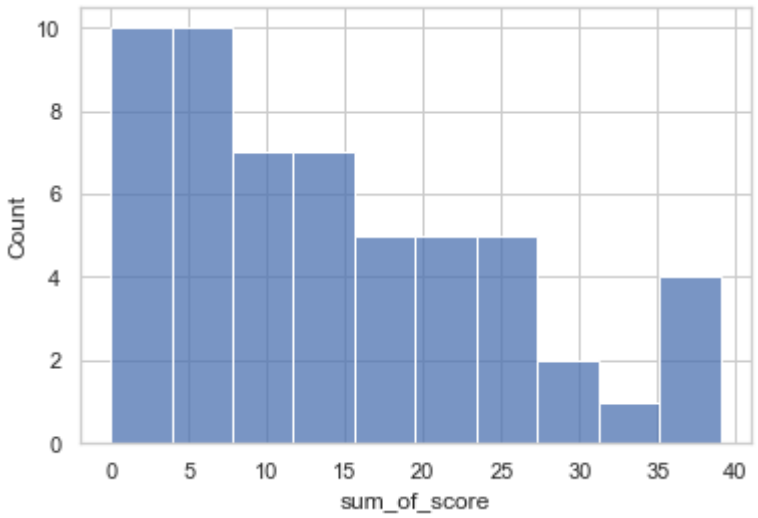


FIGURE 3.5: Pain Catastrophizing Scale

In the formal experiment, the experimental parameters are exactly the same as those in the pre-experiment. Each stimulus will randomly have a pain intensity, and the temperature corresponding to the volunteer is the current stimulus temperature. Each volunteer experienced 50 stimuli, and the total time was about 30 minutes.

3.1.9 Data Organization Stage

- Collect, sort out and back up the experimental data.
- Volunteers fill in the pain sensitivity questionnaire and the pain disaster scale questionnaire.

3.1.10 Experimental Caveats

- Before the start of all experiments, it is necessary to repeatedly test the pain stimulation equipment to make sure that there is no accident such as sudden temperature rise.
- In the formal experiment, volunteers need to perform pain simulation test in an unattended environment

- Try to ensure that volunteers have no strong personal emotions such as happiness and anxiety when participating in the experiment.
- Try to ensure that volunteers do experiments under stress-free conditions.
- Ensure the laboratory temperature and suitable lighting conditions.

3.1.11 Analysis

Whether there is a statistically significant difference in pain scores between subjects at the same temperature. To test whether there is a significant difference in pain scores among different objects at the same temperature, we can conduct a Friedman test for each temperature level. This is because the data appears to be paired (multiple scores at each temperature), and the Friedman test is suitable for non-parametric repeated measures data.

Results: According to the experimental results, there are 48 data points at 42.0°C, 40 data points at 43.0°C, 42 data points at 43.5°C, 60 data points at 44.0°C, 60 data points at 44.5°C, 60 data points at 45.0°C, 60 data points at 45.5°C, 60 data points at 46.0°C, 34 data points at 46.5°C, and 19 data points at 47.0°C. Temperatures with less than 10 data points are disregarded.

The results of the Friedman test showed the statistical significance of pain scores at each temperature. The p-values for most temperatures are far less than 0.05, indicating that there are significant differences in pain scores among different subjects at these temperatures. This means that for these specific temperatures, different subjects may experience different intensities of pain.

Statistical difference in subjects' ratings for the same temperature. Each subject underwent six heat pain stimuli at the same temperature and provided six pain ratings.

Average number of identical ratings	3.81	0.635
Mean number of different ratings	2.19	0.375
Mean difference in ratings	0.57	

This shows that the number of pain scores giving the same rating on 6 occasions with the same temperature stimulus was 3.81, the number of different number of ratings at 2.19 and the mean difference between ratings at 0.57.

Segmentation of data according to pain intervals. the total number of samples are 516.

- **Primary data**, samples satisfy Six scoring ties (not including 0 points included), the number of samples are 102, the percentage is 19.76%.
- **Secondary data**, 1) Four data with the same score (not including 0), six data in the same interval (light, medium, heavy); 2) Six Data Convergence in a Pain Interval. the number of samples are 57, the percentage is 11.04%.
- **Tertiary data**, 1) Four data with the same score (not including 0), and two others not in the same pain interval; 2) Four data converge (are in the same interval) and the other two are in adjacent intervals of this interval. the number of samples are 158, the percentage is 23.64%.

3.2 Clinical Pain Data Collection Study

The goal of this study is to investigate the pattern of pain and whether there is a clear correlation between pain and electrical signals in human skin by collecting and analyzing data from different types of subjects on pain manifestations and signs of stress.

This study was approved by the Institutional Review Board (IRB) of the Ethics Committee at Shanghai First Maternity and Infant Hospital. The IRB approval number is KS21179, and the approval was granted on March 1, 2023. The approved research protocol, which outlines the objectives, methodology, anticipated risks, and benefits of the study, was strictly followed. All participants provided informed consent prior to their inclusion in the study. The research team, led by Hongyi Zhu, Vice-president of the Jiangsu Apon Medical Technology Co.,Ltd., can be contacted for further information at hongyi.zhu@gmail.com.

The trial was randomized. Screened subjects were enrolled, and data on stress signs were collected from different types of subjects (e.g., distinguishing between "trauma patients", "laboring mothers", "ICU thoracic surgery patients", etc.) in the subject group. In the group of subjects, data were collected on different types of stress signs (e.g. "trauma patient", "laboring mother", "ICU thoracic surgery patient", etc.), with the following requirements: the presence of a strong pain process and a significant change in the pain level over a certain period of time. To investigate whether there is a clear correlation between pain and human skin electrical signals, and whether it is feasible to investigate the idea of indirectly characterizing the degree of stress pain in humans by human fingertip skin electrical signals.

3.2.1 Pain Patients

Acute trauma patients. Acutely traumatized patients are a category of patients with a high level of pain. Before the patient is admitted to the hospital without intervention, data collection of heart rate, skin conductance, body surface temperature, blood pressure, and evaluation of pain signs (including pain level) is performed for about 3-5 minutes. A second collection of the same types of signs and pain evaluation was performed during wound management and analgesia (depending on the range of the collection device). Before the patient left the hospital at the end of treatment, a final collection of the same type of sign data and pain evaluation were performed as baseline data.

Maternity pain data. The collaboration between Apon Medical Company and Shanghai First Maternity and Infant Healthcare Hospital (SFMICH) has resulted in the collection of data on the pain generated by contractions in pregnant women. So far, the dataset contains 63 observations, and the average length of the data is about 30 minutes.

Mothers in natural labor will experience regular contractions in the first and second stage of labor, with obvious pain in the contraction state, pain relief in the inter-contraction stage, and pain reduction after the implementation of labor analgesia, which will continue until the end of the labor for women in completely natural labor. After the laboring woman entered the labor room to wait for labor, we observed and selected the women who had regular contractions,

and collected the data of heart rate, skin conductance, body surface temperature, and blood pressure for the first 3-5 minutes during the intervals, and evaluated the pain signs (including the degree of pain), and collected the data of the second group of the same type of signs at the stage of obvious pain in the period of contractions, and carried out the pain evaluation. When labor entered the third stage of labor and the uterine opening was close to full, the third data collection of the same type of signs and pain evaluation were performed. At the end of labor, when maternal signs are stable and there is no pain (e.g., before leaving the hospital), a fourth set of the same type of signs is collected as baseline data (which can be strongly influenced by medications such as oxytocin).

Postoperative patient pain data . Apon Medical Company and Shanghai No. 4 People's Hospital collaborated to collect pain data from postoperative patients in scenarios including before anesthesia, in the recovery room, and in the hospital room. So far, there are 94 data items of different pain levels, and the average sampling time is 7 minutes.

Thoracic surgery targets diseases such as lung cancer, esophageal cancer, mediastinal tumors, and chest wall deformities, and these surgeries are the most severe types of surgical procedures in terms of postoperative pain due to their extensive and traumatic nature. One study reported that timely video-assisted thoracoscopic surgery still resulted in moderate to severe pain in 78% of postoperative patients: 27% with severe pain, 34% with severe pain, and 17% with very severe pain. Before the patients entered the operating room to prepare for the administration of anesthesia, the first 3-5 minutes of data collection of heart rate, skin conductance, body surface temperature, blood pressure, and evaluation of pain phenotypes (including pain level) were performed as baseline data for the subjects. After the procedure, in the observation/resuscitation room, when the patient has largely regained consciousness, a second data collection of the same type of signs is performed, as well as pain evaluation (possibly for 2 days in the ICU). Depending on the patient's second pain level, as well as the level of care evaluation, the third and fourth same type of sign data collection, and pain evaluation were performed in the ward before and after analgesic measures were taken.

3.2.2 Data Acquisition Equipment

The eVu TPS is a lightweight and small portable sensor that can be used to measure biometric data (heart rate variability, body surface temperature, skin conductance) and import the data into corresponding connected smartphones, tablets and laptops. The product has been validated by the FDA and obtained the FDA Medical Device Registration Certificate. The product consists of a portable sensor, two fabric straps, micro USB charging cable, and a carrying case; the sensor is mounted on a finger via a fabric strap and detects and transmits three highly researched psychophysiological health measures: heart rate variability, skin conductance, and body temperature. Data is transmitted in real time via Bluetooth to a companion app (phone, tablet or laptop).

3.2.3 Physiological Signals

Heart rate variability. At rest, a healthy heartbeat speeds up with inhalation and slows down with exhalation. These momentary variations between successive heartbeats are a good measure of neurological health in terms of physical, emotional and mental function. Research has shown that high heart rate variability is associated with better recovery from exercise, greater resilience to psychophysical obstacles, and a positive mood and outlook on life. Stress, especially prolonged stress, decreases heart rate variability. Training to increase HRV through slow-paced, relaxed breathing can reduce the effects of stress on the nervous system, thereby reducing negative psychophysiological problems.

Skin conductance. Emotional arousal affects the pores on the surface of our skin, which in turn affects the subtle changes in sweating. Conductance measurements of the skin surface are a biometric reflection of our mental activity and stress perception.

Body Surface Temperature. Cold hands may not only be due to cold weather, but may also be the body's response to elevated anxiety and stress. As a response to stressful moments of tension, the body shunts blood away from the fingers. The logic behind this is that blood will be needed for more important systems associated with the fight or flight response. Repeated

and sustained stress causes the body to remain in a fight-or-flight state, diverting heat-rich blood away from the hand, thereby lowering the temperature of the fingers. When the body eventually begins to relax in response, the blood flow is allowed to return to the peripheral portions of the hand, resulting in the heating of the fingers.

3.2.4 Experimenal Design

Subject selection:

- Age 18 years- 75 years (both 18 years and 75 years) and gender;
- Patients with moderate to severe pain, both pain VAS scores of 4 to 7;
- Patients who require analgesic treatment as assessed by a physician;

Specific types of subjects, different types of subjects corresponding to the intervention method and data collection and evaluation of each different. Subject types include:

- Acutely traumatized patients
- Mothers who chose to have a natural birth (or implement labor analgesia) with regular contractions
- Patients undergoing ICU thoracic surgery

Exclusion Criteria:

- Patients whose fingers are unable to wear fingertip electrical signal sensors;
- Patients with sensory dysfunction;
- Patients with hearing and speech disorders;
- People with allergy to anesthetic drugs;
- People with fabric strap allergies;

Other conditions that the investigator considers inappropriate for enrollment.

3.2.5 Evaluation Method

Visual analogue scale (VAS). VAS is a strip scale with no markings on the patient's side and 1 - 100 mm graduations on the physician's side, with "no pain" at one end and "most severe pain" at the other end, which is marked by the patient according to the intensity of the pain, and the physician determines the score.

The Wong-Baker face pain rating scale. The Wong-Baker Face Pain Rating Scale (FPRS) consists of 6 pictograms of different facial expressions ranging from smile or happiness to tears. This method is suitable for people with communication difficulties, such as children, the elderly, patients who are unconscious or unable to express themselves accurately in words, but it is easily affected by emotional, cultural, educational and environmental factors, and should be used in accordance with the specific situation.

Richmond Agitation-Sedation Score (RASS score).

Score	Description	Definition
+4	Combative	Overtly combative or violent; immediate danger to staff
+3	Very Agitated	Pulls or removes tubes or catheters; aggressive
+2	Agitated	Frequent non-purposeful movement, fights ventilator
+1	Restless	Anxious but movements not aggressive or vigorous
0	Alert and Calm	-
-1	Drowsy	Not fully alert, but has sustained awakening (eye-opening/eye contact) to voice (>10 seconds)
-2	Light Sedation	Briefly awakens with eye contact to voice (<10 seconds)
-3	Moderate Sedation	Movement or eye opening to voice (no eye contact)
-4	Deep Sedation	No response to voice, but movement or eye opening to physical stimulation
-5	Unarousable	No response to voice or physical stimulation

TABLE 3.1: Richmond Agitation-Sedation Scale (RASS)

3.3 Conclusion and Future Work

This section describes the details of the experimental design and data collection. The collected pain datasets were named Apon (experimental pain), Apon-postoperative (postoperative

patients), and Apon-labor (pregnant women). It was found that the data quality was higher for the experimental pain data because environmental variables in real medical scenarios are relatively difficult to control compared to laboratory settings. In the next technical chapters, these three datasets will be used to validate the effectiveness of automated pain assessment algorithms. Currently, there are very limited publicly available pain datasets, and the development and public availability of pain datasets is a great contribution to the overall research in the field of pain. We plan to make our pain datasets publicly available in the future and collaborate with more hospitals to collect pain data.

CHAPTER 4

Pain Personalization via Test Time Adaptation

Clinical management of pain in patients typically relies on accurate and sustainable methods of pain assessment. However, existing clinical pain assessment methods necessitate patient self-reporting to achieve personalized evaluations, which not only fails to capture continuous changes in pain during treatment but is also impractical for patients who are unable to provide self-reports. To address this issue, we propose a personalized pain assessment learning framework tailored to clinical use scenarios, which consists of three stages: pain model training, pre-test fine-tuning, and test-time adaptation. To better describe the predicted results, we use an ordinal regression neural network by transforming ordinal targets into binary classification subtasks to fit the natural order of pain labels. We also designed an unsupervised loss function to adapt all the model parameters by minimizing the entropy of the model's output distribution during the testing phase for new pain patients, thus allowing the model to adapt better to new and unseen data distributions. Our method was developed on two experimental pain datasets and validated on two additional real-world pain datasets, proving the effectiveness of our approach.

4.1 Introduction

In pain assessment and management, exploring pain management plans that are relevant to the individual patient is essential to clinical practice, which is also seen as the next logical step to accelerate innovation and development in the pain research field (Kearns, 1989). In the real world, as the first and most important part of pain management, pain assessment uses the individual patient's self-report as the gold standard (Yang *et al.*, 2021; Werner *et al.*,

2019). However, this subjective approach to pain assessment 1) Requires the patients to be cognitively competent, but some special groups of pain patients, such as infants, people with dementia or speech disorders, are unable to provide subjective reports, 2) Subjective reports of pain are not easily detected and quantified on an ongoing basis, which makes tracking pain changes and treatment outcomes difficult, 3) Pain is a complex and dynamic experience, and self-reports can vary over time, even under similar conditions, leading to inconsistent assessments (Hyun *et al.*, 2022). Therefore the development of objective pain assessment tools to assist physicians in patient pain management is an urgent need.

Combining machine learning and patient-controlled analgesic pumps demonstrates practical future scenarios for automatic pain assessment (Wang *et al.*, 2020b). There have been a number of research results showing the significant potential exhibited by biosignals as inputs to pain assessment algorithms, with the electrical skin activity response and heart rate/heart rate variability being the most characterized autonomic responses (Chae *et al.*, 2022), and in an experimental setting, the EDA can be recognized as a difference feature, and in combination with the SVR algorithm, the results show that it can distinguish experimental pain well (Pouromran *et al.*, 2021). Recently, the existing automatic pain assessment solution based on machine learning is now in the clinical validation phase, the study validated the effectiveness of postoperative pain assessment in real patients using random forests as a predictive model and electrical skin stimulation activity as a feature (Aqajari *et al.*, 2021), as well as the effectiveness of using EDA and video to assess pain in children after laparoscopic appendectomy (Susam *et al.*, 2021). These studies emphasize the potential of pain assessment tools in real-world medical scenarios; however, these traditional machine learning methods do not take into account the subjective nature of pain.

Subjectivity cannot be brought into the model as a feature due to the fact that it cannot be quantified, but personalization of the model is an option to solve this problem. From the perspective of the degree of personalization, personalization approaches in the field of pain research can be divided into two categories: 1) Fully personalized, that is, the source data when training the model and the target domain when testing it are the same subject. 2) Non-fully personalized, that is, the training process of the model may include data from other subjects

that are different from the test subject, and the effect of discriminating between personalised differences is achieved by capturing and quantifying those personal traits that may influence an individual's perception and expression of pain. Currently, there are already studies that bring personalization into the development of pain assessment algorithms, however, these methods require extra individuals' private information, such as inter-individual differences (complexion, age, gender and race) (Uddin *et al*., 2023; Liu *et al*., 2017) and past pain experience (Casti *et al*., 2020), require large model with more parameters (Lopez-Martinez and Picard, 2017) and multiple personal models (Xu and de Sa, 2021), or require extra steps, such as finding similar subjects (cluster-based) (Kächele *et al*., 2016), analysing bio-signal information (Jiang *et al*., 2024).

Therefore, we propose a stage-by-stage personalized learning framework for pain assessment called PCOR-PA. We consider the natural ordering of pain labels and the interpretability of prediction results, and we adopt COnsistent RAnk Logits (CORAL), which is a rank-consistent ordinal regression neural network where the probabilities decrease consistently, to serve as the inference model structure. Considering the subjectivity of pain, combined with the availability of data from pain patients in real medical scenarios, we fine-tuned the pre-trained model using data from a time when pain did not occur and designed a new unsupervised loss function to minimize the entropy of the output distribution of the CORAL model at the time the model was tested on data from a real occurrence of pain. The main objective is to improve the generalization ability of the model by reducing the uncertainty of the model on the test data without retraining the model.

To summarize, there are three main areas of our contribution:

- A personalized pain intensity inference model for automated pain assessment in the clinic. In terms of pain personalization, our method allows for online and passive pain assessment and is even comparable to methods that use source data
- A novel approach to test-time adaptation is incorporated to minimize the predictive entropy of ordinal regression neural networks.

- Our algorithms are developed on two datasets, BioVid (public) and Apon (private), on the public dataset we achieved the state of the art results. In addition, we take the developed algorithms further on real patients to validate the effectiveness of our approach.

4.2 Related Work

4.2.1 Personalization Techniques in the field of Pain

According to IASP's six keynotes on pain, the first one is that "Pain is always a personal experience that is influenced to varying degrees by biological, psychological, and social factors". Pain is a subjective, unpleasant sensory and emotional experience that is highly variable between individuals (Spisak *et al*., 2020). Nowadays, patient self-reporting is defined as the "gold standard" in pain assessment (Yang *et al*., 2021; Werner *et al*., 2019). This also demonstrates the significant role of personalized technology in the context of automated pain assessment algorithms. In reality, due to the privacy of medical data, researchers cannot obtain sufficient personal information for quantification; secondly, as pain is a subjective sensation, its perception and expression vary from person to person, influenced by factors such as age, gender, cultural background, and previous pain experiences. All these barriers make it very challenging to establish a universal model capable of accurately assessing pain for everyone.

Currently, in the field of pain, machine learning all provides a powerful tool in the identification of pain phenotypes (Chesler *et al*., 2002), the assessment of treatment efficacy (Lötsch *et al*., 2017), or in the analysis of clinical records (Patterson *et al*., 2015), in addition, some of the latest technologies are capable of personalized inference and prediction. These machine learning algorithms not only help doctors discover important information that may be overlooked but also lay the foundation for customized medicine and precision medicine (Lötsch and Ultsch, 2018).

Early work found that electrical changes in the skin may be used as an objective indicator to differentiate between pain and discomfort in infants (Munsters *et al*., 2012). Recent advances

in personalized pain assessment have leveraged both internal and external contexts, including health conditions and physiological activities, along with inter-individual differences such as age, gender, and race, to tailor pain prediction models more closely to individual needs (Uddin *et al.*, 2023; Jiang *et al.*, 2024).

The application of multi-task learning, especially through hard parameter sharing, has been explored to personalize output results in pain assessment (Lopez-Martinez and Picard, 2017). This approach is complemented by the design of personalized pain assessment platforms that incorporate visual, voice, and physiological cues, utilizing data such as frame-by-frame video analysis of a patient's past pain experiences for training fully personalized pain assessment models (Casti *et al.*, 2020).

Moreover, the fusion of hand-crafted personal features (complexion, age, and gender) with multi-task learning has been demonstrated in automated pain assessment algorithms based on facial expressions, offering a path towards greater personalization (Liu *et al.*, 2017). Similarly, the use of video data from individual subjects to train personal models, coupled with Tikhonov regularization to integrate predicted variance, further advances the field of personalized pain management (Xu and de Sa, 2021). The integration of personalized information, such as individuals' heart rate signals, into pain recognition and management systems, marks a significant step forward in creating more sensitive and individualized pain assessment tools (Jiang *et al.*, 2024).

4.2.2 Test Time Adaptation

Domain adaptation assumes that we have a labeled training set (source domain), and an unlabeled test set (target domain), and that unlabeled data from the target domain is utilized during model training to improve the model's ability to adapt between domains. With this in mind, domain generalization further weakens this assumption by making data from the target domain unavailable during model training, and the goal of domain generalization is to create a model with strong generalization capabilities that can be applied to any unseen target

domain. However, the most difficult aspect of domain generalization is also the unavailability of the target domain, which is not available during training.

In the past two years, there have been a series of test time adaptive methods assuming that test data comes in the form of online, attempting to allow pre-trained models to undergo target domain adaptation during the testing phase in order to leverage their information to enhance generalizability. Test-Time adaptive methods includes two main categories: Test-time training (TTT) and Test-time adaptation (TTA). TTT methods focus on additional training of the model at test time to adapt to the current task or data distribution. For example, the work (Zhang *et al*., 2021) propose a new multivariate expert learning strategy is proposed that utilizes self-supervised learning to aggregate multiple experts to deal with unknown test distributions and the work (Sun *et al*., 2020) propose a self-supervised learning based approach to solve the problem of distributional bias in test images. the test samples are augmented to form a batch. Then, the entire batch is used to update the model through self-supervised loss. The updated model is then utilized to make predictions on the current test samples.

TTA approaches do not change the training process, but usually combines pre-trained models and test cases, including entropy minimisation (Wang *et al*., 2020a; Zhang *et al*., 2022), Batchnorm Statistics Adaptation (Schneider *et al*., 2020; Lim *et al*., 2023) and Prototype-Based Method (Iwasawa and Matsuo, 2021). A non-parametric test time adaptive approach is proposed that does not require any gradient update, which solves the pressure of computational overhead. By using a nonparametric KNN classifier, instead of the traditional linear layer, the domain divergence is explicitly reduced (Zhang *et al*., 2023).

Test Time Adaptation focuses on adapting models to real-time data as they are used, aiming to improve their performance in the current environment. This approach has emerged to enhance model performance during the test phase and to address the issue of newcomers, providing a personalized approach and offering a novel perspective on crafting solutions.

4.3 Method

4.3.1 Problem settings & Notations

In this study, we have compiled a comprehensive dataset of physiological signals, denoted as D^s, with the aim of analyzing and understanding the characteristics of pain-related signals. The dataset encompasses signals from S distinct subjects, each with their series of biopotential signals and corresponding pain level annotations.

For each subject j, where j ranges from 1 to S, we define a subset of the data D_j^s that contains N_{sj} pairs of signals and their labels. This subset is formally represented as:

$$D_j^s := \left\{(x_{ij}^s, y_{ij}^s)\right\}_{i=1}^{N_{sj}}$$

where x_{ij}^s is the i-th bio-signal from subject j, and y_{ij}^s is the corresponding pain level label. The complete training dataset D^s is the union of all individual subject subsets (Multi-Source Domain):

$$D^s := \bigcup_{j=1}^{S} D_j^s$$

Let $X^t := (x^t(\tau))_{\tau \in \mathcal{T}}$, where $\mathcal{T}$ represents the entire time domain of a new coming subject (target domain).

To focus on a specific segment of this data stream within a time interval from a to b, which are points in $\mathcal{T}$, we define the segment as $X_{[a,b]}^t := (x^t(\tau))_{\tau \in [a,b]}$. The objective is to learn a prediction rule $f(\cdot;\theta)$, parameterized by θ, that can accurately predict the label $y_{[a,b]}^t$ for each segment $X_{[a,b]}^t$ in X^t, where $X_{[a,b]}^t$ represents a segment of the continuous data stream X^t within the time interval $[a, b]$, and $y_{[a,b]}^t$ is the corresponding label for this data segment.

Using traditional machine learning methods, we usually train on the source dataset to find the optimal model parameters θ^s, which is under the inductive setting. However, in the pain

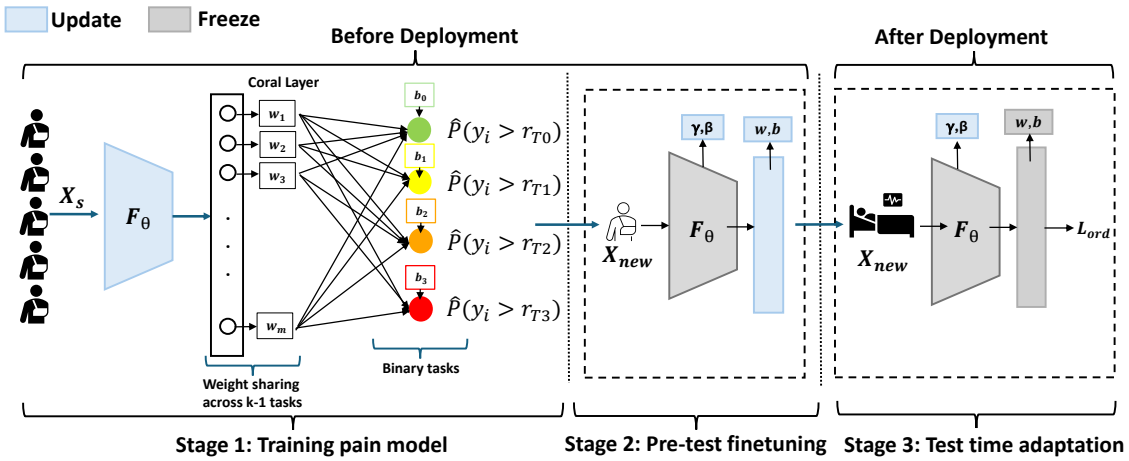


FIGURE 4.1: Overview of PCOR-PA

problem, the distribution of physiological data of the trained and tested subjects may change, the main reason of which is the individual variability of pain. In order to solve this problem, the test time adaptive methods, under transductive learning setting, has been first proposed by the works (Sun *et al.*, 2020; Wang *et al.*, 2020a). Specifically, Test time adaptive method first trains a model $f(\cdot;\theta^s)$, and then adapts to the test data during inference X^t. In this case, the data distribution of the current testing subject is progressively updated and learned by the model to improve performance.

4.3.2 Training schema

Our method comprises three stages. The first stage involves training a predictive model using a training dataset composed of data from multiple source domains. We standardize the data from each domain separately. The predictive model employs an ordered regression neural network, incorporating the natural ordering of pain labels, drawing upon previous studies. The second stage, referred to as the fine-tuning stage, addresses the clinical realities of pain. Before deploying the model for an individual test patient, we can obtain "no pain" data for this patient, which is data labeled as "0". In this stage, we utilize this data for supervised learning, primarily adjusting the parameters of the batch normalization layer. In the final stage, when testing our model on test patients, we update the model using unsupervised learning. Specifically, for each test batch, we calculate a prediction entropy loss and minimize this loss

to enhance the model's generalization capability to the current test domain. In this stage, we update both the batch normalization layer and the final linear layer. This approach is due to the fact that, while the Empirical Risk Minimization (ERM) method learns sufficient feature representations, the final linear layer often overfits the source domain (Rosenfeld *et al.*, 2022). This is why we choose to adjust the parameters of the final layer during testing. The decision to update the batch normalization layer is motivated by our aim to address the distribution shift problem (Segu *et al.*, 2023; Chang *et al.*, 2019).

Algorithm 1 Three-Stage Training and Adaptation Process

Input: Training dataset $\mathcal{D}_{train}$ comprising multiple patient domains, Test dataset $\mathcal{D}_{test}$ from a single patient domain
Output: Predictive model M adapted for the test patient domain
Phase 1: Training the Predictive Model
for each batch $B \in \mathcal{D}_{train}$ **do**
    Update M using B: $M \leftarrow M - \eta \nabla \mathcal{L}_{train}(M, B)$
end for
Phase 2: Fine-tuning with Test Dataset
Split $\mathcal{D}_{test}$ into fine-tuning subset $\mathcal{D}_{ft}$ and evaluation subset $\mathcal{D}_{eval}$
for each batch $B_{ft} \in \mathcal{D}_{ft}$ **do**
    Fine-tune M using B_{ft}: $M \leftarrow M - \eta \nabla \mathcal{L}_{ft}(M, B_{ft})$
end for
Phase 3: Test Time Adaptation
for each sample $x \in \mathcal{D}_{eval}$ **do**
    Apply unsupervised learning to update M: $M \leftarrow \text{UnsupervisedUpdate}(M, x)$
end for
Evaluate the adapted model M on $\mathcal{D}_{eval}$
return Final predictions by M on $\mathcal{D}_{eval}$

4.3.3 Rank consistent ordinal regression neural network

The concept underlying the ordinal regression algorithm consists of converting the ranking challenge into multiple binary classification tasks. As elucidated in the study (Cao *et al.*, 2020), a rank y_i is initially expanded into a vector of $K-1$ binary labels $y_i^{(1)}, \ldots, y_i^{(K-1)}$, where $y_i^{(k)} = 1\{y_i > r_k\}$ signifies if y_i surpasses rank r_k. The boolean operation $1\{\cdot\}$ returns 1 if its enclosed condition is met, and 0 otherwise. This expansion into binary labels facilitates the training of $K-1$ binary classifiers within the neural network's output layer.

Let W represent the weight parameters of the neural network, with the exception of the bias terms in the final layer, and let b_k be the bias associated with the k^{th} output neuron. To ensure rank consistency across its output layer tasks, all neurons in the final output layer share identical weight parameters. The inputs to the final layer are denoted as $\{g(x_i, W) + b_k\}_{k=1}^{K-1}$. Define the logistic sigmoid function as $\sigma(z) = \frac{1}{1+e^{-z}}$. The empirical probability that the k^{th} binary classification task predicts $y_i^{(k)} = 1$ is given by:

$$\hat{P}(y_i^{(k)} = 1) = \sigma(g(x_i, W) + b_k). \tag{4.1}$$

During model training, the objective is to minimize the loss function:

$$L(W, b) = -\sum_{i=1}^{N}\sum_{k=1}^{K-1} \lambda^{(k)} \left[\log(\sigma(g(x_i, W) + b_k))y_i^{(k)} + \log(1 - \sigma(g(x_i, W) + b_k))(1 - y_i^{(k)})\right], \tag{4.2}$$

which represents the weighted cross entropy for the $K-1$ binary classifiers. Here, $\lambda^{(k)}$ indicates the loss weight for the k^{th} classifier, a critical parameter for task k. Given the challenges in optimizing $\lambda^{(k)}$, a non-uniform task weight scheme is preferred to address the data imbalance issue.

4.3.4 Minimise prediction entropy at test time

Note that the theorem for rank-monotonicity proposed by the work (Li and Lin, 2006). By minimizing the loss function defined in Eq. (2), the optimal solution $(\mathbf{W}^*, \mathbf{b}^*)$ satisfies $b_1^* \geq b_2^* \geq \ldots \geq b_{K-1}^*$. the work (Cao *et al.*, 2020) applies this theory to ordered regression neural networks by adding the restriction of weight sharing, which ensures that the probability curves produced by each binary classifier do not want to intersect, in this case with a sigmoid function, so that if each classifier uses a different weight w, then their respective sigmoid curves will have different slopes. This may result in the curves of different classifiers intersecting at certain values of x. Intersecting curves mean that the model's prediction of the category is no longer monotonic at certain values of the feature x. For example, as x increases, the model may predict a lower and then a higher category probability, which violates the basic properties of ordered categories.

In this way, in the simplified model $p_i = \sigma(wx + b_i)$, Assume that x is the input feature and $f(x)$ represents the neural network's function mapping. We have a shared weight w and an ordered set of biases $b_1 > b_2 > \ldots > b_{K-1}$. Each bias defines a decision boundary that partitions the input space into K intervals. For a given input x, we compute $f(x)$ and compare it against the biases to determine in which interval x falls, thereby determining the category of x:

$$\text{Category} = \begin{cases} 1, & \text{if } f(x) + b_1 \leq 0 \\ 2, & \text{if } f(x) + b_1 > 0 \text{ and } f(x) + b_2 \leq 0 \\ 3, & \text{if } f(x) + b_2 > 0 \text{ and } f(x) + b_3 \leq 0 \\ \vdots & \\ K, & \text{if } f(x) + b_{K-1} > 0 \end{cases}$$

Here, $f(x)$ is typically the dot product $w^T x$, though it could be a more complex non-linear function in a neural network context. Each threshold b_i corresponds to a specific decision boundary that defines the conditions under which x falls between the ith and $i + 1$th category.

In order to minimize the prediction entropy, we want $f(x)$ to be close to the "middle" of the interval of the class it predicts.

Let m_i be the median of the i^{th} interval, defined as:

$$m_i = \begin{cases} b_1, & \text{if } i = 1 \\ \frac{b_{i-1}+b_i}{2}, & \text{if } 1 < i < K \\ b_{K-1}, & \text{if } i = K \end{cases}$$

For a given input x, let c be the predicted category, then the loss is:

$$L(x) = (f(x) - m_c)^2$$

The objective is to minimize the expected loss over all inputs:

$$\mathcal{L} = \mathbb{E}\left[L(x)\right]$$

This loss encourages $f(x)$ to be close to the center of the interval associated with its predicted category.

Algorithm 2 prediction entropy minimization on Coral

Require: x (input feature), w (shared weight), $\{b_i\}_{i=1}^{K-1}$ (ordered biases)
Ensure: Category of x and minimized loss $\mathcal{L}$

```
f(x) ← w^T x                          ▷ Assume f(x) represents the neural network function
for i = 1 to K − 1 do
    if f(x) + b_i ≤ 0 then
        Category ← i
        break
    end if
end for
if f(x) + b_{K−1} > 0 then
    Category ← K
end if
Define m_i for each interval:
```

$$m_i = \begin{cases} b_1, & \text{if } i = 1 \\ \frac{b_{i-1}+b_i}{2}, & \text{if } 1 < i < K \\ b_{K-1}, & \text{if } i = K \end{cases}$$

```
c ← Predicted category for x
L(x) ← (f(x) − m_c)^2                 ▷ Compute loss for input x
ℒ ← 𝔼[L(x)]                           ▷ Minimize expected loss over all inputs
```

4.3.5 Subject-specific Batch Normalization

Let $D = \{D_1, D_2, \ldots, D_S\}$ denote the entire dataset, where D_s represents the dataset for subject s, and S is the total number of subjects. Each subject's dataset D_s can further be defined as a set of data points, i.e., $D_s = \{(x_{s1}, y_{s1}), (x_{s2}, y_{s2}), \ldots, (x_{sn_s}, y_{sn_s})\}$, where x_{si} denotes the i^{th} input feature for subject s, y_{si} denotes the corresponding label, and n_s is the total number of data points for subject s.

During each training iteration, we construct subject-specific batches B_s instead of randomly selecting data points from the entire dataset D. Such a batch can be represented as:

$$B_s = \{(x_{sj}, y_{sj}) | j = 1, \ldots, M\}$$

where M is the batch size and $M \leq n_s$, indicating that the number of data points in each batch does not exceed the total number of data points in the subject's dataset. In this manner, all data points in batch B_s originate exclusively from one subject's dataset D_s, allowing the neural network to learn distribution characteristics specific to each subject during training.

The key advantage of this approach is its direct targeting of distribution shifts between subjects by training the neural network to recognize and adapt to the unique data distribution of each subject, potentially enhancing the model's adaptability to individual differences and its generalization capability.

4.3.6 Loss weight scheme for data unbalance

The problem of data imbalance is common in medical applications. In order to solve this problem without increasing the arithmetic requirements and guaranteeing the data integrity, we use the loss weight method to give different weights to the losses on the output layer to regulate the influence of different classes on the model parameters. Here, we describe the two strategies used in the following.

Firstly, a prevalent approach to mitigating data imbalance in machine learning, especially when dealing with tasks or datasets comprising pairs of samples, involves the application of differential loss weighting, called Inverse Frequency Weighting (IFW). This technique assigns lower weights to samples or sample pairs that are abundant, while allocating higher weights to those associated with underrepresented tasks. Mathematically, let w_i denote the weight for the i-th task, then w_i is inversely proportional to the number of samples in that task, formulated as $w_i = \frac{1}{N_i^\alpha}$, where N_i represents the number of samples in the i-th task, and α is a parameter that controls the degree of weighting adjustment. This loss weighting

schema ensures that the learning process pays greater attention to tasks with fewer samples, thereby encouraging a more balanced performance across all tasks and addressing the issue of data imbalance.

Secondly, in addressing the imbalance inherent in ordinal classification tasks, we refer to the work (Han *et al*., 2020), introducing a Task Importance Weighting (TIW) strategy. For each task derived from ordinal ranks, let $S_k = \sum_{i=1}^{N} \mathbf{1}\{y_i^k = 1\}$ denote the count of instances exceeding a rank threshold r_k, with $S_1 \geq S_2 \geq \ldots \geq S_{K-1}$ indicating decreasing instance counts. The imbalance is quantified using $M_k = \max(S_k, N - S_k)$, and the task importance, λ^k, is defined by normalizing the square root of M_k against the largest square root value across all tasks:

$$\lambda^k = \frac{\sqrt{M_k}}{\max\limits_{1 \leq i \leq K-1} \sqrt{M_i}}$$

This methodology ensures equitable weight allocation across binary classification tasks, mitigating the effects of label imbalance.

4.4 Experiment Design and Implementation

In this section, we conduct experiments to demonstrate the effectiveness of PCOR-PA algorithm in pain assessment and we apply our method to classify target labels from short segments of EDA data. Our experimental results are obtained by averaging over 5 different runs. Since the standard deviation is small in all experiments, we ignore the standard deviation.

4.4.1 Dataset

We first developed and validated the proposed method in our private dataset Apon and then further tested it externally with the publicly available pain dataset BioVid. The basic information of the BioVid and Apon datasets is very similar. Recruitment was done on healthy volunteers, and the number of people was respectively 87 and 60, respectively. They are both pain triggered by thermal stimulation experiments, and the definition of pain is Pain

stimulation at the pain threshold, pain tolerance, and two levels between them. The main differences in the experiments were the length of the stimulus and the number of stimuli per subject; Apon was a 30-second stimulus duration and a rest period of about 1 minute and 30 seconds between stimuli, with the stimulus occurring a total of 30 times. Biovid had a 5-second stimulus duration, with a rest period of 4 to 8 seconds, with the stimulus occurring a total of 100 times.

In addition, we brought the algorithm into real medical scenarios where we collected data from two different types of pain patients. The first type is post-surgical patients. For this group of patients, we performed data collection and pain assessment reports (VAS scores) in three different scenarios: pre-anesthesia, in the waking room, and in the hospital room, and in each scenario, two self-reported pain assessments were performed, which were categorized into dynamic pain scores and static pain assessments. Here, only the patient's pain scores at rest are considered due to uncertainty in dynamic pain scores. The length of data collected for each patient was determined by the medical setting of the scenario at the time, with an average length of 8 to 15 minutes. The second type of patient with pain we chose was a woman in labor who required an anesthesiologist to assess her pain and administer analgesia prior to entering into labor and delivery. At this stage, we were able to collect data including facial expression, EDA, BVP, blood pressure, blood oxygen, heart rate, respiratory rate, the intensity of contractions, the interval between contractions, and fetal heart rate. The data collection experiments were all performed without interfering with the patient's normal rest and treatment.

After collecting and organizing the data and pain labels, we asked the physicians to perform label revisions, which followed predetermined assessment rules, and the scores were continuous, i.e., in one piece of data, the physician would give different pain scores in different intervals.

4.4.2 Metric

The Leave-One-Subject-Out (LOSO) cross-validation method is designed for evaluating models on multi-subject datasets. It is particularly relevant in fields such as biomedical and psychological research, where data are collected from multiple individuals. In the LOSO approach, the dataset D consists of data from N different subjects, denoted as $D = \{D_1, D_2, \ldots, D_N\}$. For each iteration i of the cross-validation, the data from one subject D_i is used as the test set, and the data from all remaining subjects, $D \setminus D_i$, are used for training the model. This process is repeated N times, with each subject's data being used as the test set exactly once. The performance of the model is then averaged across all N iterations, providing an estimate of its generalizability to new, unseen subjects. This method ensures a stringent evaluation of the model's ability to generalize across the variability inherent in different subjects.

Combining Test-Time Adaptation (TTA) with a "Leave One Subject Out" approach involves using TTA while systematically excluding one subject (or a group of data points) from the training data and using it as a test set. In this setup, TTA fine-tunes the model for each test subject individually, adapting the model to the specific characteristics of the left-out subject's data. This approach can be particularly useful in fields like medical imaging, where personalized adaptation to individual patient data is crucial. By applying TTA in a "Leave One Subject Out" scenario, the model can potentially achieve better performance and generalization for each individual test subject.

Some previous pain studies have used metrics for regression questions because Pain labels have a natural ordering. The researchers hope that the predictions of the model can be closer to the true values rather than just considering whether the predictions match the true results. In the evaluation of regression models, two commonly used metrics are the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE), which measure the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

- **Mean Squared Error (MSE):** The MSE calculates the average of the squares of the errors or deviations; that is, the difference between the estimator and what is estimated. MSE provides a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better. The MSE is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

  where n is the number of observations, Y_i is the actual value of the observation, and $\hat{Y}_i$ is the predicted value.

- **Root Mean Squared Error (RMSE):** The RMSE is the square root of the mean square error. It measures the standard deviation of the residuals or prediction errors. Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. The RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

  RMSE is particularly useful when you want to gauge the error magnitude in the same units as the response variable.

Both MSE and RMSE are crucial for understanding the accuracy and performance of a regression model, with RMSE being more interpretable in terms of the original units of the output variable.

The Quadratic Weighted Kappa (QWK) is a statistical measure used to assess the degree of agreement between two raters on an ordinal scale, with a particular focus on the severity of disagreements. It modifies the traditional Cohen's kappa by incorporating a weight matrix W, where the weights are defined as:

$$W_{ij} = (i - j)^2$$

These weights exponentially penalize discrepancies based on the squared distance between categories. The QWK is calculated using the formula:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

where p_0 is the observed weighted agreement and p_e is the expected weighted agreement under random chance. Quadratic Weighted Kappa thus provides a more sensitive measure of agreement that accounts for the ordinal nature of the data and the varying seriousness of different types of disagreements.

For continuous prediction of accuracy, we introduce Maximum Consecutive High Error (MCHE) defined as follows:

$$MCHE = \max_{i,j:i\leq j} (j - i + 1) \text{ subject to } x_k = 1 \text{ for all } i \leq k \leq j$$

where, x_t is defined as:

$$x_k = \begin{cases} 1 & \text{if } |y_k - \hat{y}_k| > \theta \\ 0 & \text{otherwise} \end{cases}$$

In this formula, the sequence $X = \{x_1, x_2, \ldots, x_T\}$ represents whether the prediction error at each time point k exceeds a predefined threshold θ. If the absolute error $|y_k - \hat{y}_k|$ is greater than θ, then $x_k = 1$ indicating a significant error, otherwise $x_k = 0$. $MCHE$ is the length of the longest consecutive sequence of $x_k = 1$. The $\max$ function finds the maximum length $(j - i + 1)$ among all possible consecutive intervals $[i, j]$ where $x_k = 1$ throughout. This metric effectively evaluates the stability and reliability of the prediction model under consecutive extreme conditions.

Here, we also introduce a metric, called WMAE, that adjusts the weights based on the continuity and magnitude of the prediction error. Such an indicator would better reflect the continuous large errors you wish to avoid. Assume you have a time series with true values y_t and predicted values $\hat{y}_t$ at time t, for $t = 1, 2, \ldots, T$. Define the absolute error at time t as $e_t = |y_t - \hat{y}_t|$.

Introduce a threshold θ to identify "large errors." If the error e_t exceeds this threshold, it's considered significant. We'll define a sequence of weights w_t that increase with consecutive occurrences of large errors:

1. If $e_t > \theta$, then $n_t = n_{t-1} + 1$. 2. If $e_t \leq \theta$, then $n_t = 0$.

We can then define the weight w_t as $w_t = 1 + \alpha n_t$, where α is a positive parameter that modulates the impact of consecutive large errors on the weight.

The overall weighted error metric E can be computed as:

$$E = \frac{\sum_{t=1}^{T} w_t e_t}{\sum_{t=1}^{T} w_t}$$

This formula weights each time point's error e_t by w_t, enhancing the effect of continuous large errors on the overall assessment.

4.4.3 Domain generalization benchmarks

Compared Baselines. To validate the effectiveness of our algorithms, we implemented the more common algorithms in domain adaptation and domain generalization, such as TENT (Wang *et al*., 2020a), MEMO (Zhang *et al*., 2022), Domain-Adversarial Neural Networks (DANN) (Ganin and Lempitsky, 2015), Mixup (Zhang *et al*., 2017), Correlation Alignment (CORAL) (Sun and Saenko, 2016), Maximum Mean Discrepancy (MMD) (Borgwardt *et al*., 2006), Disentanglement (Bui *et al*., 2021).

- **TENT:** It is designed for test-time adaptation, allowing a model to adjust to new, unseen data during testing by minimizing the prediction entropy. It operates by updating only the normalization statistics and affine parameters of a model's feature representations, without altering the core model parameters. This approach is efficient and stable, as it focuses on low-dimensional, linear modulations (scales and shifts) that are less prone to causing divergence from the model's trained state. TENT's effectiveness is demonstrated through its application to various domain adaptation scenarios, including image classification tasks with different datasets.
- **MEMO:** algorithm represents an innovative approach to enhancing model robustness at test time through adaptation and augmentation. This technique is particularly designed to address challenges related to distribution shifts between training and deployment environments, a common scenario in real-world applications. Memo

operates by dynamically adjusting the model's parameters or applying strategic data augmentations during the test phase, without the need for retraining or access to the original training data. This process aims to align the model more closely with the current data distribution encountered at test time, thereby improving prediction accuracy and reliability.

- **DANN:** It addresses domain adaptation by introducing a domain classifier that learns to distinguish between source and target domain features while the feature extractor learns to generate domain-invariant features. This adversarial process is facilitated by a gradient reversal layer, which updates the feature extractor to fool the domain classifier, thereby encouraging domain invariance. DANN's approach to generating features that are indistinguishable across domains makes it a powerful tool for domain adaptation.
- **Mixup:** It is a data augmentation technique that operates by creating virtual training examples through the linear interpolation of pairs of examples and their labels. This method encourages the model to behave linearly in-between training examples, which can improve the model's generalization performance and robustness to adversarial examples. By training on these mixed examples, Mixup helps to reduce overfitting and improve the model's performance on unseen data
- **CORAL:** It aims to minimize domain shift by aligning the second-order statistics (covariances) of source and target distributions, without requiring explicit domain labels. This is achieved by adjusting the source features to have the same covariance as target features, thereby making the features more similar across domains. CORAL is particularly effective in situations where the domains are related but exhibit variation in their data distributions.
- **MMD:** It is a measure of the distance between the distributions of source and target domains. It works by computing the difference in expectations across domains for a given function class. MMD can be used as a loss function to train models that minimize the distributional discrepancy, ensuring that the learned representations are domain-invariant. This makes MMD a valuable tool for both domain adaptation and domain generalization tasks.

- **Disentanglement:** It aims to learn representations that separate the underlying factors of variation in the data. By disentangling these factors, the model can achieve better generalization by focusing on the relevant features that are invariant across domains. This approach is beneficial for tasks where the model needs to generalize across different, unseen domains by capturing the essence of the data in a way that is not specific to the domain it was trained on.

4.5 Result and Analysis

4.5.1 Comparison with domain adaptation algorithms

We compare our model with some traditional machine learning methods for EDA-based pain assessment (support vector machine, Random forest). Also, we compare with other DG algorithms designed for more general fields (TENT (Wang *et al*., 2021), DANN (Ganin and Lempitsky, 2015), Mixup (Zhang *et al*., 2017), CORAL (Sun and Saenko, 2016), MMD (Tzeng *et al*., 2014), Disentanglement (Bui *et al*., 2021), GroupDRO (Sagawa *et al*., 2019), VERx (Krueger *et al*., 2021)). The baseline is Empirical Risk Minimization (ERM) which minimizes the average loss across all the training examples from all the subjects. In our experiments, we combined the ERM and TENT methods with different settings for the updated model parameters in the TENT method (BN for Batch normalization layer, BN-L for Batch normalization layer and linear layer, and C for the full set of parameters). We conducted 10 rounds of experiment and reported means and standard deviations.

As can be seen from Table 2, the performance of PCOR-PA is superior to other methods. PCOR-PA's improved performance is attributed to the fact that it takes into account the pain specificity between different patients throughout the training process. In addition, PCOR-PA also breaks the "fair" to "good" kappa qualifier (Landis and Koch, 1977). This performance was seen on both the biovid and apon datasets. This shows that PCOR-PA can effectively improve the consistency of the pain level prediction model and match the actual situation more

TABLE 4.1: Results of comparative study.(Bolding denotes the best.)

	BioVid			Apon		
Method	**MAE**	**RMSE**	**QWK**	**MAE**	**RMSE**	**QWK**
Random Forest	0.930 ± 0.005	$\mathbf{1.112 \pm 0.005}$	0.436 ± 0.004	1.198 ± 0.006	1.482 ± 0.007	0.342 ± 0.008
SVM	0.915 ± 0.009	1.118 ± 0.008	0.445 ± 0.008	1.114 ± 0.008	1.320 ± 0.007	0.325 ± 0.009
ERM	1.019 ± 0.008	1.346 ± 0.009	0.389 ± 0.005	0.857 ± 0.005	1.111 ± 0.007	0.310 ± 0.005
ERM with TENT-BN	0.959 ± 0.004	1.295 ± 0.002	0.442 ± 0.005	0.838 ± 0.008	1.112 ± 0.010	0.418 ± 0.002
ERM with TENT-BN-L	0.959 ± 0.004	1.295 ± 0.002	0.442 ± 0.005	0.858 ± 0.005	1.112 ± 0.007	0.408 ± 0.007
ERM with TENT-C	0.958 ± 0.007	1.301 ± 0.005	0.425 ± 0.006	0.853 ± 0.006	1.114 ± 0.008	0.328 ± 0.002
ERM with Coral layer	1.053 ± 0.002	1.357 ± 0.005	0.362 ± 0.005	0.782 ± 0.004	1.018 ± 0.006	0.465 ± 0.004
DANN	0.901 ± 0.010	1.182 ± 0.008	0.476 ± 0.002	0.780 ± 0.005	0.998 ± 0.008	0.462 ± 0.008
Mixup	1.102 ± 0.013	1.115 ± 0.015	0.259 ± 0.012	1.005 ± 0.010	1.118 ± 0.011	0.298 ± 0.019
CORAL	1.230 ± 0.018	1.452 ± 0.012	0.198 ± 0.025	1.120 ± 0.009	1.479 ± 0.012	0.276 ± 0.010
MMD	1.284 ± 0.015	1.312 ± 0.016	0.223 ± 0.009	1.280 ± 0.012	1.396 ± 0.014	0.221 ± 0.001
Disentanglement	1.440 ± 0.005	1.798 ± 0.008	0.102 ± 0.008	1.774 ± 0.009	2.018 ± 0.012	0.070 ± 0.003
GroupDRO	1.119 ± 0.001	1.515 ± 0.001	0.366 ± 0.010	1.253 ± 0.007	1.121 ± 0.003	0.302 ± 0.006
VERx	0.986 ± 0.001	1.236 ± 0.001	0.392 ± 0.015	0.911 ± 0.004	1.213 ± 0.005	0.328 ± 0.007
Ours (PCOR-PA)	$\mathbf{0.889 \pm 0.018}$	1.202 ± 0.020	$\mathbf{0.492 \pm 0.017}$	$\mathbf{0.775 \pm 0.006}$	$\mathbf{0.996 \pm 0.004}$	$\mathbf{0.513 \pm 0.002}$

accurately. This is particularly valuable for clinically important decision-making systems that require accuracy and reliability.

Table 3 shows the performance of the Apon-heat dataset when we consider the predictions of the data as continuous rather than independent. Only the Apon-heat dataset was used because the length of time for a single sample is about 30 or so, whereas a single sample in the Biovid dataset is only 5.5 seconds of data for each independent thermal stimulus. For this analysis, we introduce two new metrics for evaluating the continuity of the continuous prediction problem: WMAE emphasizes the performance of the model at different time points, while MCHE monitors the overall stability of the model. Under the setting of continuous prediction, we use a sliding window to slice the EDA sequence, which may result in inconsistent label-data pairing. In this scenario, we observed that PCOR-PA still outperforms the baseline method and maintains a relatively competitive performance level compared to other DG methods.

We comparatively evaluated PCOR-PA and state-of-the-art methods on the Biovid Heat Pain Database. To be consistent with previous work, we used MAE and RMSE to measure regression problems, Also, we chose to compare BLN with P4 to compare binary classification performance, as higher pain intensities are more likely to be recognized and may be considered the upper limit of performance when recognizing a single pain intensity. To do this, we modify

TABLE 4.2: Results of comparative study. (Bolding denotes the best, underlining denotes the second best)

Method	Apon			
	MAE	**RMSE**	**MCHE**	**WMAE**
Random Forest	1.152 ± 0.007	1.250 ± 0.009	1.358 ± 0.010	1.126 ± 0.008
SVM	1.011 ± 0.005	1.118 ± 0.006	1.102 ± 0.009	0.995 ± 0.007
ERM	0.857 ± 0.008	1.111 ± 0.012	1.210 ± 0.015	0.938 ± 0.005
ERM with TENT-BN	0.820 ± 0.004	1.024 ± 0.012	1.030 ± 0.003	0.953 ± 0.006
ERM with TENT-BN-l	0.825 ± 0.005	1.024 ± 0.009	1.030 ± 0.013	0.990 ± 0.012
ERM with TENT-C	0.850 ± 0.005	1.118 ± 0.007	0.985 ± 0.003	0.910 ± 0.015
ERM with Coral layer	0.802 ± 0.005	1.051 ± 0.003	0.955 ± 0.009	0.989 ± 0.001
DANN	0.790 ± 0.007	$\mathbf{1.001 \pm 0.010}$	$\mathbf{0.921 \pm 0.003}$	0.975 ± 0.009
Mixup	0.920 ± 0.012	1.212 ± 0.015	1.312 ± 0.003	1.152 ± 0.008
CORAL	0.924 ± 0.010	1.213 ± 0.012	1.321 ± 0.008	1.160 ± 0.006
MMD	0.924 ± 0.010	1.213 ± 0.012	1.321 ± 0.008	1.160 ± 0.006
Disentanglement	1.021 ± 0.020	1.444 ± 0.018	1.890 ± 0.012	1.212 ± 0.006
GroupDRO	0.910 ± 0.013	1.212 ± 0.008	1.218 ± 0.005	1.116 ± 0.012
VERx	0.853 ± 0.001	1.019 ± 0.003	0.958 ± 0.012	0.998 ± 0.005
Ours(PCOR-PA)	$\mathbf{0.784 \pm 0.003}$	$\underline{1.015 \pm 0.008}$	$\underline{0.928 \pm 0.008}$	$\mathbf{0.925 \pm 0.002}$

the output of the last coral layer into a single binary classification task. From the table 4.3, we can observe that in the case where I use only EDA for individual physiological signal data, we outperform the state-of-art methods in retaining all "pain" samples in the data. For the *BLN vs P4* task, our method also shows some improvement, which indicates the effectiveness of our method.

TABLE 4.3: Comprison with the state of art methods. The regression and classification results are provided in the form of [MAE, RMSE] and Accuracy $\pm$ standard deviation respectively

BioVid				
Model	Signal	Regression [0-4]	Regression [1-4]	BLN vs P4 Accuracy (%)
LSTM-NN (Lopez-Martinez and Picard, 2018)	Physiological(EDA + ECG)	1.05, 1.29	NA	NA
NN (Lopez-Martinez *et al*., 2017)	Physiological (EDA + ECG) +Video	0.77, 1.15	NA	82.75 ± 1.86
Random Forest (Kächele *et al*., 2015)	Physiological(EDA + ECG + EMG)+Video	NA	0.84, 0.98	84.20 ± 13.70
DDCAE (Thiam *et al*., 2021)	Physiological(EDA + ECG + EMG)	0.97, 1.16	NA	83.99
NN (Jiang *et al*., 2024)	Physiological (EDA + ECG)	0.93, 1.12	0.84, 1.00	84.58 ± 13.28
Our (PCOR-PA)	Physiological (EDA)	0.88, 1.20	0.76, 1.03	85.12 ± 2.13

4.5.2 Ablation study

We conduct an ablation study to evaluate the effect of finetuning and test time adaptation on the performance of PCOR-PA. We have four variants: (a) PCOR-PA w/o pre-test finetuning (FT), (b) PCOR-PA w/o test time adaptation (TTA), (c) PCOR-PA w/o pre-test finetuning and

test time adaptation, (d) PCOR-PA w/o any personalizing operation (PL). The personalization operation is used in conjunction with both FT and TTA, which serves to do a personal standarlization of the patient data, and ensures that the samples in a batch are all from the same subject by using the batch in the training and testing of the model (see details in Appendix c).

Table 4.4 demonstrates the results. we observe that PCOR-PA without FT or PCOR-PA without TTA achieved the lowest MAE and kappa index in the BioVid and Apon datasets respectively, suggesting that adapting the model using the target subject's data is essential for making accurate personalized pain assessments. There is a significant drop in performance for PCOR-PA w/o PL, highlighting the importance of taking into account differences in data distribution between patients.

TABLE 4.4: Results of Ablation study. (Bolding denotes the best, underlining denotes the second best)

	BioVid			**Apon**		
Method	**MAE**	**RMSE**	**QWK**	**MAE**	**RMSE**	**QWK**
PCOR-PA w/o FT	1.060 ± 0.004	1.359 ± 0.004	0.365 ± 0.005	0.910 ± 0.008	1.131 ± 0.010	0.333 ± 0.002
PCOR-PA w/o TTA	1.047 ± 0.008	1.344 ± 0.009	0.365 ± 0.012	0.928 ± 0.005	1.202 ± 0.003	0.302 ± 0.001
PCOR-PA w/o FT and TTA	0.891 ± 0.007	1.212 ± 0.005	$\mathbf{0.506 \pm 0.006}$	0.882 ± 0.003	1.112 ± 0.002	0.412 ± 0.005
PCOR-PA w/o PL	1.042 ± 0.001	1.314 ± 0.002	0.339 ± 0.002	0.876 ± 0.010	1.111 ± 0.013	0.425 ± 0.007
PCOR-PA	$\mathbf{0.888 \pm 0.002}$	$\mathbf{1.198 \pm 0.005}$	$\underline{0.497 \pm 0.002}$	$\mathbf{0.775 \pm 0.006}$	$\mathbf{0.996 \pm 0.004}$	$\mathbf{0.513 \pm 0.002}$

4.5.3 Finetuning on hyperparameters

1D-CNN for feature extraction. The model first processes the input through a sequence of features containing two convolutional layers. The first convolutional layer uses 1 input channel and 3 output channels, with a convolutional kernel size of 3, a step size of 1, and a padding of 1. It is followed by a bulk normalization layer and a LeakyReLU activation function, and then halves the data dimensionality through an average pooling layer. The second convolutional layer receives 3 input channels and outputs 6 channels, with the same configuration as the first convolutional layer, and is again passed through a bulk normalization, LeakyReLU activation, and average pooling.

Coral Layer. Coral layer is just a linear layer and accepts the input features x, represented as a tensor with shape=(num_examples, num_features). It processes these features using the *coral_weights* to generate a linear output for each example, which is then combined with the *coral_bias* to produce logits. The shape of the logits is (num_examples, num_classes-1), representing the model's estimation of the ordinal thresholds for each class in an ordered classification task.

Hyperparameters.

- **LOSS weighting method**: The output of the final layer of the model can be interpreted as the prediction result of a series of dichotomous classification problems.Considering the imbalance of the data and the uncertainty of the data with different pain levels, we experimented with three different approaches to the modulation of the effect of the loss function on the updating of the model parameters.
- **Learning Rates**:
  - **0.0001 for SGD**: Used in the pre-test finetuning to optimize specific parameters, likely those of the CORAL layer. This learning rate is crucial for making small, precise updates to the model.
  - **0.001 for Adam**: Used during the test time adaptation phase to optimize a subset of parameters for final adjustments based on the test data.
- **Weight Decay**: Set at $1e-4$ for SGD during fine-tuning, helping to prevent overfitting by penalizing larger weights.
- **Upper and lower bounds**: In our design for calculating the loss of predictive entropy for ordered regression networks, we want to minimize the median of the predicted values and their prediction intervals, the intermediate classes can calculate the median based on their neighboring classes, but the two largest and smallest classes require an upper and lower bound in which the median can be calculated.

ablation study on parameters updating on pre-test finetuning stage.

ablation study on parameters updating on TTA stage.

TABLE 4.5: Average performance (mean and standard deviation) on finetuning different layers

Method	BioVid			Apon		
	MAE	RMSE	QWK	MAE	RMSE	QWK
Finetuning on BN layer	0.895 ± 0.001	1.209 ± 0.002	0.507 ± 0.004	0.804 ± 0.008	1.142 ± 0.006	0.425 ± 0.003
Finetuning on BN + Coral layer	0.901 ± 0.003	1.201 ± 0.004	0.501 ± 0.004	0.776 ± 0.005	1.032 ± 0.008	0.479 ± 0.004
Finetuning on Coral layer	1.021 ± 0.010	1.288 ± 0.016	0.380 ± 0.011	0.825 ± 0.010	1.210 ± 0.008	0.422 ± 0.014
Finetuning on all layer	0.912 ± 0.017	1.250 ± 0.019	0.342 ± 0.016	0.858 ± 0.015	1.345 ± 0.013	0.410 ± 0.014

TABLE 4.6: Average performance (mean and standard deviation) on finetuning different layers

Method	BioVid			Apon		
	MAE	RMSE	QWK	MAE	RMSE	QWK
Finetuning on BN layer	0.895 ± 0.011	1.034 ± 0.012	0.483 ± 0.007	0.829 ± 0.008	1.207 ± 0.010	0.442 ± 0.012
Finetuning on BN + Coral layer	0.887 ± 0.006	1.028 ± 0.011	0.502 ± 0.009	0.795 ± 0.005	1.130 ± 0.007	0.452 ± 0.003
Finetuning on Coral layer	1.040 ± 0.015	1.350 ± 0.021	0.394 ± 0.007	0.814 ± 0.009	1.155 ± 0.005	0.432 ± 0.011
Finetuning on all layer	1.152 ± 0.007	1.312 ± 0.005	0.364 ± 0.006	0.952 ± 0.015	1.352 ± 0.018	0.328 ± 0.011

ablation study on TTA range.

TABLE 4.7: Average performance (mean and standard deviation) on finetuning upper and lower bound for entropy loss (TTA stage)

Method	BioVid			Apon		
	MAE	RMSE	KAPPA	MAE	RMSE	KAPPA
-10, 10	0.895 ± 0.011	1.109 ± 0.012	0.488 ± 0.006	0.782 ± 0.008	1.103 ± 0.005	0.501 ± 0.007
-20, 20	0.887 ± 0.006	1.103 ± 0.016	0.498 ± 0.008	0.779 ± 0.003	1.118 ± 0.003	0.508 ± 0.007
-50, 50	0.902 ± 0.006	1.203 ± 0.010	0.472 ± 0.009	0.793 ± 0.004	1.137 ± 0.004	0.462 ± 0.009
-100, 100	0.896 ± 0.009	1.135 ± 0.009	0.490 ± 0.005	0.785 ± 0.003	0.995 ± 0.004	0.499 ± 0.004
-1000, 1000	0.913 ± 0.009	1.211 ± 0.007	0.462 ± 0.008	0.810 ± 0.005	1.013 ± 0.004	0.452 ± 0.009

ablation study on loss weighting methods.

TABLE 4.8: Comparison of validation loss (MAE and RMSE) of all three task important weighting schemes on Biovid database

Method	BioVid		
	Uniform	Random	Hard
ERM	0.960, 1.360	0.968, 1.359	0.941, 1.333
ERM with TENT-BN	0.956, 1.348	1.011, 1.389	0.938, 1.274
ERM with TENT-C	1.014, 1.419	1.050, 1.450	0.958, 1.301
ERM(Coral)	0.910, 1.162	0.884, 1.186	0.896, 1.171
DANN	0.890, 1.156	0.888, 1.180	0.887, 1.182
Mixup	1.102, 1.115	1.207, 1.399	1.155, 1.527
CORAL	1.230, 1.452	1.320, 1.521	1.212, 1.401
MMD	1.284, 1.312	1.012, 1.432	1.144, 1.448
Disentanglement	1.440, 1.798	1.485, 1.832	1.312, 1.603
GroupDRO	0.982, 1.302	1.018, 1.452	0.975, 1.210
VERx	0.912, 1.201	0.962, 1.333	0.901, 1.105
Ours (PCOR-PA-bn)	0.890, 1.189	0.899, 1.177	0.889, 1.159
Ours (PCOR-PA-c)	0.897, 1.214	0.915, 1.240	0.896, 1.189

TABLE 4.9: Comparison of validation loss (MAE and RMSE) of all three task important weighting schemes on Apon database

Method	Apon		
	Uniform	Random	Hard
ERM	0.869, 1.116	0.861, 1.124	0.857, 1.111
ERM with TENT-BN	0.869, 1.117	0.862, 1.127	0.858, 1.112
ERM with TENT-C	0.859, 1.122	0.860, 1.118	0.853, 1.114
ERM(Coral)	0.777, 1.017	0.786, 1.019	0.782, 1.018
DANN	0.785, 1.031	0.791, 1.120	0.780, 0.998
Mixup	1.012, 1.230	1.085, 1.334	1.005, 1.118
CORAL	1.118, 1.358	1.115, 1.288	1.120, 1.479
MMD	1.110, 1.247	1.155, 1.286	1.280, 1.396
Disentanglement	1.874, 2.210	1.596, 2.230	1.774, 2.018
GroupDRO	1.010, 1.201	1.201, 1.352	0.988, 1.102
VERx	0.965, 1.155	1.215, 1.448	0.965, 1.201
Ours (PCOR-PA-bn)	0.779, 1.001	0.782, 1.121	0.775, 0.996
Ours (PCOR-PA-c)	0.798, 1.155	0.785, 1.130	0.780, 1.106

4.6 External Validation on Real Pain Patients

In this section, we used two datasets collected from real patients, which are Apon-postoperative and Apon-labor. In order to ensure the quality of the datasets, we performed data cleansing and filtering on these two data respectively. The data collection methods for these two datasets have been described in Section 4.4.1. Here, I will describe both datasets in detail, as well as the process of preparing the datasets.

4.6.1 Data exploration

In a real medical environment, using Evu sensor as the acquisition device, the collected eda device has anomalous data with a value <1 within a few minutes. As a reference, the mean value of the eda signal in apon's internal experimental dataset is 5.59, and the mean value of the eda signal in biovid's public dataset is 3.44. In addition to this, the bvp signal also suffers from missing data and uneven data profiles. bvp data is representative of heart rate, which should fluctuate continuously and evenly. we applied four different method to solve this problem: 1) contact the manufacturer of the device collection to try to solve the problem. 2) in the process of data collection, pay more attention to the wearing of the device and the position of skin contact, to find the most appropriate contact method, as well as real-time monitoring of dynamic data changes, if there are unusual changes in the value of the collection process, you can record the scene that occurred at that time (such as changes in the wearing position of the device, etc.) 3) the need for pre-experimentation to find the most appropriate way to wear the device in the current medical environment. This takes into account that in a hospital, the patient may be doing something that causes the sensor to shift. 4) the use of multiple devices with simultaneous data acquisition.

4.6.2 Data cleaning

Due to the inability to control environmental variables under experimental conditions in real medical scenarios, we conducted a separate data quality analysis for each case (subject). For privacy reasons, each case is identified by an ID instead of a name. The following are some examples:

- ID 415138: The maximum values of the skin electrical signals in the pre-anesthesia and awakening rooms were 0.8 and 0.2, respectively. In contrast, the skin electrical signals in the wards showed complete performance. For the ECG data, significant missing data were observed in all three data sessions in the awakening room only.

- ID 415734: All skin electrical signals showed numerically small values. The complete waveform of the SCR event was visible only in the pre-anesthesia data. Only a few data points were missing in the BVP data.
- ID 416139: The tonic component in the pre-anesthesia phase was distinct and complete, showing an upward trend with a complete SCR response. However, in the resuscitation room and ward, the BVP data were invalid, possibly due to signal interference or incomplete wear of the equipment.
- ID 416130: The output format of Data No.1 differed from the rest of the data.
- ID 415315: Pre-anesthesia data were normal, with a monotonically rising trend in skin electrical signals during the awakening and ward phases. However, the VAS score indicated that the patient did not respond to pain.
- ID 413343: A sudden rise in the pre-anesthesia phase occurred simultaneously in one of the EDA and BVP signals. No significant fluctuations were observed in the galvanic skin data in the awakening room or ward.

After evaluating each patient's pain, we summarized some obvious data quality issues as follows:

- The values of skin electrical data were too small in some cases. Possible reasons include: 1) They represent normal data; 2) The data was collected in a position where valid skin electrical readings could not be obtained without affecting cardiac electrical signals; 3) Systemic problems with the device, where there was no noticeable change in the patient's data for up to a couple of minutes.
- Data format inconsistencies.
- Obvious sections of missing data where there is a sudden jump in ECG values. It is possible to determine the timing of the missing data so that algorithms can more accurately fill in the gaps.

To address these issues, we employed the following data cleaning and filtering process to ensure the dataset's quality: 1) Checking data integrity, identifying and dealing with missing values. For physiological signals, missing values may result from equipment malfunction or

sensor detachment. Short-term missing values are filled using interpolation methods (e.g., linear or nearest neighbor interpolation), and data are marked as unavailable for long periods. 2) Identifying outliers using statistical methods (e.g., box plot analysis) or machine learning models (e.g., isolated forests), which may result from sensor errors, data entry mistakes, or atypical physiological states of the patient. Outlier data points are corrected or removed as appropriate. 3) Improving signal quality for measures such as skin resistance, blood volume pulse (BVP), and skin temperature using signal processing techniques, including filtering and denoising. For instance, a low-pass filter is applied to eliminate high-frequency noise, and a moving average or median filter is used to smooth the data. 4) Integrating data from disparate sources into a consistent format to ensure completeness of each patient's data record, with accurate timestamps and other relevant information.

4.6.3 Data preparation

The length of real patients' data is different, and we need to do data slicing with sliding window before feeding the data into the model. Here we introduce the method of slicing window for data segmentation.

Consider a physiological signal represented as a time series, $\{x(t)\}$, where $t = 1, 2, \ldots, N$ indexes time and N is the total count of data points. To analyze this signal, we apply a sliding window technique, which partitions the data into overlapping segments.

The procedure involves two parameters: the window length, W, indicating the number of data points in each segment, and the step size, S, specifying how many points to move forward between segments. Generally, S is less than or equal to W, allowing for segments to overlap when $S < W$.

For any segment i, the data it contains, X_i, is defined as:

$$X_i = \{x(t)\}_{t=iS+1}^{min(iS+W,N)}$$

with i ranging from 0 up to $\left\lfloor \frac{N-W}{S} \right\rfloor$, ensuring coverage of the entire dataset while accommodating for the final segment that may contain fewer than W points.

4.6.4 Related Results

Finally, we present the results of externally validating real pain patients to highlight PCOR-PA's ability to provide valuable pain evaluation, which ultimately contributes to better personal pain management. In the following, we describe two real pain patient datasets: Apon-postoperative and Apon-pregnancy.

Apon-postoperative. Postoperative pain management has been a challenge for a number of reasons including complex pathomechanisms, inappropriate use of opioids and resource constraints in pain management . We conducted data collection at Union Hospital for patients undergoing ICU and thoracic surgery in the following three scenarios: pre-operative, anesthesia awakening room, and post-operative ward. While data collection was being conducted, physicians recorded pain levels through subjective VAS scores.

Apon-pregnancy. We apply our algorithms to a medical scenario of maternal pain in which an anesthesiologist needs to assess a pregnant woman's pain and analgesia before she enters the delivery room to give birth. Without interfering with the normal analgesic process, we perform data collection, including facial expression, EDA, blood volumn pulse, blood pressure, blood oxygen, heart rate, respiratory rate, contraction intensity, contraction interval, and fetal heart rate. When organizing the data, we asked obstetricians and gynecologists to perform continuous pain intensity labeling based on the information we provided in the data, where the pain labels were all derived from doctors' objective judgments of maternal paroxysms.

Table 4.10 shows the performance of result on patient on real world database. From the results of the Apon-postoperative dataset, we found a decrease in accuracy. The main reason for this may be that the postoperative pain dataset was collected from three different scenarios, and due to the many uncontrolled variables in real healthcare environments, pooling these

data from different scenarios can lead to noise in the data. However, our algorithm still outperforms the baseline model. On the contrary, in the Apon-pregnancy dataset, we found good results, probably because the maternal pain is acute and the pain response is obvious, and another reason is that for the maternal pain data, the pain labels are evaluated by experts.

TABLE 4.10: Results of comparative study

	Apon-postoperative			Apon-pregnancy		
Method	**MAE**	**RMSE**	**QWK**	**MAE**	**RMSE**	**QWK**
Random Forest	1.182 ± 0.002	1.412 ± 0.009	0.269 ± 0.009	0.653 ± 0.004	0.825 ± 0.010	0.528 ± 0.009
SVM	1.011 ± 0.003	1.213 ± 0.002	0.278 ± 0.005	0.722 ± 0.004	0.989 ± 0.005	0.519 ± 0.005
ERM	1.117 ± 0.011	1.385 ± 0.013	0.391 ± 0.004	0.810 ± 0.008	1.016 ± 0.006	0.502 ± 0.007
ERM with TENT-BN	1.052 ± 0.005	1.222 ± 0.006	0.358 ± 0.002	0.792 ± 0.005	0.952 ± 0.006	0.473 ± 0.005
ERM with TENT-BN-L	1.116 ± 0.004	1.314 ± 0.003	0.329 ± 0.007	0.833 ± 0.003	1.028 ± 0.004	0.402 ± 0.005
ERM with TENT-C	1.209 ± 0.009	1.482 ± 0.015	0.201 ± 0.009	0.988 ± 0.002	1.259 ± 0.007	0.359 ± 0.007
ERM with Coral layer	1.082 ± 0.003	1.203 ± 0.002	0.386 ± 0.002	0.774 ± 0.004	0.910 ± 0.002	0.497 ± 0.002
DANN	0.880 ± 0.006	1.025 ± 0.003	0.499 ± 0.003	0.724 ± 0.005	0.976 ± 0.004	0.518 ± 0.005
Mixup	1.018 ± 0.010	1.263 ± 0.009	0.297 ± 0.005	0.974 ± 0.008	1.176 ± 0.009	0.368 ± 0.005
Ours (PCOR-PA)	0.858 ± 0.003	0.974 ± 0.006	0.505 ± 0.005	0.602 ± 0.004	0.819 ± 0.004	0.582 ± 0.004

4.6.4.1 Comparison with research results on real-world pain databases

This table 4.11 presents a comparative analysis of various methods applied to a real-world pain database, focusing on their performance across both regression and classification tasks. The comparison spans different sources of pain, such as heat (DUG-CORAL), sickle cell disease, electric pulses, shoulder pain, laparoscopic appendectomy, and two conditions related to the Apon dataset (post-operative and labor).

For the regression task, the effectiveness of pain assessment methods is evaluated using Mean Absolute Error and Root Mean Square Error as metrics. Specifically focusing on the Apon data, the post-operative scenario shows an MAE of 0.858 and an RMSE of 0.974, indicating the level of deviation from the actual pain levels. These relatively high values suggest challenges in precisely estimating post-operative pain using Electrodermal Activity (EDA) as the sensor. In contrast, the labor pain scenario under Apon presents a lower MAE of 0.602 and an RMSE of 0.819, demonstrating a better performance in estimating pain levels more accurately for labor pain with the same type of sensor.

Pain sources	Input sensors	Regression		Classification
		MAE	RMSE	ACC
Heat (DUG-CORAL)	EDA	0.755	0.942	53.79%
Sickle cell disease (Yang *et al*., 2019)	HR, RR, GSR, SkinTemp, etc.	N/A	1.526	N/A
Electric pulses (Kong *et al*., 2021)	EDA	0.885	1.129	N/A
Shoulder Pain (Xu and de Sa, 2021)	video	0.63	1.13	N/A
laparoscopic appendectomy (Susam *et al*., 2021)	EDA	N/A	N/A	45.45%
laparoscopic appendectomy (Susam *et al*., 2021)	Video	N/A	N/A	77.27%
laparoscopic appendectomy (Susam *et al*., 2021)	EDA, Video	N/A	N/A	36.36%
Apon-postoperative (PCOR-PA)	EDA	0.858	0.974	40.20%
Apon-labor (PCOR-PA)	EDA	0.602	0.819	60.81%

TABLE 4.11: Comparison with the existing methods on real-world pain database

In terms of classification tasks, the accuracy percentage provides insights into how well the pain assessment methods can classify pain levels correctly. The Apon data reveals a stark contrast between the post-operative and labor scenarios. The post-operative pain assessment achieves a classification accuracy of only 40.20%, which is significantly low, highlighting the difficulty in accurately classifying pain levels in post-operative conditions. On the other hand, the labor pain scenario shows a much higher classification accuracy of 60.81%. This indicates that, despite the challenges, the method used for labor pain assessment is considerably more effective at classifying pain levels correctly compared to the post-operative scenario, though there is still room for improvement.

4.7 Conclusion and Future Work

The high subjectivity of pain, coupled with its resistance to quantification, presents significant challenges to the development of automated pain assessment algorithms. In this paper, we introduce a test-time adaptive learning framework for personalized pain models, PCOR-PA, which fully leverages the availability of pain data in medical settings and minimizes the prediction entropy of ordered regression networks. This approach enables pain prediction models to adapt to previously unseen test subjects, effectively aligning the model with the current test data distribution. PCOR-PA offers a novel solution for personalized pain assessment, surpassing existing methods in domain generalization.

In the future, we will focus on integrating facial expressions and physiological signals through multimodal learning; exploring how to construct data-balanced batches for each subject while preserving the advantages of subject-independent batches; and conducting in-depth analyses of the learned models to enhance their interpretability.

CHAPTER 5

A Data-Uncertainty Guided Approach for Effective Pain Assessment

---

Pain, a primary reason for seeking medical help, requires essential pain assessment for effective management. Studies have recognized electrodermal activity (EDA) signaling's potential for automated pain assessment, but traditional algorithms often ignore the noise and uncertainty inherent in pain data. To address this, we propose a learning framework predicated on data uncertainty, introducing two forms: a) subject-level stimulation-reaction drift; b) ambiguity in self-reporting scores. We formulate an uncertainty assessment using Heart Rate Variability (HRV) features to guide the selection of responsive pain profiles and reweight subtask importance based on the vagueness of self-reported data. These methods are integrated within an end-to-end neural network learning paradigm, focusing the detector on more accurate insights within the uncertainty domain. Extensive experimentation on both the publicly available BioVid dataset and the proprietary Apon dataset demonstrates our approach's effectiveness. In the BioVid dataset, we achieved a 6% enhancement over the state-of-the-art methodology, and on the Apon dataset, our method outperformed baseline approaches by over 20%.

5.1 Introduction

Pain assessment is crucial for healthcare professionals in devising effective pain management plans. Effective pain assessment involves a blend of subjective reports, objective measures, and computational models to gauge pain levels. Despite being the current gold standard, self-reported pain scales are inherently variable and subjective (Kong *et al.*, 2021), and are challenging to administer with infants or individuals with communication difficulties.

Complementing self-reports, objective measures like physiological indicators (e.g., heart rate, blood pressure, facial expressions) and behavioral indicators (e.g., mobility, sleep patterns) can provide valuable pain-related information. Leveraging Artificial Intelligence-Assisted Patient-Controlled Analgesia (AI-PCA) (Wang *et al.*, 2020b), objective and continuous pain state evaluation can be achieved through algorithmic approaches. However, due to the complexity of pain as a phenomenon (von Hehn *et al.*, 2012) and the lack of identified biomarkers reflecting its fundamental mechanisms (Van Der Miesen *et al.*, 2019), the search for objective biomarkers to elucidate pain mechanisms remains a pressing need (Tracey *et al.*, 2019).

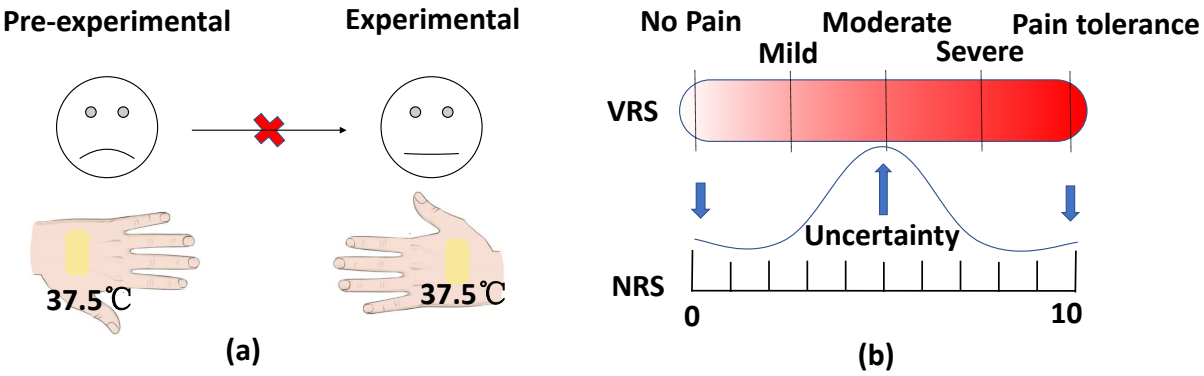


FIGURE 5.1: Examples of uncertainty in pain stimulation experiments include: (a) the inconsistency in stimulation-reaction, where pain perception varies between the pre-experiment and the formal experiment, and changes over time; (b) the uncertainty in subjective pain assessment, which increases as the rating approaches the middle of the pain scale.

The AI-PCA system utilizes facial expressions and bio-signals as inputs. While facial expression data can be noisy and resource-intensive in real medical settings, physiological signals, e.g., electrocardiograms (ECG), EDA, electromyograms (EMG), electroencephalograms (EEG), respiratory rates, and blood pressures are preferred due to their objectivity and lower equipment requirements. Biosignal-based methods have emerged as effective means for pain assessment, focusing on feature learning in time, frequency, and wavelet domains, along with sequence learning methods like LSTM and attention mechanisms.

Existing methods often overlook the intrinsic uncertainty present in pain data, potentially compromising the accuracy of pain assessment algorithms. These algorithms, based on

noisy data and approximate models, can produce inaccurate results, influencing physician decision-making and interpretation of intervention effects (Madden *et al.*, 2021). To fully harness the potential of machine learning algorithms in the AI-PCA system, it is crucial to quantify this uncertainty, typically represented as data uncertainty. Data uncertainty in pain data manifests as sensor noise, cognitive ambiguity in active reports, and drift in experimental pain response-stimulus. Each one could significantly impact pain assessment. It is unrealistic to assume noiseless sensors or control for changes in subject responses to experimental stimuli, let alone account for uncontrollable environmental variables in clinical data and their impact on the patient.

We present a new learning framework to tackle the challenges in pain assessment by considering uncertainty in labeled data. Our approach comprises three key components. First, subject batch normalization is utilized to stabilize data distribution by exploiting inherent similarities within data from the same subject. This reduces inter-subject data distribution differences. Second, we avoid incorporating ultrashort HRV features into the prediction model training due to their lower predictive capability compared to EDA signals. Instead, we leverage the low-noise nature of HRV features for assessing subjects with prior uncertainty. The dispersion of HRV features for each individual is calculated and used as a scaling factor during pain assessment model training. Third, to address pain ambiguity characteristics (high-low-high), we propose a pain-specific task weighting scheme combined with the coral framework. This conversion transforms the regression problem into a series of binary classification tasks (T_0 vs. T_1, T_1 vs. T_2, T_2 vs. T_3), and for tasks with high ambiguity, their loss weight is actively increased.

5.2 Data Uncertainty in Pain Assessment

Pain data uncertainty stems from two main sources. The first, subjectivity, arises from individual variations in pain sensitivity, influenced by personal characteristics and context, complicating objective measurement (Lundberg *et al.*, 2022). The second, cognitive ambiguity,

occurs when individuals find it difficult to discern intermediate pain states, neither pain-free nor at pain tolerance levels.

Pain studies gather data from two sources: experimental pain stimuli in healthy individuals and actual pain experiences in patients. The thermal stimulation pain experiment process consists of calibrating the stimulus and collecting data. Assuming consistency in stimulus intensities across both stages overlooks intraindividual variability (Madden *et al*., 2021), leading to shifts in stimulus-response relationships and changes in pain-affecting factors (Mosley and Butler, 2017). This results in data-label noise and uncertainty in automatic pain assessment modeling. In clinical settings, uncertainties arise from factors like patient physiology, movement, ongoing treatment, and external interference, necessitating careful evaluation.

Traditional subjective pain rating scales, such as the Verbal Rating Scale (VRS), Visual Analogue Scale (VAS), and Numerical Rating Scale (NRS), are standard in pain studies. VAS and NRS measure pain intensity on scales of 0 to 100 mm and 0 to 10 points, respectively, while VRS uses adjectives (Williamson and Hoggart, 2005). In thermal stimulation experiments, the Standard-Method-of-Limits calibrates stimulus intensity, identifying painless and pain tolerance thresholds but lacking intermediate pain level reports (Kamper-Fuhrmann *et al*., 2023). Assuming equidistant intermediate pain thresholds between these extremes is unreasonable (Walter *et al*., 2013). In clinical settings, patients' self-reported pain usually identifies no pain and pain tolerance accurately, but assessing intermediate levels introduces ambiguity.

Patient pain perception often involves significant uncertainty due to subjectivity (Zaman *et al*., 2021). While data uncertainty can be reduced with sufficient data (Kendall and Gal, 2017), the unique nature of pain experiments and limited pain-related databases leave uncertainty unaddressed in automated pain assessment algorithms. Recent work (Xu and de Sa, 2021) incorporates uncertainty into pain assessment algorithms, using the optimal linear combination model to reduce epistemic uncertainty from facial expressions, enhancing results.

5.3 Problem Statement

This paper addresses the practical issue of pain assessment in real clinical settings. We define a dataset as $D = \langle X, Y \rangle$, where $X = \{x_0, x_1, ..., x_T\}$ denotes a continuous stream of physiological signals, and $Y = \{y_0, y_1, ..., y_T\}$ corresponds to sequence of pain intensity over time. At any given time t, we have $y_t \in \{r_1, r_2, ..., r_k\}$, where r_1 denotes no pain and r_k signifies the most severe pain level. The task is to determine, at any time, whether a patient is in pain and, if so, the intensity of the pain. Accurate pain intensity estimation requires analyzing the sequence of signals, not just a single time point. The goal is to find a function $f : \mathcal{X} \to \mathcal{Y}$ that minimizes

$$\sum_{i=1}^{N} \sum_{t=0}^{T} l(f(X_{0:t}), y_t^i) \tag{5.1}$$

where N is the number of observed subjects and l denotes an error function, measuring the difference between $f(X_{0:t})$ and y_t^i. As per equation (5.1), the input of f includes a signal sequence within the time range $[0, t]$. The input sequence length, denoted as $O(t)$, increases proportionally with t. This setting becomes impractical due to the escalating computational and storage demands with the increase in t. Hence, we propose a model denoted as f with the following property:

$$f(X_{0:t}) = f(X_{t-\Delta t:t}) \tag{5.2}$$

for some small $\Delta t > 0$ and for all $t \geq 0$. We refer to the model with this property as a *short-term Markovian model* model. In this work, we consider models with $\Delta t \leq 6s$ as practical models, as suggested in (Kong *et al.*, 2021).

5.4 Our Approach

In this section, we present our DUG-CORAL framework for pain assessment. This framework includes two essential features for robust pain intensity modeling: 1) Uncertainty incorporation with automatic pain intensity prediction and 2) Addressing pain-specific self-reporting

ambiguities. As shown in Figure 6.2, the modeling process in our design is divided into two phases: a) conduct HRV analysis to estimate the subject's physiological response uncertainty, and b) consistent rank ordinal regression neural network training, where this uncertainty is integrated into the training process. We also include subject batch normalization to quantify uncertainty further. Additionally, we outline the design of a pain-specific task weighting scheme to handle self-reporting ambiguity.

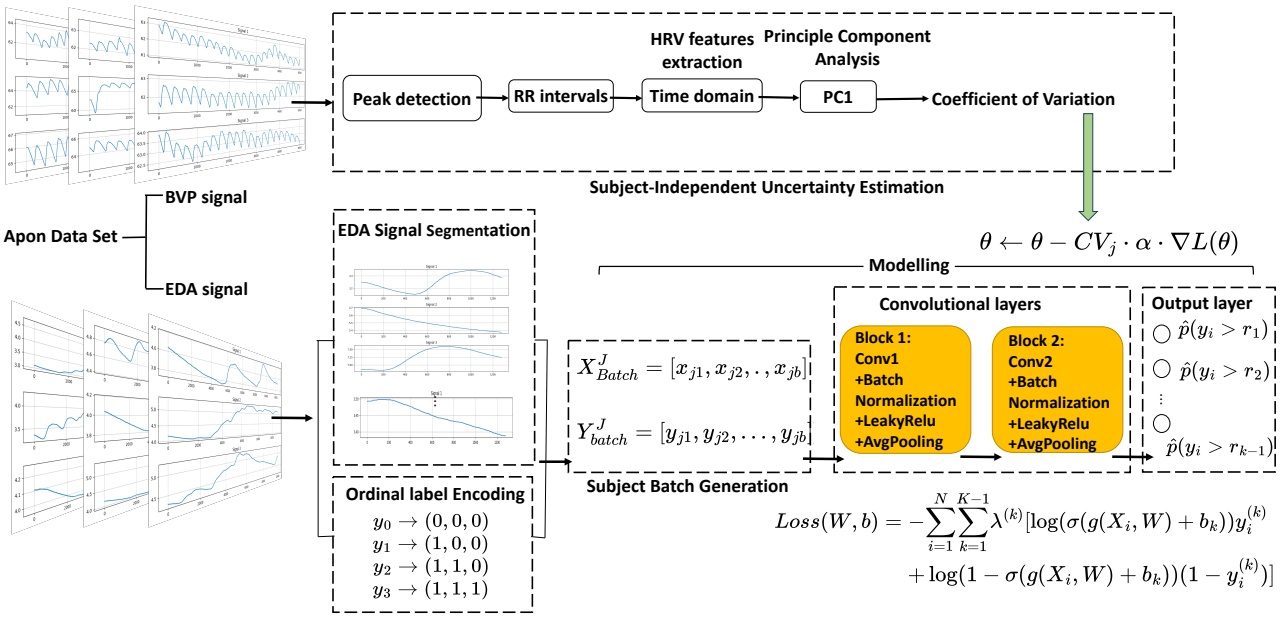


FIGURE 5.2: Framework of DUG-CORAL ranker on Apon database

5.4.1 Subject-Independent Uncertainty Estimation

Many algorithms for multi-modality-based pain assessment have discovered that EDA signals are more sensitive to pain than ECG signals. This sensitivity is not significantly enhanced by multi-modal fusion, likely due to the length of the ECG signal. In the study by (Jiang *et al.*, 2017), HRV analysis of 40-second data was found to be significant only for pain/no pain classification problems. However, a 40-second pain assessment time falls short of actual medical requirements in objective pain assessment algorithms. In this work, we thus depart from a multi-modal approach and utilize ECG data to assess patient uncertainty. This strategy aims to enhance the accuracy and generalization of the EDA-signal-based pain assessment algorithm.

Algorithm 3 Subject-independent uncertainty estimation

Input: A dataset consisting of n subjects, where each subject i has m_i signals (Heart related indicators used for HRV analysis).
Output: CV_i for each subject i where i from $0, 1, ..., n$

Let D be the entire dataset
for $i = 1$ to n **do**
    Let D_i be the subset corresponding to subject i from D
    for each *Signal* s_i^j in D_i **do**
        Detect peaks:
        peaks ← DETECTIONPEAK(s_i^j)
        Calculate intervals:
        RR_intervals ← CALCULATEDIFF(peaks)
        Extract features:
        features["MEAN", "SDNN", "RMSSD"] ← FEATUREEXTRACTION(RR_intervals)
    end for
    Apply PCA:
    PC1 ← PCA(features)
    Calculate coefficient of variation:
    CV_i ← CALCULATECV(PC1)
end for

Suppose that we are given a heart-related signals data set of N subjects which can be represented as $D = \{S_1, S_2, ..., S_N\}$. For each subject S_i, it contains a collection of m_i signals, the set of signals of each subject $S_i = \left\{x_1^i(t), x_2^i(t), ..., x_{m_i}^i(t)\right\}$. Each $x_j^i(t)$ is an time series data belonging to subject S_i, where $i = 1, 2, ..., n$, $j = 1, 2, ..., m_i$ and t is total length of signal. For each subject's signals, we first partition each signal $x_j^i(t)$ into segments of equal length $\{seg_1, seg_2, ..., seg_k\}$. Then, we merge these segments and extract each segment's HRV related features. Since the segmented signal fragments are relatively short, less than 6 seconds, we focus on extracting HRV-related features: $\{MEAN, SDNN, RMSSD\}$ in the time domain and nonlinear domain while disregarding frequency domain features. Let F_{S_i} be the HRV feature set for subject S_i, and its elements can be denoted as $F_{S_i} = \{F_{S_i,1}, F_{S_i,2}, F_{S_i,3}\}$. Later, to comprehensively assess the set of features extracted from these, Principal Component Analysis (PCA) is used to determine the major differences in these features. Standardizing the features to have a mean of zero and a variance of one ensures that features with varying scales do not wield disproportionate influence over the PCA process. Let $F_{std} = \frac{F-\mu}{\sigma}$, where μ is the mean of F, and σ is the standard deviation of F. Let C be the covariance matrix of

the standardized HRV feature F_{std}, which is provided as:

$$C = cov(F_{std}) \tag{5.3}$$

Let $\lambda_1, \lambda_2, ..., \lambda_n$ be the eigenvalues of C, and $v_1, v_2, ..., v_n$ be the corresponding eigenvectors and sort the eigenvectors based on their corresponding eigenvalues in descending order. Let v_{pc1} be the eigenvector corresponding to the highest eigenvalue λ_1. Let $pc1_{scores}$ be the scores of HRV feature set F_{std} projected onto $pc1$.

$$PC1_{scores} = F_{std} \times v_{PC1} \tag{5.4}$$

For physiological signaling, each subject's physiological state is different because the metric of their physiological data is different. Here, we choose to use the coefficient of variation instead of variance or range, which can well eliminate the effect of the scale of the physiological data of different subjects. We define the subject uncertainty as:

$$CV = \frac{STD_{PC1_{score}}}{MEAN_{PC1_{score}}} \tag{5.5}$$

which will be used in the further modelling phase. The overall process is described in Algorithm 1.

5.4.2 DUG-CORAL Training

For the final pain intensity inference model, we are aware of the ordinality of pain intensity labels and uncertainty of pain-related databases, i.e., observation uncertainty on experimental pain and label ambiguity. Here, we introduce the data uncertainty-guided ordinal regression algorithm for pain assessment. The only input to the pain assessment algorithm is the EDA signal. Feature learning for EDA signals is not our focus in this article. For simplicity, we use two layers of 1-dimensional CNN as the feature extractor. As an ordered regression problem, we employ CORAL (Cao *et al.*, 2020) for classification Loss .

5.4.2.1 CORAL framework

The idea behind the ordinal regression algorithm is to transform the ranking problem into a series of binary classification problems. According to the work in (Cao *et al*., 2020), a rank y_i is first extended into a vector of K-1 binary labels $y_i^{(1)}, ..., y_i^{(K-1)}$ such that $y_i^{(k)} = 1\{y_i > r_k\}$ indicates whether y_i exceeds rank r_k. The boolean test $1\{\cdot\}$ equals 1 if the inner condition is true and otherwise. By incorporating the expanded binary labels in the model training process, we can train the K-1 binary classifier in the output layer of the neural network.

Algorithm 4 Uncertainty guided training procedure

Input: EDA signal set X, pain rating y, Initial model parameterθ, learning rate α, Number of training epochs T, subject uncertainty $CV_{s_1,s_2,...,s_n}$
Output: trained model with updated parameters: θ_t

```
Initialize θ_t = θ_s
for epoch = 1 to T do
    for (X_batch_i, Y_batch_i) from subject i do
        compute the forward pass: f(X_batch, θ)
        compute the coral loss: L(F(X_batch, θ), y_batch)
        compute the gradient: ΔL(θ)
        Update model parameters: θ ← θ − CV_i · α · ∇L(θ)
    end for
end for
```

Let W denote the weight parameters of the neural network, excluding the bias terms in the final layer, and b_k denote the bias corresponding to the k_{th} output neuron. All neurons in the final output layers share the same weight to achieve rank consistency among its output layer tasks. The inputs of final layer is denoted $\{g(x_i, W) + b_k\}_{k=1}^{k-1}$. Let $\sigma(z) = \frac{1}{1+e^{-z}}$ be the logistic sigmoid function. The predicted empirical probability for binary classification task k is defined as:

$$\hat{P}(y_i^{(k)} = 1) = \sigma(g(x_i, W) + b_k). \tag{5.6}$$

For model training, we minimize the loss function:

$$\begin{aligned} L(W, b) = -\sum_{i=1}^{N}\sum_{k=1}^{K-1}\lambda^{(k)}[\log(\sigma(g(X_i, W) + b_k))y_i^{(k)} \\ + \log(1 - \sigma(g(X_i, W) + b_k))(1 - y_i^{(k)})] \end{aligned} \tag{5.7}$$

which is the weighted cross entropy of $K-1$ binary classifiers in the above equation. $\lambda^{(k)}$ denotes the weight of loss associated with the k^{th} classifier, which is the important parameter for task k. Considering the difficulty in optimizing $\lambda^{(k)}$, we opt for its selection through a non-uniform task weight scheme. This paper introduces our novel approach: an uncertainty-guided non-uniform weighting scheme tailored for automatic pain assessment.

5.4.2.2 Subject-independent standardization and batch normalization

In neural network training, there are two common techniques that can be combined to enhance training effectiveness. In this work, we propose subject-independent standardization and batch normalization. Firstly, regarding feature standardization, considering our dataset comprises data from various individuals, we have taken into account the diversity in data distribution across subjects. Rather than applying feature standardization across the entire dataset as typically done, we have chosen a different approach. We have individually processed the data of each subject, ensuring that the mean of their data features is centered around zero, with a variance of one. Secondly, during the preparatory phase of neural network training, we have adopted an alternative approach from the conventional method. Instead of randomly dividing the data into batches, we organize the data into batches based on individual subjects. To elaborate, each batch exclusively contains data from a single individual. This methodology elevates the significance of batch normalization. By calculating the mean and variance on each small batch of data, we normalize the inputs for every neuron. This practice efficiently mitigates internal covariate shifts during the training process, consequently enhancing the stability of the model's training process.

Suppose B_j is jth batch, we denote each batch as:

$$B_j = \{(X_i, y_i) \mid S_i = s_j\} \tag{5.8}$$

where s_j is the subject of jth batch

The operation of Batch Normalization can be represented as:

$$\text{BN}(X_i) = \frac{X_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \odot \gamma + \beta \tag{5.9}$$

where $\mu_i = \frac{1}{N_i}\sum_{n=1}^{N_i} X_i^{(n)}$ denotes the mean of the ith batch, $\sigma_i^2 = \frac{1}{N_i}\sum_{n=1}^{N_i}(X_i^{(n)} - \mu_i)^2$ denotes the variance of the ith batch, γ and β are the learnable parameters of the Batch Normalization layer, ϵ is a small positive value to avoid dividing by zero.

Constructing batches exclusively from the same subject for batch normalization provides two key benefits: 1) It reduces internal covariate shift by ensuring that the signals in each batch have more consistent distributions. 2) It avoids batch effects that can occur when combining data from multiple heterogeneous subjects in the same batch. Calculating batch statistics on these homogeneous batches enables more effective normalization of network activations by reducing variance across batches. This leads to more stable training and faster convergence, as demonstrated in our ablation study (Figure 3). To further enhance the performance of our model, we will investigate the construction of data-balanced batches for each subject while preserving the advantages of subject-independent batches.

5.4.2.3 Pain-Specific Non-uniform Weighting Scheme for Solving Label Ambiguity

The recognition system of humans indicates that people are more aware of the two limits of pain: no pain and pain tolerance, than of the intensity of pain in between. Therefore, we hypothesized that in the subjective report of pain (i.e., the labelling of the data), there is a smaller likelihood of greater ambiguity for two extremes of pain intensity; conversely, there is a high likelihood of greater ambiguity for intermediate pain scores. For this reason, we introduce a non-uniform weighting scheme to emphasize the ambiguity degree of data labels in automatic pain assessment. For training, we minimize the loss function.

5.4.2.4 Uncertainty-Guided Factor for Weight Adjustment

Previously, we obtained the data uncertainty for each subject by calculating the degree of discretization of their heart-related features, and we brought this uncertainty into the training

Model	BioVid		
	Input sensors	MAE	RMSE
Multi-stage ensemble classifier(Kächele *et al*., 2017)	EDA, ECG, EMG and videos	0.99	1.16
DDCAE (Thiam *et al*., 2021)	EDA, ECG, EMG	0.97	1.16
SVR (Pouromran *et al*., 2021)	EDA	0.93	1.16
Domain Adaptation (Rajasekhar *et al*., 2021)	Facial expression (Video)	1.16	N/A
SVR with trajectories (Szczapa *et al*., 2022)	Facial landmark coordinates (Video)	1.13	1.47
CORN (Ji *et al*., 2023)	EDA	0.90	1.22
DUG-CORAL (ours)	ECG, EDA	**0.83**	1.25

TABLE 5.1: Comparison with the existing methods on BioVid database

of the EDA-based ordinal regression neural network. To intern this, we used subject batch for the training of the neural network, where training batches are belongs to the same subject. The weight adjustment factor is generated by the uncertainty of the data for each object and it belongs to a certain range $(W_{min}, ..., W_{max})$, where $0 < W_{min} < 1$ and $1 < W_{max} < 2$ because we want to control the effect of a single subject on the model weights within a reasonable range to ensure the generalization ability of the model. Here we use the tanh function for scaling, which moves the mean of the data around zero.

$$CV_{scale} = \left(\frac{\tanh(CV) + 1}{2}\right) \cdot (W_{max} - W_{min}) + W_{min} \tag{5.10}$$

For these subjects $W < 1$, we consider the uncertainty in their data to be larger and therefore reduce the influence of these data on the parameters of the model. Conversely, for subjects whose $W > 1$, we consider their data to be of relatively high quality and therefore increase their impact on the model parameters.

We use HRV features as a scaling factor to reduce the impact of data from subjects with high uncertainty on the model parameters. In conjunction with the subject-independent batch training, the data batch of a subject with high uncertainty will have less influence on parameter updating and vice versa. In response to the second question, the dispersion of HRV features (Mean, SDNN, RMSSD) is used as a surrogate for uncertainty without the help of other signals. HRV tends to be relatively robust during pain experiments, with limited influence from environmental factors or individual variations in pain sensitivity or anxiety. A high

dispersion in HRV features could signify autonomic nervous system responses to varying levels of induced thermal pain.

5.5 Experiments

5.5.1 Datasets and Experimental Settings

The Apon dataset comprises data from 59 participants, including 30 males and 29 females, all healthy and inexperienced in pain assessment. The experiment involved three pain levels (mild, moderate, and severe), each subjected to 20 sessions of 30-second thermal stimulation, followed by a 90-second interval. Forty-seven subjects completed all three levels, while the remaining 11 only finished the mild and moderate levels. We utilized the eVu TPS (Thought Technology, 2023) wearable sensor to record psychological data, such as skin conductance level and blood pressure pulse, at a 256 Hz sampling rate. Data recording began 10 seconds before and ended 30 seconds after the heat stimulation. During the data preparation phase, we segmented the data using a 5-second sliding window, labeling segments before stimulus onset as "no pain" and those after as the corresponding pain level ("Mild", "Moderate", "Severe") based on the current temperature.

The BioVid heat pain dataset, containing bio-signals (ECG, EMG, and EDA), is the only publicly available resource of its kind. It is divided into two parts; we utilize part A, comprising 87 subjects with 100 data pieces each. These data correspond to different pain levels (T0, T1, T2, T3, T4), with 20 thermal stimulation experiments conducted for each level. Each signal recording lasts 5.5 seconds, with intervals of 8-12 seconds between adjacent stimuli. The original sampling rate is 512 HZ. To align with the Apon dataset's frequency, we downsampled the EDA signal to 256 HZ during data preparation. We excluded the EMG signal, as it corresponds to facial expression changes, and facial information related to pain is not considered in our work. We chose to use only Part A of BioVid to enable direct comparison to existing benchmarks, as most SOTA results are based on Part A inputs. Additionally, the

Method	Apon				
	Time	Frequency	Wavelet	Catch22	Tsfresh
SVR	1.44, 1.84	1.10, 1.40	1.15, 1.50	1.18, 1.44	0.95, 1.47
RF	0.96, 1.10	0.97, 1.13	0.98, 1.14	0.97, 1.10	0.94, 1.38
AdaBoosting	0.96, 1.08	0.96, 1.09	0.97, 1.09	0.97, 1.09	0.96, 1.45
CORAL	0.78, 1.01	**0.77**, 0.99	0.78, 0.98	0.80, 1.07	0.76, 0.95
DUG-CORAL	**0.77**, **0.98**	**0.77**, **0.98**	**0.77**, **0.97**	**0.79**, **1.05**	**0.75**, **0.94**

TABLE 5.2: Comparison with the baseline methods on Apon database. The results are provided in the form of (MAE, RMSE)

other parts B, C, and D differ in video (B – partially occluded faces, C – longer videos, D – posed pain and basic emotions) but not in ECG and EDA signals.

For both datasets, pain is simulated using thermal stimulation, which may constitute a bias for the study. To our knowledge, thermal stimuli is still the common approach to simulate pain as it can produce a reproducible pain experience in a controlled setting. BioVid is the only publicly available experimental pain dataset that contains the bioelectrical signals (ECG and EDA) we need. Other publicly available pain-related datasets, such as BP4D+ (cold stimuli), are too small to be used in our study, as they have only 140 samples compared to BioVid's 8700 samples.

5.5.1.1 Evaluation Protocols

In alignment with real healthcare scenarios, we investigate the subject-independent pain assessment problem, where training and testing biosignal observations come from different subjects. We utilize the leave-one-subject-out (LOSO) cross-validation method, consistent with existing literature. This approach involves cyclically selecting one subject's biosignals for testing and using the remaining subjects' biosignals for training. After each subject has been tested once, the average prediction performance is calculated. Our study targets an ordered regression challenge, leading us to choose Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as our regression metrics.

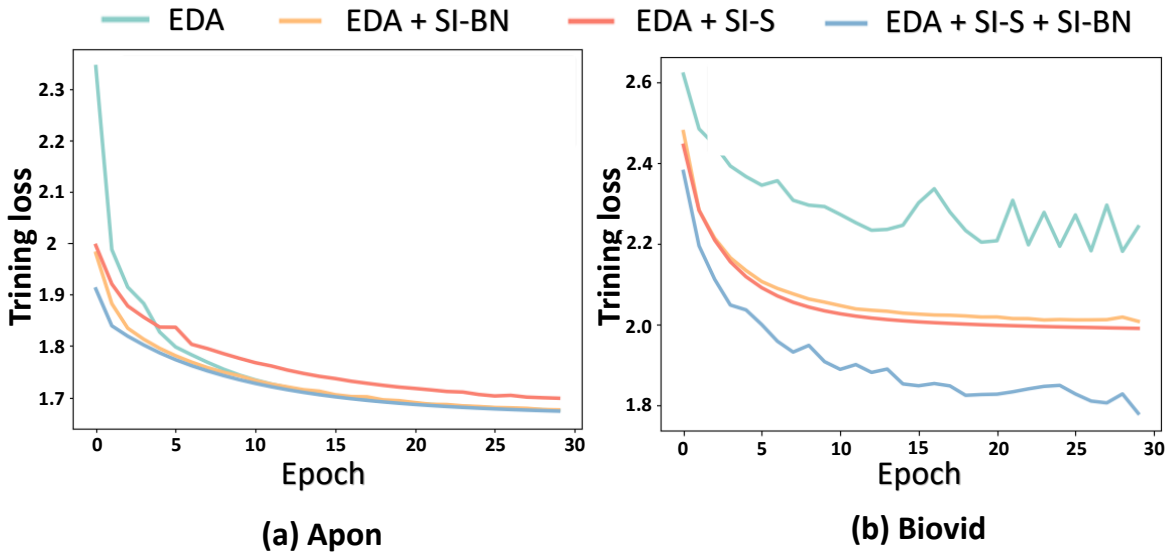


FIGURE 5.3: Ablation study and training loss. EDA refers to training process of DUG-CORAL without SI-standardization (SI-S) and SI-Batch normalization (SI-BN)

5.5.2 Results

5.5.2.1 Tests on BioVid

In line with previous work, we evaluate DUG-CORAL against state-of-the-art methods using the BioVid heat pain database, employing MAE and RMSE to assess regression performance. The comparisons are detailed in Table 5.1. It is worth noting that in our algorithm, the ECG signal was not utilized as an input for the prediction model but was employed to estimate each subject's uncertainty prior to model training. The table reveals that our algorithm significantly improves MAE, achieving state-of-the-art results. However, when considering RMSE, our method's effectiveness diminishes in comparison to other techniques. Given that RMSE emphasizes larger errors more than MAE, a plausible interpretation is that our method may produce a residual number of predictions with substantial error, although this error remains minimal in the majority of cases.

	BioVid				Apon			
HRV feature	Initial	[0.5, 1.5]	[0.8, 1.2]	[0.9, 1.1]	Initial	[0.5, 1.5]	[0.8, 1.2]	[0.9, 1.1]
MEAN	0.955, 1.562	0.852, 1.310	0.840, 1.311	0.835, 1.285	0.965, 1.418	0.885, 1.230	0.858, 1.166	0.854, 1.142
SDNN	0.913, 1.474	0.883, 1.399	0.856, 1.323	0.838, 1.362	0.947, 1.350	0.841, 1.160	0.842, 1.122	0.835, 1.135
RMSSD	0.903, 1.457	0.885, 1.401	0.852, 1.313	0.839, 1.288	0.860, 1.170	0.861, 1.168	0.825, **1.099**	0.836, 1.109
PC1 (time)	N/A	0.873, 1.401	**0.825**, 1.262	**0.826**, **1.248**	N/A	0.833, 1.155	**0.804**, 1.120	0.813, 1.135

TABLE 5.3: Comparison of performance using different scale range of HRV features (Mean, SDNN, RMSSD and PC1) on both BioVid and Apon databases. The results are provided in the form of (MAE, RMSE)

5.5.2.2 Tests on Apon

In order to compare the efficacy of our algorithms, we chose some benchmark methods, including support vector regressor (SVR), Random forest (RF), and adaptive boosting method (Ada boosting), which are some classic machine learning algorithms and some of them were used in the state of the art method in pain research. In order to explore the efficiency of our algorithm in different latent spaces, we extracted features from various domains, the time domain, frequency domain, and wavelet domain. Additionally, we utilized time series feature libraries such as Catch22 and TSfresh for feature extraction. Details are provided in Appendix (Xinwei Ji, 2023). As seen in Table 5.2, CORAL's prediction results demonstrate significant improvement across all feature spaces, while DUG-CORAL exhibits subtle enhancements based on the CORAL algorithm. We also evaluated the efficacy of our proposed algorithm using the Apon private dataset and compared the outcomes with those from other studies in real healthcare settings. The detailed results are provided in Appendix (Xinwei Ji, 2023). We used SVR as a baseline for both datasets. Three baseline methods from Table 1 (Multi-stage ensemble classifier, Domain Adaptation, SVR with trajectories) were excluded from the Apon dataset test because they include the video modality.

The details of feature extraction methods are provided as follows: we compare different feature extraction methods on the Apon dataset, focusing on features from the time, frequency, and wavelet domains, alongside two comprehensive feature libraries: Catch22 and Tsfresh.

Time Domain Features: Given a signal X with N samples, we compute the following time domain features:

- Mean of X: $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, representing the average value.
- Maximum value in X: $\max(X) = \max(X_i)$, the highest amplitude in the signal.
- Root Mean Square (RMS) of X: $\text{RMS}(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_i^2}$, quantifying the signal's power.
- Variance of X: $\text{Var}(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$, indicating signal variability.
- Standard Deviation of X: $\text{Std}(X) = \sqrt{\text{Var}(X)}$, the spread of the signal amplitude.
- Further features include Peak, Peak-to-Peak, Crest Factor, Skewness, Kurtosis (Excess Kurtosis), Form Factor, Pulse Indicator, Interquartile Range, and Waveform Length, each providing unique insights into the signal's characteristics.

Frequency Domain Features: After converting the signal X to its frequency representation using the Fast Fourier Transform $\mathcal{F}(X)$, we obtain the spectral power $S(k) = |\mathcal{F}(X)(k)|^2 / N$ for each frequency component k. Based on this, we compute the following features to characterize the frequency content of the signal:

- Maximum spectral power: $\text{freq\_max} = \max(S(k))$, highlighting the frequency with the maximum power.
- Sum of all spectral power values: $\text{freq\_sum} = \sum_{k=1}^{N} S(k)$, indicating the total power across all frequencies.
- Mean of the spectral power: $\text{freq\_mean} = \frac{1}{N}\sum_{k=1}^{N} S(k)$, offering an average power per frequency component.
- Variance of the spectral power: $\text{freq\_var} = \frac{1}{N}\sum_{k=1}^{N}(S(k) - \text{freq\_mean})^2$, measuring the dispersion of power across frequencies.
- Peak of the spectral power: $\text{freq\_peak} = \max(|S(k)|)$, identifying the peak power irrespective of frequency.
- Skewness of the spectral power: $\text{freq\_skew} = \frac{\frac{1}{N}\sum_{k=1}^{N}(S(k) - \text{freq\_mean})^3}{(\text{freq\_std})^3}$, with freq_std being the standard deviation of $S(k)$, captures the asymmetry of the power distribution across frequencies.
- Kurtosis of the spectral power: $\text{freq\_kurtosis} = \frac{\frac{1}{N}\sum_{k=1}^{N}(S(k) - \text{freq\_mean})^4}{(\text{freq\_std})^4} - 3$, assessing the "tailedness" of the power distribution, where a higher kurtosis indicates a distribution with heavier tails.

Wavelet Domain Features: Using the Discrete Wavelet Transform (DWT) with Daubechies wavelets, we decompose X into approximation and detail coefficients, cA and cD_n, respectively. From these, we compute:

- Mean and standard deviation for each order of Daubechies wavelets (cD_1 through cD_5), capturing both the average and variability of the detail coefficients, providing insight into the signal's hierarchical structure.

Catch22: The Catch22 library is a concise yet powerful tool, generating 22 features from the Hctsa (Highly Comparative Time Series Analysis) toolbox. These features, organized into six categories—distributions, simple time statistics, linear autocorrelation, nonlinear autocorrelation, continuous variance, and fluctuations—provide a broad analysis spectrum from a compact feature set.

Tsfresh: Tsfresh systematically extracts high-quality features from time series data, categorizing them into five main groups: distribution (both parametric and non-parametric measures), autocorrelation, frequency domain characteristics, linearity tests, and information content measures. This library focuses on comprehensive feature engineering to capture the essence of time series in various aspects.

5.5.3 Ablation Study

5.5.3.1 Impact of Subject-Independent Data Normalization

To explore the impact of subject-independent standardization (SI-standardization) and batch normalization (SI-Batch normalization), we conducted an ablation study, the results of which are depicted in Figure 5.3. Both SI-standardization and SI-Batch normalization expedite convergence in the Apon and BioVid datasets, with the effect being particularly pronounced in the BioVid dataset. The less apparent impact on the Apon dataset may be attributed to the nature of the Apon data, which was derived by slicing from a lengthy signal sequence, leading to inherent correlations within the data. Since SI-standardization and SI-batch normalization show notable performance improvement, we have externally validated the effectiveness of

this technique with real patient data. Further details can be found in Appendix (Xinwei Ji, 2023).

5.5.3.2 Impact of Scale Range of Subject Uncertainty Factor

We performed HRV analysis on each subject, calculating a discrete coefficient as an uncertainty factor for each subject. To explore the scaling range of this factor, we conducted experiments and compared the effects of three individual HRV features (MEAN, SDNN, and RMSSD) in the time domain with their main principal component. As shown in Table 5.3, the principal component's effect outperforms that of using MEAN, SDNN, and RMSSD individually. We also determined that the optimal scaling range is between 0.8 and 1.2.

5.5.3.3 Impact of Task Important Weighting Scheme

DUG-CORAL views the ordinal regression problem as a series of $K-1$ binary classification tasks (T_0 vs. T_1, T_1 vs. T_2..., T_{k-1} vs. T_k), and we assume that these tasks have different levels of difficulty due to the ambiguity of human cognition. Here, we compare DUG-CORAL using different task weighting schemes as below: 1) Scheme I: In the first scheme, the model is trained for tasks with an equal weighting scheme. 2) Scheme II: In the second scheme, we first randomly sample K-1 numbers from the standard normal distribution (Mean value is 0, the standard deviation is 1) and put them into the softmax function. Finally, we obtain numbers as task weight factors. 3) Scheme III: In the third scheme, we adopted an adaptive loss weighting scheme considering the homoscedastic uncertainty (Kendall *et al.*, 2018). 4) Scheme IV: In the fourth scheme, hard weighting scheme for BioVid is applied as $w_{T_0 vs T_1}$ = 0.4, $w_{T_1 vs T_2}$ = 0.1, $w_{T_2 vs T_3}$ = 0.1, $w_{T_3 vs T_4}$ = 0.4. For Apon, $w_{T_0 vs T_1}$ = 0.4, $w_{T_1 vs T_2}$ = 0.2, $w_{T_2 vs T_3}$ = 0.4. We adopted the principle that for tasks with low ambiguity, the importance is high. 5) Scheme V: In the fifth scheme, a hard weighting scheme for BioVid is applied as $w_{T_0 vs T_1}$ = 0.1, $w_{T_1 vs T_2}$ = 0.4, $w_{T_2 vs T_3}$ = 0.4, $w_{T_3 vs T_4}$ = 0.1. For Apon, $w_{T_0 vs T_1}$ = 0.3, $w_{T_1 vs T_2}$ = 0.4, $w_{T_2 vs T_3}$ = 0.3. We adopted the principle that for tasks with high ambiguity, the importance is high.

Dataset	Scheme	MAE↓	RMSE↓
BioVid	IV	0.865	1.351
	II	0.858	1.339
	I	0.845	1.294
	III	0.842	1.30
	V	**0.839**	**1.271**
Apon	III	0.785	0.998
	IV	0.782	0.996
	II	0.781	0.996
	I	0.776	0.986
	V	**0.773**	**0.980**

TABLE 5.4: Comparison of validation loss (MAE and RMSE) of all five task important weighting schemes on both BioVid and Apon databases

Results on both BioVid and Apon datasets are summarized in Table 5.4. Weighting scheme V outperforms the other schemes by giving more weight to the more challenging binary classification tasks. This approach balances the optimization process between tasks, ensuring that neither difficult nor easy tasks are neglected.

5.5.4 Additional comparison with other pain-related studies in real healthcare settings

From Table I, we can see that using the EDA signal alone, our experimental results are better than the work (Kong *et al*., 2021) on both MAE and RMSE. In contrast to the study using video as input, our algorithm is slightly better than the work (Yang *et al*., 2019) on RMSE, which indicates that there are fewer cases of huge errors in our algorithm. Additionally, we treat the problem as a classification problem and replace the validation loss with accuracy, and our algorithm achieved 53.79% accuracy on the 4-level pain intensity assessment, which is better than the work using only EDA signal (Susam *et al*., 2021).

5.5.5 External validation on data from real patients

In the experimental section, we have shown the efficacy of SI-standarlization and SI-batch normalization on apon and biovid datasets. Here, we present its efficiency on real patient database. The real patient dataset consists of two parts, postoperative patients and label

Pain sources	Input sensors	Regression		Classification
		MAE	RMSE	ACC
Heat (ours)	EDA	0.755	**0.942**	53.79%
Sickle cell disease (Yang *et al*., 2019)	HR, RR, GSR, SkinTemp, etc.	N/A	1.526	N/A
Electric pulses (Kong *et al*., 2021)	EDA	0.885	1.129	N/A
Shoulder Pain (Xu and de Sa, 2021)	video	**0.63**	1.13	N/A
laparoscopic appendectomy (Susam *et al*., 2021)	EDA	N/A	N/A	45.45%
laparoscopic appendectomy (Susam *et al*., 2021)	Video	N/A	N/A	**77.27%**
laparoscopic appendectomy (Susam *et al*., 2021)	EDA, Video	N/A	N/A	36.36%

TABLE 5.5: Comparison with the existing methods on Biovid database

pain patients, with a total of 17 postoperative patients and 23 label pain patients. Here, we categorized the VAS scores collected by medical workers into 4 levels, no pain, mild, moderate, and severe, corresponding to VAS score intervals of 0, 1-3, 4-6, and 7-10, respectively. From Figure 5.5.5, we can see that SI-standarlization and SI-batch normalization can still work well on real patient datasets.

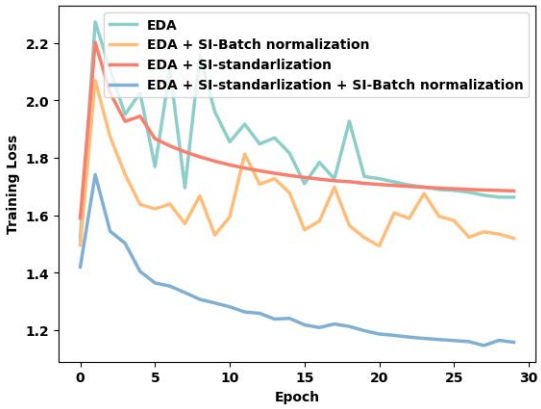


FIGURE 5.4: Ablation study and training loss on real patient database. EDA refers to training process of DUG-CORAL without SI-standardization and SI-Batch normalization

5.6 Conclusion and Future Work

Pain assessment and monitoring demand an objective, precise predictive model, but uncertainties in pain datasets and the absence of robust biological pain biomarkers hinder real-world

medical applications. In this paper, we introduce DUG-CORAL, a model based on the Ordered Regression Neural Network, enhanced with uncertainty assessment through HRV analysis and label ambigurity-based task weighting scheme. By relying solely on EDA signals, DUG-CORAL surpasses existing methods in pain prediction, including those utilizing facial expressions and multimodal data.

In the future, we will focus on integrating facial expressions and physiological signals through multimodal learning; investigating the construction of data-balanced batches for each subject while preserving the advantages of subject-independent batches; and conducting in-depth analyses of the learned models to enhance their interpretability.

CHAPTER 6

Automatic Pain Assessment with Ultra-short Electrodermal Activity Signal

---

Automatic pain assessment systems can help patients get timely and effective pain relief treatment whenever needed. Such a system aims to provide the service with pain identification and pain intensity rating functions. Among the physiological signals, the electrodermal activity (EDA) signal emerges as a promising feature to support both functions in pain assessment. In this work, we propose a machine learning framework to implement pain identification and pain intensity rating using only EDA and its derived features. Our solution also explores the feasibility of using ultra-short EDA segmentation of about 5 seconds to meet real-time requirements. We evaluate our system on two datasets: Biovid, a publicly available dataset, and Apon, the one we build. Experimental results demonstrate that using just the ultra-short EDA signal as input, our algorithm outperforms state-of-the-art baselines and achieves a low regression error of 0.90.

6.1 Introduction

Suffering in pain is an unpleasant experience for all patients. Nowadays, pain assessment has become an essential component in pain management. Unfortunately, the existing methods are laborious and require great care in handling (Kappesser and Williams, 2010). They are generally completed by either patient self-reporting or observed by others. For patient self-reporting, the subjective pain assessment often leads to not the best treatment plan for patients, particularly for those with communication disorders, e.g., dementia patients (Jonsdottir and

Gunnarsson, 2021). As observed by others, the observers could underestimate the pain suffered by the patients and create barriers to adequate pain relief (Achterberg *et al*., 2013).

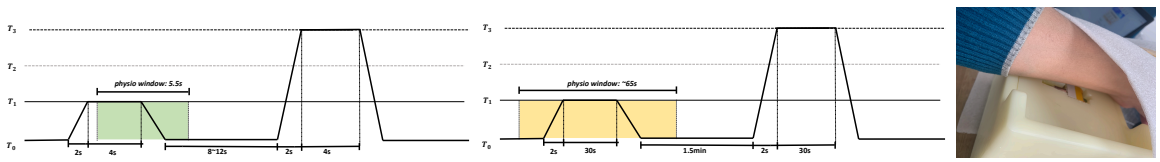
Automatic pain assessment is a new proposal to provide objective pain measurement while eliminating potential human errors. Some preliminary studies (Rajasekhar *et al*., 2021; Susam *et al*., 2021) have been proposed with the help of computing techniques. In these studies, facial expression and physiological signals are the two main approaches to achieving the goal and facilitating a proper response in time (Achterberg *et al*., 2020). Compared with facial expression, physiological signals are a more objective and quantitative measure to reveal pain as they directly reflect physiological changes in our bodies. Using machine learning to analyze physiological signals recognized as a valuable approach for pain assessment (Bhatkar *et al*., 2022; Posada-Quintero *et al*., 2021; Chu *et al*., 2017). It is important to note that these works like simultaneously using multiple physiological signals over a long period (lasting more than 10 seconds) to make decisions. Intuitively, a long signal often contains more information than a short one (lasting less than 10 seconds) to reflect patient conditions better (Leiner *et al*., 2012; Posada-Quintero and Chon, 2020). However, the acquisition of the needed and well-aligned physiological signals often relies on the availability of the corresponding sensors in a controlled environment. Sometimes, the physical conditions of the patients can also limit the monitoring of physiological signals. Patients with heart-related diseases could have abnormal electrocardiogram signals in nature.

In this work, we propose a low-latency automatic pain assessment framework with an ultra-short single physiological signal to address the above issues. The EDA signal is cherry-picked from numerous physiological signals to serve as the sole input for our framework. Our selection centers on the pain stimulus intensity being sensitively correlated with the human sympathetic nervous system. As a surrogate measure of sympathetic function, the EDA response is the most well-characterized body response signal to pain (Chae *et al*., 2022). To reduce latency, we design a lightweight pain identification model as a background daemon to detect if a patient undergoes pain by using only a limited number of EDA features. Upon pain is detected, we perform a pain intensity rating model using the EDA signal with a length of 5 seconds, which we refer to as the ultra-short signal. The ultra-short signal we used is

the segment starting from the pain detected. As a natural stress response, we believe this signal segment contains rich information about the pain experience beyond that provided by the other segments. Meanwhile, we integrate the feature libraries and feature engineering techniques to derive insightful features from the ultra-short EDA signal as input for the pain intensity rating model. With these strategies, our solution can significantly reduce the reliance on computational resources while still having accuracy assurance.

6.2 Data

Few publicly available datasets exist for pain research. Among the available ones, the BioVid dataset (Walter *et al*., 2013) is the most popular one. Due to the experimental design, the BioVid dataset contains physiological signals of only 5.5 seconds under a controlled environment. The length of the signals is not good enough to reflect real-world scenarios. We thus design and conduct new pain stimuli experiments to collect the physiological signals of 30 seconds to allow better reflection of the real-world scenarios. Our private dataset names Apon. Please note that we only use the EDA data from these datasets in this work.



(a) Biovid (b) Apon (c) Apon test

FIGURE 6.1: Exemplary temperature curve with alternating stimuli and pauses for (a) BioVid dataset, (b) Apon dataset, and (c) the test environment for building Apon dataset.

Part A of the BioVid Database focuses on pain intensity recognition. It contains physiological data (EDA, ECG, EMG) and videos from 87 participants. Data collection has two phases: heat threshold calibration and pain stimulation. The heat-induced pain stimulation is given to the participants at each phase. In the heat threshold calibration phase, each individual is requested to determine two critical points, T_s and T_t. The first point denotes the participant starts to feel pain when the pain stimulation device reaches the temperature of T_s, and the second point

denotes the participant fails to tolerate pain when set to T_t. In the pain stimulation phase, every participant experiences pain stimulation of 5 levels (20 times per level), resulting in a total of 100 observations per participant. BioVid labels the 5 levels as $T_0 < T_s < T_m^1 < T_m^2 < T_t$, where T_0 denotes no pain, T_m^1 denotes pain intermediate 1, and T_m^2 denotes pain intermediate 2. The time window for data collection is 5.5 seconds (as shown in Figure 6.1 (a)). The data sampling rate is 512Hz.

We create the Apon dataset with the assistance of Apon medical company. The Apon dataset includes data from 59 participants, of which 30 are male, and 29 are female. 47 out of 59 participants complete the experiment with three pain levels (mild, moderate, and severe). The remaining 11 participants complete the mild and moderate levels only. These participants are healthy and have no experience in pain assessment.

The pain induction is via thermal stimulation by placing an electrode pad under the forearm of the participants. When heat stimulation begins, the electrode rises and conducts heat directly to the body. The test environment is shown in Figure 6.1 (c). We conduct a pre-experiment as the threshold calibration phase in BioVid to reduce participant's subjectivity to thermal stimuli. A round of pain stimulation lasts 30 seconds. Once completed, the participant rests for 1.5 minutes before entering the next test round. The process is shown in Figure 6.1 (b). The level of pain stimulation at each round is randomized. Every level is repeated 10 times for each participant during the experiment. We record the start and the end timestamps for every stimulation.

A eVu TPS (unk, 2021) wearable sensor is used for recording psychological data (EDA and blood pressure pulse) with a sampling rate of 256Hz. We start recording the data 10 seconds before the test and stop recording 30 seconds after the test. The Apon dataset has 1363 effective observations, where 519 observations are for mild pain, 517 for moderate pain, and 312 for severe pain.

6.3 Method

The overall design of our low-latency automatic pain assessment framework is shown in Figure 6.2. It has three sequential stages, signal pre-processing, pain identification, and pain intensity recognition.

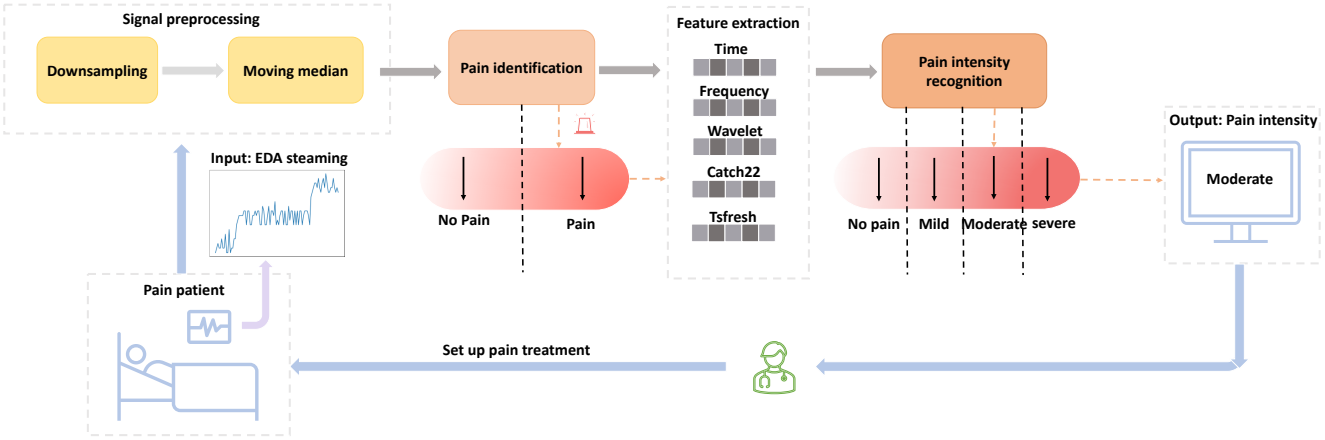


FIGURE 6.2: Framework of pain identification and intensity rating.

The signal pre-processing stage has two steps, downsampling and moving median. The former reduces time and memory complexity, while the latter smooths the data by removing body gestures and movements. We downsample the EDA signal to 32Hz to reduce the data size for analysis. Meanwhile, the signal still offers almost the same information (IMotions, 2017). Then, we filter the signal using a sliding window median of 0.5 seconds. Compared with a sliding window average, the median window can help reduce high intrinsic statistical noises and undesired irregularities (Leiner *et al.*, 2012). Once completed, we have the needed EDA signal for further processing.

The pain identification stage aims to indicate if a patient is in pain. In our framework, the pain identification model is a long-running process to determine whether the patient undergoes pain and segment an ultra-short EDA signal from a long EDA sequence to represent the pain. This segment will then be used as the input for the pain intensity rating task. These requirements impose the constraint that our design has to be computationally lightweight to provide real-time feedback.

We thus choose seven features consisting of MEAN and STD across three domains (time, frequency, wavelet) and the normalized SCL level (nSCL) for completing this task.

We apply Fourier and wavelet transforms to extract the features from the frequency and wavelet domains. The features from the frequency domain can provide us with discriminative and distinct phenotypes that are not found in the time domain. The wavelet features, on the other hand, explore the instantaneous frequency at each time point while retaining the time features of the signal. In addition, the normalized MEAN of EDA amplitude is a clinical means to distinguish between no pain and pain. We compute nSCL for an ultra-short signal using the following formula, which modifies the one proposed in (Treister *et al.*, 2012) by reducing both the time window in the pre-stimulus period (previously 60 seconds) and the current average skin conductance (previously 10 seconds) to 4 seconds and 1 second respectively:

$$\text{nSCL} = 100 \times \frac{\text{current average SCL} - \text{pre-stimulus average SCL}}{\text{pre-stimulus average SCL}}.$$

With these enhanced features, we can thus employ those machine learning algorithms with relatively simple structures rather than relying on complex neural networks to exact embedded information from the signal on the fly. In this work, we use the random forest (RF) model to determine whether the patient is in pain and segment the representative ultra-short EDA signal because it has good resistance to noise and introduces randomness, making it less likely to overfit the data (Aqajari *et al.*, 2021; Kächele *et al.*, 2015).

Pain intensity evaluation is the key to understanding the feeling of a patient. This stage aims to use only the EDA signal to estimate pain intensity. We argue that the existing subjective pain scales are short on clear boundaries between the defined levels. In practice, this issue could easily lead to different estimated levels (such as a 1- or 2-point difference) from the baseline ranked by various personnel. We thus propose a 4-level pain intensity rating to enable clear definitions of each pain level and the details are shown below.

At a time point t, a patient's EDA signal is $S = (S_1, ..., S_m)$ – $m = f_s \cdot T$, where f_s is the sampling rate, and T is a pre-defined small constant less than 10 seconds. Our task is to find a model F to estimate the probability of the patient's experiencing a pain intensity $I_t \in \{0, 1, 2, 3\}$. The pain intensity I_t represents no pain if $I_t = 0$, mild pain if $I_t = 1$, moderate pain if $I_t = 2$, and severe pain if $I_t = 3$. The pain intensity rating stage contains a sequence of three tasks: feature extraction, feature selection, and pain intensity modeling.

The following sections provide the details. In this step, we extract the feature representations in two ways: statistical features and feature libraries (tsfresh (Christ *et al.*, 2018) and catch22 (Lubba *et al.*, 2019)).

Statistical features. For the EDA signal, morphology over a long period, including its increase to the peak and decrease back to normal. The traditional processing method decomposes the EDA signal into phasic and tonic components. However, we focus on the ultra-short EDA signal, which is not likely to have a complete morphology. This need motivates us to explore other views to extract features that represent the EDA signal more completely. We thus explore the signal from time, frequency, and wavelet domains.

For the time domain, the features we used include MIN, MAX, RANGE, MEAN, MEDIAN, MODE, STD, RMS, MS, K ORDER MOMENT (CENTER AND ORIGINAL), SKEW, KURT, KURT FACTOR, WAVE FACTOR, PULSE FACTOR, MARGIN FACTOR and WAVEFORM LENGTH. For the frequency domain, the features include MAX, SUM, MEAN, KURT, SKEW, PEAK, and VARIANCE of the coefficients of the Fourier transformation on the signal. For the wavelet domain, the features include MEAN and STD of the detail coefficient in the wavelet decomposition to the signal using the Daubechies wavelet family of order 1–5 (db1 - db5). The wavelet transformation can extract both local spectral information and local temporal information. It also contains information about the transient signal segments of our interest.

Feature libraries. In addition to basic statistical features from multiple domains, we view the EDA signal as time series and use feature libraries catch22 (Lubba *et al.*, 2019) and Tsfresh (Christ *et al.*, 2018) to extract features. The employment of these feature libraries

can produce more diverse features rather than hand-crafted statistical features to describe the ultra-short signals. During the use, library catch22 produces 22 features from the Hctsa (highly comparative time-series analysis) toolbox (Fulcher *et al.*, 2013). These 22 features fall into six categories: distribution, simple temporal statistics, linear autocorrelation, non-linear autocorrelation, successive differences, and fluctuation. Similarly, library Tsfresh generates quality features from the input time series by systematic feature engineering. The extracted features fall into multiple categories: distribution (parametric and non-parametric), autocorrelation, frequency, linearity, and information content.

It is not hard to envision the generated features overlapping between the two feature libraries and statistic features in multiple ways. If these features are used directly as input for the training model, they could incur a high computational cost and violate our design principle. Hence, we conduct feature engineering on both databases to eliminate some features to build the final feature representation for pain intensity rating.

Feature Selection. In this step, We propose our feature selection approach following the competitive learning schema to select the key features from the raw ones generated above. Figure 6.4 gives the feature selection workflow diagram.

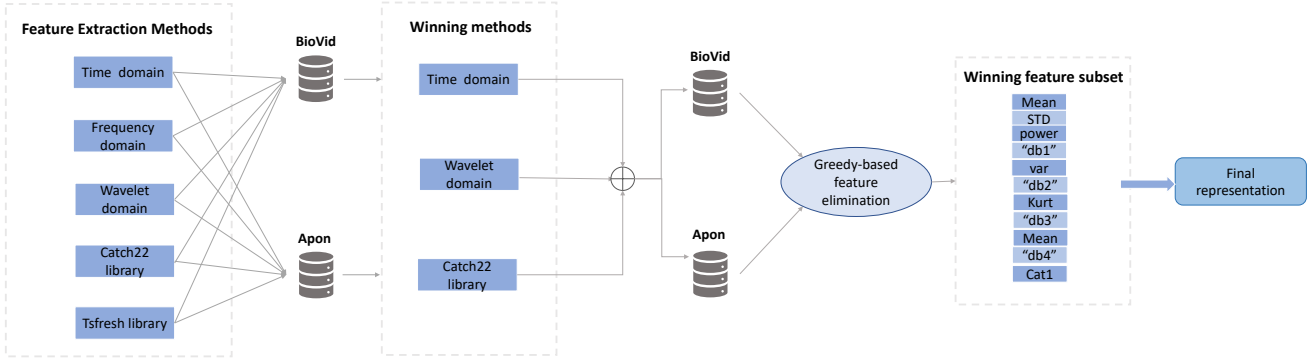


FIGURE 6.3: Competitive learning workflow in pain intensity rating task.

To observe the effectiveness of features in the pain intensity rating task, we build a random forest classifier with 1000 trees and use 10-fold cross-validation to examine the feature sets independently on BioVid and Apon datasets. Once completed, we can identify the feature

sets with the lowest contribution to the prediction accuracy on each dataset. The feature extraction methods for these identified feature sets will then be removed from the list. We call the remaining ones the winning methods.

To have a lightweight and trustworthy model to predict pain intensity, we continue to remove those features with a marginal contribution to the prediction accuracy in the feature sets. By marginal contribution, we mean that the removal of the feature does not decrease the prediction accuracy. We propose a greedy feature elimination algorithm to eliminate the features with a marginal contribution in rounds. Algorithm 5 gives the pseudo-code of the approach. In the pseudo-code, X denotes the dataset (a $n \times m$ matrix if there are n samples each with m features), R denotes a feature set, and $X(R)$ denotes X with just the columns defined by a feature set R. A metric function J takes a dataset and a trained model as input and returns the accuracy of the model. Finally, f is any predictive model that can learn the function from the feature space to the labels. In our work, we use the random forest as f.

Algorithm 5 Greedy feature elimination

Input: Feature set $R = \{R_1, ..., R_m\}$, sample features X, sample labels y, a metric function J, and a model f with adjustable features.
Output: A subset of feature space: the features of high relevance with the pain intensity.

train f using data $X(R)$ and labels y.
score $\leftarrow J(X(R), y, f)$
while true **do**
    found $\leftarrow$ *false*
    for $R_i \in R$ **do**
        train f, from the scratch, using data $X(R - \{R_i\})$ and labels y.
        score$'$ $\leftarrow J(X(R - \{R_i\}), y, f)$
        if score$'$ $=$ score **then**
            found $\leftarrow$ *true*
            $R \leftarrow R - \{R_i\}$
            break
        end if
    end for
    if not found **then**
        break
    end if
end while
return R

With R initially set to the winning features, we run Algorithm 5 to eliminate the features with marginal contributions. The left ones form the final learning representation. The final feature

set are standardized by the personal data to have zero-mean and unit-variance. In essence, this step removes the variations across subjects.

Pain intensity modeling. Our pain intensity rating module operates on the final representation with the rank-consistent ordinal regression neural network (CORN) (Shi *et al.*, 2021). For our datasets, we have modelled the pain intensities ranking into a discrete ordered order $no\ pain < mild\ pain < moderate\ pain < severe\ pain$. Typical ordinal regression follows the divide-and-conquer schema and converts the multi-class regression problem to a series of binary classification problems. However, the rank inconsistency problem in ordinal regression approaches may result in multiple non-adjunct ranks when their probabilities are above the threshold. For example, the patient gets no pain and severe pain simultaneously. The work (Cao *et al.*, 2020) achieves rank consistency among its output layer tasks by imposing a weight-sharing constraint. CORN takes one step further to solve the weight-sharing constraint by using conditional training sets to obtain the unconditional rank probabilities by applying the chain rule for conditional probability distributions. As mentioned, the pain intensity was encoded and represented by discrete values. The CORN ordinal regression algorithm is thus employed to fully explore the encoding to capture the ordinal relationship on pain intensity rating.

Complexity Analysis. Given a L-length signal with frequency f, the input of our design is an array of $n = L \cdot f$ numbers. In the signal pre-processing part, the downsampling step uses Fast Fourier Transformation (FFT) filter to the array and its complexity is $O(n \log n)$. The moving median also has $O(n \log n)$ complexity for maintaining a balanced binary search tree of w elements, where w is the window size, as the algorithm scans the windows through the array.

The pain identification algorithm first takes $O(n \log n)$ time to compute each of the D features. The random forest then takes $O(T \cdot H)$ time to infer the decision, where T denotes the decision trees in the random forest, and H is the maximum height of a tree. The complexity of this step is $O(n \log n)$ with pre-defined constants $D = 7$, $T = 1000$, and $H \leq D = 7$.

The pain intensity rating has $O(D' \cdot n \log n + D' \cdot M_1 + M_1 \cdot M_2 + M_2 \cdot K)$ complexity for inference, where D' is the feature number for pain intensity rating, M_1 the first hidden layer size, M_2 the second hidden layer size, and K the intensity level number (our CORN has 2 hidden layers). Computing all features requires $O(D' \cdot n \log n)$ time; with pre-defined constants $D' = 40$, $M_1 = M_2 = 256$, and $K = 4$ for Apon or 5 for BioVid, the overall complexity is $O(n \log n)$.

To summarize, the overall computational complexity of our design, including signal pre-processing, pain identification, and pain intensity rating, is $O(n \log n) + O(D \cdot n \log n + T \cdot H) + O(D' \cdot n \log n + D' \cdot M_1 + M_1 \cdot M_2 + M_2 \cdot K)$ in general. It can achieve $O(n \log n)$ with pre-defined constants D, T, H, D', M_1, M_2, and K.

6.4 Experimental settings and results

To verify our design, we evaluate our framework from individual modules, pain identification and pain intensity rating, to a whole.

Experiments on the BioVid dataset. Pain intensity in the BioVid dataset is divided into five levels. Like the previous work on BioVid database, we transformed the pain identification problem into a series of binary classification problems: T_0 vs T_1, T_0 vs T_2, T_0 vs T_3, T_0 vs T_4. Representation learning and pain intensity rating have been converted to five-class classification and regression tasks.

Experiments on the Apon dataset. There are 3-levels of pain intensity included in the Apon database: mild, moderate, and severe. Each observation contains signals before the pain stimulus's onset and after the pain stimulus's end. Two pointers, S_s and S_e, represent the start point and end point of a sliding window. We set the start of thermal stimulation as the pain start point P_s and the end of thermal stimulation as the pain endpoint P_e. When $S_s > P_s$ and $S_e < P_e$, we mark the current time segment as a pain segment, otherwise as a non-pain segment. We set sliding windows at different sizes, 5s, 6s, 7s, 8s, and 9s, to extract pain and non-pain segments from the signal stream. Using the extracted non-pain segments and pain

segments under different pain levels, we have the following binary classification tasks for pain identification: Non-pain vs. Mild pain, Non-pain vs. Moderate pain, and Non-pain vs. Severe pain. Representation learning and pain intensity recognition have been converted to four-class classification and regression tasks.

Evaluation metric. To evaluate the regression performance of different models, we used the metrics MAE and RMSE. Their definitions are shown below.

$$MAE = \sum_{i=1}^{n} |x_i - y_i| \tag{6.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{6.2}$$

where x_i is the true value, y_i is the prediction, and n is the number of samples.

To analyze the classification results, we calculate the confusion matrix. The element of this matrix includes true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which need to be calculated beforehand. Further, we calculate estimators: accuracy (ACC), precision (PREC), recall (SENS), and F1-score based on these elements. Given below the definitions of the performance metrics

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.3}$$

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{6.4}$$

$$F1score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{6.5}$$

We apply the leave-one-subject-out (LOSO) cross-validation method, which is the most robust way to test a model that contains data on a participant level. The BioVid database and Apon

TABLE 6.1: Pair-wise binary classification with LOSO validation.

Method	BioVid			
	T_0 vs T_1	T_0 vs T_2	T_0 vs T_3	T_0 vs T_4
EDA:7 features (ours)	58.1	64.7	71.3	80.4
All video + bio (Werner *et al*., 2014)	49.6	60.5	72.0	80.6
DDCVAE (Thiam *et al*., 2021)	N/A	N/A	N/A	84.2
SVM with EDA (Pouromran *et al*., 2021)	N/A	N/A	N/A	83.3
Video + biomedical (Kächele *et al*., 2015)	N/A	N/A	N/A	83.1

database have 87 and 59 subjects, respectively. Hence, there are 87 and 59 rounds for each dataset where each subject has been used as the testing subject once.

6.4.1 Performance of Pain Identification

The first set of experiments aims to evaluate the effects of the computational feature set for pain identification. Classification experiments for both BioVid and Apon databases were performed. We used the random forest with parameters of 1000 trees.

We compared our result to four recent studies (Thiam *et al*., 2021; Pouromran *et al*., 2021; Werner *et al*., 2014; Kächele *et al*., 2015) on the BioVid dataset. Table 6.1 presents the mean accuracy across all subjects on each prediction task mentioned above. As shown in the results, our model is sensitive to catching slight pain feelings. It outperformed the benchmarks in T_0 vs T_1 and T_0 vs T_2 tasks. In the rest tests, our model still shows comparable performance to the benchmarks but only requires a small number of features to support the decision-making. This element implies that our model is simple and lightweight in resource consumption, thus having better hardware adaptation.

The results of the same test on the Apon dataset are given in Table 6.2, which compares the pain identification performance on the binary classification tasks using the different window sizes to segment the signal. It is not hard to see that the window size of 5.0 seconds has the best performance in pain identification compared with other sizes. This result states that the ultra-short signal choice used in our design is practical in real-world scenarios.

TABLE 6.2: Pair-wise binary classification with LOSO validation.

	Apon		
window size	T_0 vs T_1	T_0 vs T_2	T_0 vs T_3
5s	58.16	62.75	58.69
6s	57.74	62.78	57.87
7s	56.80	61.45	57.87
8s	55.14	60.63	57.85
9s	55.83	59.36	57.62

6.4.2 Performance of Pain Intensity Rating

The second set of experiments were evaluated from three aspects: feature extraction, feature selection, and modeling. We transformed the pain intensity rating task into a classification and a regression problem to validate our design on both BioVid and Apon datasets.

6.4.2.1 Comparison of Feature Extraction

The comparison of feature extraction approaches were performed in two rounds. In the first round, the goal is to evaluate each approach individually. In the second round, the combinations of winning feature sets are evaluated. Here, we use the RF classifier trained with 1000 trees.

1) Feature elimination. Tables 6.3 and 6.4 show the prediction performance of different feature extraction methods on the BioVid and Apon datasets. By comparing the performance of the feature extraction methods on the two datasets, we see that the results are inconsistent. As shown in Table 6.3, the features extracted from the Tsfresh library showed the best prediction results in the evaluated parameters, followed by Catch22, time, and wavelet features. The frequency features performed the worst. Conversely, from the Apon results shown in Table 6.4, we find that Tsfresh had the worst performance. We believe such inconsistency is due to the difference between the experimental settings of the two datasets. We removed the worst-performing feature set from each test. The remaining feature extraction methods are the Catch22 feature, time domain features, and wavelet features using wavelet transformation.

TABLE 6.3: Evaluation of feature extraction on BioVid dataset.

	BioVid			
Features	ACC	PREC	SENS	F1-SCORE
Time (14)	28.2	28.8	28.2	26.8
Frequency (7)	26.0	25.9	26.1	25.1
Wavelet (10)	31.2	30.9	31.2	29.3
Catch22 (22)	31.7	31.1	31.7	30.1
Tsfresh (93)	33.3	32.7	33.3	31.2

TABLE 6.4: Evaluation of feature extraction on Apon dataset.

	Apon			
Features	ACC	PREC	SENS	F-SCORE
Time (14)	44.5	44.9	44.5	44.0
Frequency (7)	37.5	37.5	37.5	37.2
Wavelet (10)	36.5	36.5	36.5	34.7
Catch22 (22)	34.1	33.6	34.1	31.6
Tsfresh (93)	29.7	29.1	26.5	25.5

TABLE 6.5: The winning features on BioVid and Apon datasets.

	BioVid chance score: 20%				Apon chance score: 25%			
	ACC	PREC	SENS	F1-SCORE	ACC	PREC	SENS	F1-SCORE
catch22 + wavelet + time	35.8	33.6	35.8	34.0	58.2	59.1	58.2	57.8
catch22 + wavelet	35.3	33.3	35.3	33.6	50.5	51.6	50.5	49.6
catch22 + time	35.1	33.3	35.1	33.5	55.2	55	55.2	54
time + wavelet	34.8	32.6	34.8	32.7	47.6	48	47.6	47.2

2) Combination of winning feature sets. The winning feature sets from the first round experiments were Catch22 and the statistic features in time and wavelet domains. We then explored all the combinations of these winning features set to study their performance. The combination cases were defined as Catch22 + time domain, Catch22 + wavelet domain, time domain + wavelet domain, and Catch22 + time domain + wavelet domain. Table 6.5 shows the test results for the pain intensity rating with different feature combinations. The feature combination of Catch22, time, and wavelet features achieved the best performance on all evaluation metrics on both BioVid and Apon databases.

6.4.2.2 Evaluation of Greedy Feature Selection

We executed the greedy feature elimination algorithm on two datasets separately. We present the percentage improvement in accuracy over the chance score of corresponding multi-classification tasks. The definition is given below:

$$\text{accuracy improvement (\%)} = 100\% \times \frac{\text{accuracy} - \text{chance score}}{\text{chance score}}$$

As shown in Figure 6.4, the algorithm was run for 14 iterations on the BioVid dataset. When the algorithm iteratively eliminated 14 features, the model's prediction performance kept increasing until the algorithm traversed all the remaining features. Accuracy improved from 74% to 76.3% compared to the chance score for five pain-level classifications.

Similarly, the algorithm iterated nine times on the Apon dataset, and the result is shown in Figure 6.5. As shown, the prediction accuracy improved as the features that do not contribute to the decision-making were continuously eliminated. Accuracy improved from 111.6% to 114.1% compared to the chance score for four pain-level classifications.

From the results of the two databases, we filtered out the six features that overlapped and removed them from the final feature representation. The eliminated features were: (1) mode of z-scored distribution (5-bin histogram), (2) first 1/e crossing of autocorrelation function, (3) first minimum of autocorrelation function, (4) time-reversibility statistic, $(x_{t+1} - x_t)^3$, (5) proportion of successive differences exceeding 0.04σ, and (6) first minimum of the automutual information function.

6.4.2.3 Performance of Pain Intensity Rating (Regression Problem)

To verify the pain intensity evaluation module, we transferred the final pain assessment problem into a regression problem because its results are closer to the pain scores from self-reporting. We employed SVM (support vector machine) and RF (random forest) as the performance benchmarks since they have been used in pain intensity rating studies (Chu *et al.*, 2017), (Susam *et al.*, 2021), (Vijayakumar *et al.*, 2017) and achieved noticeable results. The parameters of the benchmarks were as follows: an RF with 100 trees and entropy as split

FIGURE 6.4: Accuracy improvement of the greedy feature elimination on BioVid dataset.

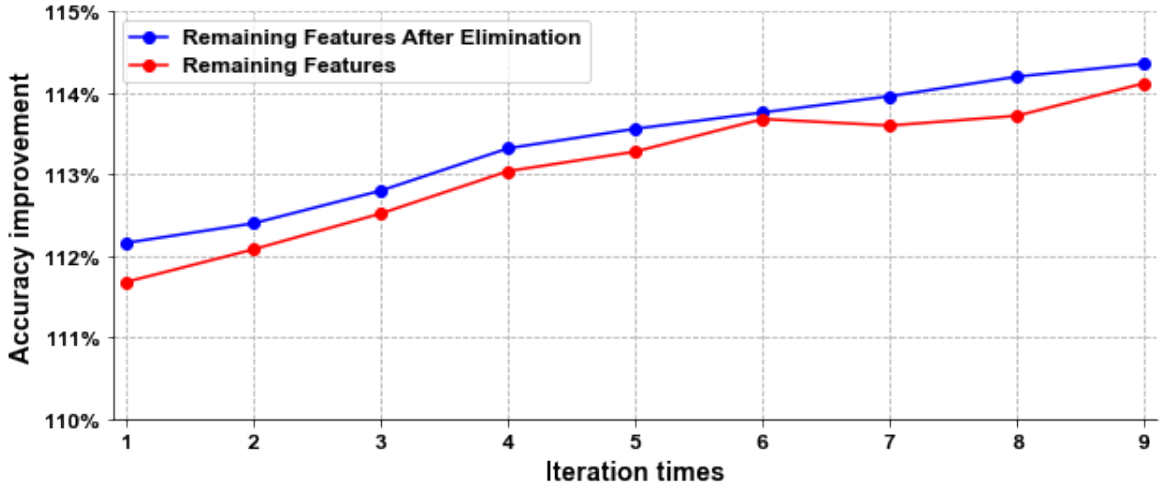
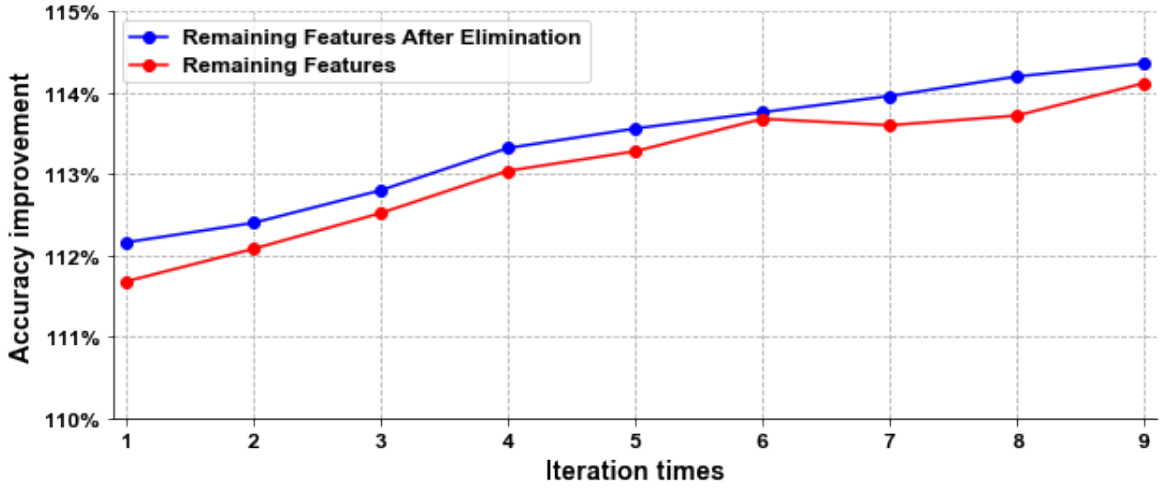


FIGURE 6.5: Accuracy improvement of the greedy feature elimination on Apon dataset.



criteria, an SVM with Radial basis kernel and cost = 1, and CORN with two hidden layers, 256 hidden units, leaky ReLU as the activate function, Adam optimizer, 0.001 learning rate, and 20 epochs in training. As shown in Table 6.6, CORN achieved the lowest MAE of 0.90 and 1.11 on BioVid and Apon datasets, respectively. The highest MAE of 0.93 and 1.19 were obtained by using RF.

TABLE 6.6: Performance of pain intensity rating regression models with personal feature representation (mean ± standard deviation).

	BioVid		Apon	
Model	MAE	RMSE	MAE	RMSE
RF Regressor	0.93 ± 0.19	1.11 ± 0.21	1.19 ± 0.21	1.48 ± 0.19
SVM Regressor	0.91 ± 0.21	1.11 ± 0.23	1.14 ± 0.11	1.31 ± 0.10
CORN	0.90 ± 0.23	1.22 ± 0.24	1.11 ± 0.08	1.55 ± 0.08

TABLE 6.7: Running time and regression error of the overall system on Apon dataset.

	Apon	
Method	Running time	Regression error
Direct pain evaluation	0.0030	1.13
Pain detection + evaluation	0.0020	1.23

TABLE 6.8: Pain prediction performance on BioVid dataset.

Study	Method	Sensor	Performance	
			MAE	RMSE
Thiam et al. (Thiam *et al*., 2021)	DDCAE	EDA, ECG, EMG	0.97	1.16
Pouromran et al. (Pouromran *et al*., 2021)	SVR	EDA	0.93	1.16
Our study	CORN	EDA	0.90	1.22

6.4.3 Overall System Performance

To verify the overall performance of the system, we conducted a comparison study. We first set a sliding window of about 5 seconds. We performed the automated pain assessment in two different ways. The first one is under each sliding window, a comprehensive feature library was extracted for pain level assessment. The second one is we first identified pain under each sliding window, then intercepted the signal segmentation as input for the pain intensity rating module if identified. We analyzed the results from two perspectives: time efficiency and prediction. Table 6.7 shows that, compared with continuously evaluating pain intensity, our proposed system achieved a $(0.0030 - 0.0020)/0.0030 = 33.3\%$ improvement in running time with a loss of only $(1.23 - 1.13)/1.13 = 8.8\%$ on regression error. Table 6.8 shows that, our study, using just an EDA signal, outperforms the state-of-the-art in MAE, with negligible higher cost in RMSE.

6.5 Conclusion and Future Work

This work presents an automatic pain assessment framework for pain identification and intensity classification. Our framework uses the EDA signal only. Our system aims to provide timely pain identification and accurate intensity classification with low computational needs. To do so, we use downsampling, ultra-short signal, small feature sets, and lightweight models to achieve the goal while still delivering high-accuracy results. We evaluate our framework on two EDA datasets, Bovid and Apon, built by heat-induced pain experiments. The experimental results show that the ordinal neural network performed the best on both BioVid and Apon databases. Our system outperforms the state-of-the-art performance. Additionally, the result of an ablation study shows that our proposed system can effectively improve the efficiency of the automatic pain assessment algorithm.

Despite the promising results, our work needs further study to be well understood. The datasets used in this work include just the pain induced by thermal stimulation, where such data do not adequately represent everyday pain but experimental pain. Studying how to use ultra-short EDA signals in managing everyday pain is our future work. We are also keen on adjusting our model to real-world scenarios and making it more robust in practice.

CHAPTER 7

Conclusion

In this thesis, we propose a framework called "Personalized, Uncertainty-Aware, Trustworthy Algorithm" to address the difficulties of traditional patient self-report-based pain assessment. Our framework uses only two physiological signals as input data, EDA and ECG, as they provide more objective metrics than facial expressions. ECG is used as an input to the framework only to assess the uncertainty of the patient's data and is not involved in the reasoning process. We present the collection of pain data in detail, which includes an experimental pain dataset induced by thermal stimulation and real pain patient data, mainly from post-surgical patients and women in labor. We also present two key issues unique to pain assessment algorithms: uncertainty and subjectivity. We design separate modules to address these two issues. First, we introduce a method for assessing uncertainty in individual data using HRV features as a reference and bringing uncertainty as a scale factor into the training of pain inference models, which can effectively allow models to focus on data with high confidence and reduce the impact of "dirty" data on model training under the premise of ensuring the amount of data. Considering that pain, as a specific human emotion, is highly subjective, we introduce a test-time adaptive approach to make the trained model predictive even on new patient data. We show how the model can be updated in real-time to adapt to the data distribution of a new subject when a patient who has never appeared in the training set arrives. The framework not only takes into account the algorithmic level design but also the application of the algorithms to improve the efficient solution. Considering the trade-off between the algorithmic effectiveness and computational complexity, we give a hierarchical prediction process with the premise of guaranteeing the algorithmic prediction of the effectiveness of the algorithm in the low computational requirements of the normal

operation of the algorithm. The algorithms in the whole framework show how to advance valid information in objective physiological signal metrics while being aware of unperfect data and labels, as well as patient-to-patient inconsistencies. Together, they provide new solutions for the application of automatic pain assessment.

7.1 Future outlook

There is still a lot of experimental and theoretical work to be done in the future.

7.1.1 Pain database

The first step is the creation of a pain dataset, which is essential for the development of pain algorithms, and pain itself has a certain complexity and diversity. The first step in the collection of pain data is to design the data collection protocol and get it approved by the ethics committee, followed by finding the target participants and signing the informed consent form. The data collection process need to be standardized so that the pain data collected can be reused and compared across studies, algorithm development, and clinical practice. The selection of participants also took into account the diversity of demographics, including different ages, genders, ethnicities, and cultural backgrounds, to increase the generalizability and fairness of the algorithm.

7.1.2 Multi-modal learning

For pain, there is no biomarker for pain in medical science, and the EDA signal responds to the change of skin conductivity, which has been found to have good predictive effect on emotion prediction, pain assessment, and stress assessment from previous studies, and is better than other modalities, such as ECG, EEG, EMG, or facial expression. Multimodal learning must be the ultimate form of pain assessment algorithms in the future, because pain is a multidimensional experience, including not only physiological sensations, but also psychological states, emotional changes, and social environment effects. Second, relying

on a single data source may overlook other important aspects of the pain experience or lead to inaccurate conclusions due to data quality issues. Multimodal schools can improve the reliability of their algorithms by using multiple data sources to compensate for the limitations of their respective data sources.

References

[Achterberg *et al*.2013] Wilco P Achterberg, Marjoleine JC Pieper, Annelore H van Dalen-Kok, Margot WM De Waal, Bettina S Husebo, Stefan Lautenbacher, Miriam Kunz, Erik JA Scherder, and Anne Corbett. 2013. Pain management in patients with dementia. *Clinical interventions in aging*, 8:1471.

[Achterberg *et al*.2020] Wilco Achterberg, Stefan Lautenbacher, Bettina Husebo, Ane Erdal, and Keela Herr. 2020. Pain in dementia. *Pain reports*, 5(1).

[Aqajari *et al*.2021] Seyed Amir Hossein Aqajari, Rui Cao, Emad Kasaeyan Naeini, Michael-David Calderon, Kai Zheng, Nikil Dutt, Pasi Liljeberg, Sanna Salanterä, Ariana M Nelson, and Amir M Rahmani. 2021. Pain assessment tool with electrodermal activity for post-operative patients: method validation study. *JMIR mHealth and uHealth*, 9(5):e25258.

[Banganho *et al*.2021] Antonio Rodrigues Banganho, Marcelino Bicho dos Santos, and Hugo Placido da Silva. 2021. Design and evaluation of an electrodermal activity sensor (eda) with adaptive gain. *IEEE Sensors Journal*, 21(6):8639–8649.

[Bhatkar *et al*.2022] Viprali Bhatkar, Rosalind Picard, and Camilla Staahl. 2022. Combining electrodermal activity with the peak-pain time to quantify three temporal regions of pain experience. *Frontiers in Pain Research*, 3.

[Borgwardt *et al*.2006] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57.

[Bui *et al*.2021] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. 2021. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201.

[Cao *et al*.2020] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.

[Casti *et al*.2020] Paola Casti, Arianna Mencattini, Joanna Filippi, Michele D'Orazio, Maria Colomba Comes, Davide Di Giuseppe, and Eugenio Martinelli. 2020. A personalized assessment platform for non-invasive monitoring of pain. In *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–5. IEEE.

[Chae *et al.*2022] Younbyoung Chae, Hi-Joon Park, and In-Seon Lee. 2022. Pain modalities in the body and brain: current knowledge and future perspectives. *Neuroscience & Biobehavioral Reviews*, page 104744.

[Chang *et al.*2019] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. 2019. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 7354–7362.

[Chesler *et al.*2002] Elissa J Chesler, Sonya G Wilson, William R Lariviere, Sandra L Rodriguez-Zas, and Jeffrey S Mogil. 2002. Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. *Neuroscience & Biobehavioral Reviews*, 26(8):907–923.

[Christ *et al.*2018] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. 2018. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77.

[Chu *et al.*2017] Yaqi Chu, Xingang Zhao, Jianda Han, and Yang Su. 2017. Physiological signal-based method for measurement of pain intensity. *Frontiers in neuroscience*, 11:279.

[Dawson *et al.*2017] Michael E Dawson, Anne M Schell, and Diane L Filion. 2017. The electrodermal system.

[Fulcher *et al.*2013] Ben D Fulcher, Max A Little, and Nick S Jones. 2013. Highly comparative time-series analysis: the empirical structure of time series and their methods. *Journal of the Royal Society Interface*, 10(83):20130048.

[Ganin and Lempitsky2015] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

[Han *et al.*2020] Wenjing Han, Tao Jiang, Yan Li, Björn Schuller, and Huabin Ruan. 2020. Ordinal learning for emotion recognition in customer service calls. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6494–6498. IEEE.

[Hyun *et al.*2022] Jinshil Hyun, Jiyue Qin, Cuiling Wang, Mindy J Katz, Jelena M Pavlovic, Carol A Derby, and Richard B Lipton. 2022. Reliabilities of intra-individual mean and intra-individual variability of self-reported pain derived from ecological momentary assessments: Results from the einstein aging study. *The Journal of Pain*, 23(4):616–624.

[IMotions2017] IMotions. 2017. Galvanic skin response—the complete pocket guide.

[Iwasawa and Matsuo2021] Yusuke Iwasawa and Yutaka Matsuo. 2021. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural*

Information Processing Systems, 34:2427–2440.

[Ji *et al*.2023] Xinwei Ji, Tianming Zhao, Wei Li, and Albert Zomaya. 2023. Automatic pain assessment with ultra-short electrodermal activity signal. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 618–625.

[Jiang and Chen2021] Run-Qiang Jiang and Lan-Lan Chen. 2021. Driving stress estimation in physiological signals based on hierarchical clustering and multi-view intact space learning. *IEEE Transactions on Intelligent Transportation Systems*.

[Jiang *et al*.2017] Mingzhe Jiang, Riitta Mieronkoski, Amir M Rahmani, Nora Hagelberg, Sanna Salanterä, and Pasi Liljeberg. 2017. Ultra-short-term analysis of heart rate variability for real-time acute pain monitoring with wearable electronics. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1025–1032. IEEE.

[Jiang *et al*.2024] Mingzhe Jiang, Riitta Rosio, Sanna Salanterä, Amir M Rahmani, Pasi Liljeberg, Daniel S da Silva, Victor Hugo C de Albuquerque, and Wanqing Wu. 2024. Personalized and adaptive neural networks for pain detection from multi-modal physiological features. *Expert Systems with Applications*, 235:121082.

[Jonsdottir and Gunnarsson2021] Thorbjorg Jonsdottir and Esther Christina Gunnarsson. 2021. Understanding nurses' knowledge and attitudes toward pain assessment in dementia: a literature review. *Pain Management Nursing*, 22(3):281–292.

[Kächele *et al*.2015] Markus Kächele, Patrick Thiam, Mohammadreza Amirian, Philipp Werner, Steffen Walter, Friedhelm Schwenker, and Günther Palm. 2015. Multimodal data fusion for person-independent, continuous estimation of pain intensity. In *International Conference on Engineering Applications of Neural Networks*, pages 275–285. Springer.

[Kächele *et al*.2016] Markus Kächele, Patrick Thiam, Mohammadreza Amirian, Friedhelm Schwenker, and Günther Palm. 2016. Methods for person-centered continuous pain intensity assessment from bio-physiological channels. *IEEE Journal of Selected Topics in Signal Processing*, 10(5):854–864.

[Kächele *et al*.2017] Markus Kächele, Mohammadreza Amirian, Patrick Thiam, Philipp Werner, Steffen Walter, Günther Palm, and Friedhelm Schwenker. 2017. Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evolving Systems*, 8:71–83.

[Kamper-Fuhrmann *et al*.2023] Elisa Kamper-Fuhrmann, Alexander Winkler, Alannah Hahn, and Christiane Hermann. 2023. The hand-withdrawal-method-an adapted and simplified method of limits for behavioral heat pain assessment. *The Journal of Pain*, 24(5):888–900.

[Kappesser and Williams2010] Judith Kappesser and Amanda C de C Williams. 2010. Pain estimation: asking the right questions. *Pain*, 148(2):184–187.

[Kearns1989] M. J. Kearns. 1989. *Computational Complexity of Machine Learning*. Ph.D. thesis, Department of Computer Science, Harvard University.

[Kendall and Gal2017] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

[Kendall *et al*.2018] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

[Kong *et al*.2021] Youngsun Kong, Hugo F Posada-Quintero, and Ki H Chon. 2021. Sensitive physiological indices of pain based on differential characteristics of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, 68(10):3122–3130.

[Krueger *et al*.2021] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR.

[Landis and Koch1977] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

[Leiner *et al*.2012] Dominik Leiner, Andreas Fahr, and Hannah Früh. 2012. Eda positive change: A simple algorithm for electrodermal activity to measure general audience arousal during media exposure. *Communication Methods and Measures*, 6(4):237–250.

[Li and Lin2006] Ling Li and Hsuan-Tien Lin. 2006. Ordinal regression by extended binary classification. *Advances in neural information processing systems*, 19.

[Lim *et al*.2023] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. 2023. Ttn: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*.

[Liu *et al*.2017] Dianbo Liu, Peng Fengjiao, Rosalind Picard, et al. 2017. Deepfacelift: interpretable personalized models for automatic estimation of self-reported pain. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, pages 1–16. PMLR.

[Lopez-Martinez and Picard2017] Daniel Lopez-Martinez and Rosalind Picard. 2017. Multi-task neural networks for personalized pain recognition from physiological signals. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 181–184. IEEE.

[Lopez-Martinez and Picard2018] Daniel Lopez-Martinez and Rosalind Picard. 2018. Continuous pain intensity estimation from autonomic signals with recurrent neural networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5624–5627. IEEE.

[Lopez-Martinez *et al.*2017] Daniel Lopez-Martinez, Ognjen Rudovic, and Rosalind Picard. 2017. Physiological and behavioral profiling for nociceptive pain estimation using personalized multitask learning. *arXiv preprint arXiv:1711.04036*.

[Lötsch and Ultsch2018] Jörn Lötsch and Alfred Ultsch. 2018. Machine learning in pain research. *Pain*, 159(4):623.

[Lötsch *et al.*2017] J Lötsch, G Geisslinger, S Heinemann, F Lerch, BG Oertel, and A Ultsch. 2017. Qst response patterns to capsaicin-and uv-b-induced local skin hypersensitization in healthy subjects: a machine-learned analysis. *Pain*, 159:11–24.

[Lubba *et al.*2019] Carl H Lubba, Sarab S Sethi, Philip Knaute, Simon R Schultz, Ben D Fulcher, and Nick S Jones. 2019. catch22: Canonical time-series characteristics. *Data Mining and Knowledge Discovery*, 33(6):1821–1852.

[Lundberg *et al.*2022] Adrian Lundberg, Nicola Fraschini, and Renata Aliani. 2022. What is subjectivity? scholarly perspectives on the elephant in the room. *Quality & Quantity*, pages 1–21.

[Madden *et al.*2021] Victoria J Madden, Peter R Kamerman, Mark J Catley, Valeria Bellan, Leslie N Russek, Danny Camfferman, and G Lorimer Moseley. 2021. Variability in experimental pain studies: nuisance or opportunity? *British journal of anaesthesia*, 126(2):e61–e64.

[Mosley and Butler2017] G Lorimer Mosley and David S Butler. 2017. *Explain pain supercharged*. NOI.

[Munsters *et al.*2012] Josanne Munsters, Linda Wallström, Johan Ågren, Torgny Norsted, and Richard Sindelar. 2012. Skin conductance measurements as pain assessment in newborn infants born at 22–27 weeks gestational age at different postnatal age. *Early human development*, 88(1):21–26.

[Ollander *et al.*2016] Simon Ollander, Christelle Godin, Aurélie Campagne, and Sylvie Charbonnier. 2016. A comparison of wearable and stationary sensors for stress detection. In *2016 IEEE International Conference on systems, man, and Cybernetics (SMC)*, pages 004362–004366. IEEE.

[Patterson *et al.*2015] Olga V Patterson, Makoto Jones, Yiwen Yao, Benjamin Viernes, Patrick R Alba, Theodore J Iwashyna, and Scott L DuVall. 2015. Extraction of vital signs from clinical notes. *Studies in health technology and informatics*, 216:1035–1035.

[Posada-Quintero and Chon2020] Hugo F Posada-Quintero and Ki H Chon. 2020. Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors*, 20(2):479.

[Posada-Quintero *et al.*2021] Hugo F Posada-Quintero, Youngsun Kong, and Ki H Chon. 2021. Objective pain stimulation intensity and pain sensation assessment using machine

learning classification and regression based on electrodermal activity. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 321(2):R186–R196.

[Pouromran *et al.*2021] Fatemeh Pouromran, Srinivasan Radhakrishnan, and Sagar Kamarthi. 2021. Exploration of physiological sensors, features, and machine learning models for pain intensity estimation. *Plos one*, 16(7):e0254108.

[Rajasekhar *et al.*2021] Gnana Praveen Rajasekhar, Eric Granger, and Patrick Cardinal. 2021. Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos. *Image and Vision Computing*, 110:104167.

[Rosenfeld *et al.*2022] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. 2022. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*.

[Sagawa *et al.*2019] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

[Schneider *et al.*2020] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. 2020. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551.

[Segu *et al.*2023] Mattia Segu, Alessio Tonioni, and Federico Tombari. 2023. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, 135:109115.

[Shi *et al.*2021] Xintong Shi, Wenzhi Cao, and Sebastian Raschka. 2021. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *arXiv preprint arXiv:2111.08851*.

[Shukla *et al.*2019] Jainendra Shukla, Miguel Barreda-Angeles, Joan Oliver, Gora Chand Nandi, and Domenec Puig. 2019. Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing*, 12(4):857–869.

[Spisak *et al.*2020] Tamas Spisak, Balint Kincses, Frederik Schlitt, Matthias Zunhammer, Tobias Schmidt-Wilcke, Zsigmond T Kincses, and Ulrike Bingel. 2020. Pain-free resting-state functional brain connectivity predicts individual pain sensitivity. *Nature communications*, 11(1):187.

[Sugimine *et al.*2020] Satomi Sugimine, Shigeru Saito, and Tomonori Takazawa. 2020. Normalized skin conductance level could differentiate physical pain stimuli from other sympathetic stimuli. *Scientific reports*, 10(1):1–12.

[Sun and Saenko2016] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*,

pages 443–450. Springer.

[Sun *et al*.2020] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR.

[Susam *et al*.2021] Busra T Susam, Nathan T Riek, Murat Akcakaya, Xiaojing Xu, Virginia R de Sa, Hooman Nezamfar, Damaris Diaz, Kenneth D Craig, Matthew S Goodwin, and Jeannie S Huang. 2021. Automated pain assessment in children using electrodermal activity and video data fusion via machine learning. *IEEE Transactions on Biomedical Engineering*, 69(1):422–431.

[Szczapa *et al*.2022] Benjamin Szczapa, Mohamed Daoudi, Stefano Berretti, Pietro Pala, Alberto Del Bimbo, and Zakia Hammal. 2022. Automatic estimation of self-reported pain by trajectory analysis in the manifold of fixed rank positive semi-definite matrices. *IEEE transactions on affective computing*, 13(4):1813–1826.

[Thiam *et al*.2019] Patrick Thiam, Peter Bellmann, Hans A Kestler, and Friedhelm Schwenker. 2019. Exploring deep physiological models for nociceptive pain recognition. *Sensors*, 19(20):4503.

[Thiam *et al*.2021] Patrick Thiam, Heinke Hihn, Daniel A Braun, Hans A Kestler, and Friedhelm Schwenker. 2021. Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective. *Frontiers in Physiology*, 12.

[Thought Technology2023] Thought Technology. 2023. evu tps package - t4500. https://thoughttechnology.com/evu-tps-package-t4500/. Accessed: 2023-08-07.

[Tracey *et al*.2019] Irene Tracey, Clifford J Woolf, and Nick A Andrews. 2019. Composite pain biomarker signatures for objective assessment and effective treatment. *Neuron*, 101(5):783–800.

[Treister *et al*.2012] Roi Treister, Mark Kliger, Galit Zuckerman, Itay Goor Aryeh, and Elon Eisenberg. 2012. Differentiating between heat pain intensities: the combined effect of multiple autonomic parameters. *PAIN®*, 153(9):1807–1814.

[Tzeng *et al*.2014] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

[Uddin *et al*.2023] Md Taufeeq Uddin, Ghada Zamzmi, and Shaun Canavan. 2023. Cooperative learning for personalized context-aware pain assessment from wearable data. *IEEE Journal of Biomedical and Health Informatics*.

[unk2021] 2021. eVu TPS Package - T4500.

[Van Der Miesen *et al.*2019] Maite M Van Der Miesen, Martin A Lindquist, and Tor D Wager. 2019. Neuroimaging-based biomarkers for pain: state of the field and current directions. *Pain reports*, 4(4).

[Vijayakumar *et al.*2017] Vishal Vijayakumar, Michelle Case, Sina Shirinpour, and Bin He. 2017. Quantifying and characterizing tonic thermal pain across subjects from eeg data using random forest models. *IEEE Transactions on Biomedical Engineering*, 64(12):2988–2996.

[von Hehn *et al.*2012] Christian A von Hehn, Ralf Baron, and Clifford J Woolf. 2012. Deconstructing the neuropathic pain phenotype to reveal neural mechanisms. *Neuron*, 73(4):638–652.

[Walter *et al.*2013] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. 2013. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *2013 IEEE international conference on cybernetics (CYBCO)*, pages 128–131. IEEE.

[Wang *et al.*2020a] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020a. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

[Wang *et al.*2020b] Rui Wang, Shaoshuang Wang, Na Duan, and Qiang Wang. 2020b. From patient-controlled analgesia to artificial intelligence-assisted patient-controlled analgesia: practices and perspectives. *Frontiers in Medicine*, 7:145.

[Wang *et al.*2021] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*.

[Werner *et al.*2014] Philipp Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C Traue. 2014. Automatic pain recognition from video and biomedical signals. In *2014 22nd international conference on pattern recognition*, pages 4582–4587. IEEE.

[Werner *et al.*2019] Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind Picard. 2019. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*.

[Williamson and Hoggart2005] Amelia Williamson and Barbara Hoggart. 2005. Pain: a review of three commonly used pain rating scales. *Journal of clinical nursing*, 14(7):798–804.

[Winslow *et al.*2022] Brent D Winslow, Rebecca Kwasinski, Kyle Whirlow, Emily Mills, Jeffrey Hullfish, and Meredith Carroll. 2022. Automatic detection of pain using machine learning. *Frontiers in Pain Research*, 3:1044518.

[Xinwei Ji2023] Xinwei Ji. 2023. Supplementary material. https://www.researchgate.net/publication/376617915_Supplementary_material_for_Unraveling_Pain_Levels_A_Data-Uncertainty_Guided_Approach_for_Effective_Pain_Assessment_Appendix_A_Feature_Extraction. Accessed: 2023-12-15.

[Xu and de Sa2021] Xiaojing Xu and Virginia R de Sa. 2021. Personalized pain detection in facial video with uncertainty estimation. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4163–4168. IEEE.

[Yang *et al.*2019] Fan Yang, Tanvi Banerjee, Mark J Panaggio, Daniel M Abrams, and Nirmish R Shah. 2019. Continuous pain assessment using ensemble feature selection from wearable sensor data. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 569–576. IEEE.

[Yang *et al.*2021] Ruijing Yang, Ziyu Guan, Zitong Yu, Xiaoyi Feng, Jinye Peng, and Guoying Zhao. 2021. Non-contact pain recognition from video sequences with remote physiological measurements prediction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), 19-27 August, Monteral, Canada*. International Joint Conferences on Artificial Intelligence.

[Zaman *et al.*2021] Jonas Zaman, Lukas Van Oudenhove, and Johan WS Vlaeyen. 2021. Uncertainty in a context of pain: disliked but also more painful? *Pain*, 162(4):995–998.

[Zhang *et al.*2017] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

[Zhang *et al.*2021] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. 2021. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv e-prints*, pages arXiv–2107.

[Zhang *et al.*2022] Marvin Zhang, Sergey Levine, and Chelsea Finn. 2022. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35:38629–38642.

[Zhang *et al.*2023] Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. 2023. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International Conference on Machine Learning*, pages 41647–41676. PMLR.