Unraveling Pain Levels: A Data-Uncertainty Guided Approach for Effective Pain Assessment

Xinwei Ji[1], **Xiaomin Chang**[1], **Wei Li**[1], **Albert Y. Zomaya**[1]

[1]The University of Sydney

xiji3593@uni.sydney.edu.au, xcha8737@uni.sydney.edu.au, weiwilson.li@sydney.edu.au, albert.zomaya@sydney.edu.au

Abstract

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%.

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, Posada-Quintero, and Chon 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. 2020), objective and continuous pain state evaluation can be achieved through algorithmic approaches. However, due to the complexity of pain as a phenomenon (von Hehn, Baron, and Woolf 2012) and the lack of identified biomarkers reflecting its fundamental mechanisms (Van Der Miesen, Lindquist, and Wager 2019), the search for objective biomarkers to elucidate pain mechanisms remains a pressing need (Tracey, Woolf, and Andrews 2019).

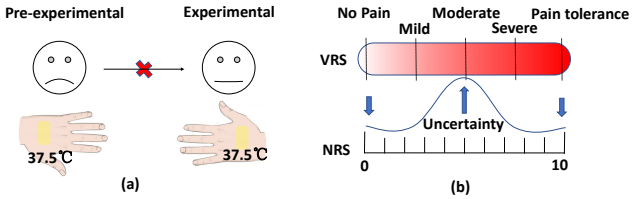

Figure 1: Examples of uncertainty in pain stimulation experiments include: (a) the inconsistency in stimulation-reaction, where pain perception varies between the pre-experiment and the formal experiment, and changes over time; (b) the uncertainty in subjective pain assessment, which increases as the rating approaches the middle of the pain scale.

The AI-PCA system utilizes facial expressions and biosignals 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 ultra-short 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.

Our main contributions can be summarized as follows:

- We propose DUG-CORAL, a novel method for automatic pain assessment that integrates pain data uncertainty and the ordinal regression algorithm (CORAL). We are the first to investigate the possibility of using heart-related indicators and their derived heart rate variability (HRV) features to estimate data uncertainty for individuals, and then incorporating this uncertainty into the training of an electrodermal activity (EDA)-based pain prediction model.
- We use a task-importance weighting strategy in the loss calculation process to alleviate the ambiguity problem specific to pain labeling.
- Extensive experiments on pain datasets, BioVid (public) and Apon (private), demonstrate the superior performance of our method compared to commonly used approaches. We also test our method on pain patients to further establish its effectiveness.

Related Work

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, Park, and Lee 2022). Differential characteristics of EDA have been identified as sensitive indices for classifying experimental pain stimulation induced by electric pulse (Kong, Posada-Quintero, and Chon 2021), heating (Pouromran, Radhakrishnan, and Kamarthi 2021), and cold pressor test (Winslow et al. 2022). Furthermore, normalized skin conductance level can distinguish physical pain stimuli from other sympathetic stimuli (Sugimine, Saito, and Takazawa 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.

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, Fraschini, and Aliani 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, Van Oudenhove, and Vlaeyen 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.

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 $f : \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes

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

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 (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:

$$f(X_{0:t}) = f(X_{t-\Delta t:t}) \tag{2}$$

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, Posada-Quintero, and Chon 2021).

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 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.

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 1: 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:

$$C = cov(F_{std}) \tag{3}$$

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 $pc1$.

$$PC1_{scores} = F_{std} \times v_{PC1} \tag{4}$$

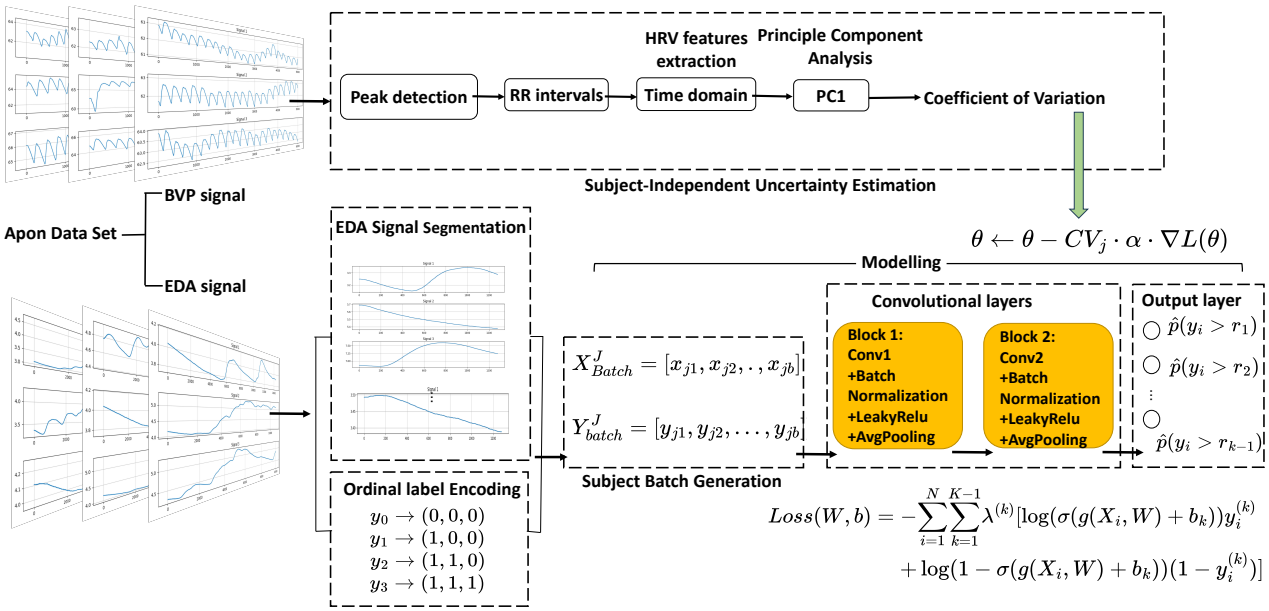

Figure 2: Framework of DUG-CORAL ranker on Apon database

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:

$$CV = \frac{STD_{PC1_{score}}}{MEAN_{PC1_{score}}} \tag{5}$$

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

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, Mirjalili, and Raschka 2020) for classification Loss .

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, Mirjalili, and Raschka 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 2: 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 $CV_{s_1,s_2,\ldots,s_n}$
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:

$$\hat{P}(y_i^{(k)} = 1) = \sigma(g(x_i, W) + b_k). \tag{6}$$

For model training, we minimize the loss function:

$$L(W,b) = -\sum_{i=1}^{N}\sum_{k=1}^{K-1}\lambda^{(k)}[\log(\sigma(g(X_i,W)+b_k))y_i^{(k)} + \log(1-\sigma(g(X_i,W)+b_k))(1-y_i^{(k)})] \tag{7}$$

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.

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:

$$B_j = \{(X_i, y_i) \mid S_i = s_j\} \tag{8}$$

where s_j is the subject of jth batch

The operation of Batch Normalization can be represented as:

$$\mathrm{BN}(X_i) = \frac{X_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \odot \gamma + \beta \tag{9}$$

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.

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.

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 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.

$$CV_{scale} = \left(\frac{\tanh(CV) + 1}{2}\right) \cdot (W_{max} - W_{min}) + W_{min} \tag{10}$$

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.

Experiments

Datasets and Experimental Settings

Apon 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.

BioVid The BioVid heat pain dataset, containing biosignals (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.

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, Radhakrishnan, and Kamarthi 2021)	EDA	0.93	1.16
Domain Adaptation (Rajasekhar, Granger, and Cardinal 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 1: Comparison with the existing methods on BioVid database

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

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.

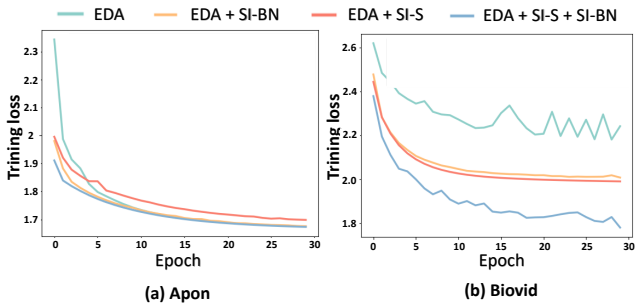

Figure 3: Ablation study and training loss. EDA refers to training process of DUG-CORAL without SI-standardization (SI-S) and SI-Batch normalization (SI-BN)

Results

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 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.

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 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).

	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 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)

Ablation Study

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 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).

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 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.

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, Gal, and Cipolla 2018). 4) Scheme IV: In the fourth scheme, hard weighting scheme for BioVid is applied as $w_{T_0vsT_1} = 0.4$, $w_{T_1vsT_2} = 0.1$, $w_{T_2vsT_3} = 0.1$, $w_{T_3vsT_4} = 0.4$. For Apon, $w_{T_0vsT_1} = 0.4$, $w_{T_1vsT_2} = 0.2$, $w_{T_2vsT_3} = 0.4$. 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 $w_{T_0vsT_1} = 0.1$, $w_{T_1vsT_2} = 0.4$, $w_{T_2vsT_3} = 0.4$, $w_{T_3vsT_4} = 0.1$. For Apon, $w_{T_0vsT_1} = 0.3$, $w_{T_1vsT_2} = 0.4$, $w_{T_2vsT_3} = 0.3$. 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 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 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.

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.

Acknowledgements

Dr. Wei Li acknowledges the support of the Australian Research Council (ARC) through the Discovery Early Career Researcher Award (DE210100263). Professor Zomaya and Dr. Wei Li acknowledge the support of an ARC Discovery Project (DP200103494) and the support from Australia-China Centre for Energy Informatics and Demand Response Technologies through Department of Industry, Innovation and Science, Australia (ACSRIII000004).

We extend our gratitude to Jiangsu Apon Medical Technology Co.,Ltd., particularly to Hongyi Zhu, Vice-president of the company, for their invaluable collaboration and support in this research. Their expertise and insights have significantly contributed to the development and success of our study.

References

Aqajari, S. A. H.; Cao, R.; Kasaeyan Naeini, E.; Calderon, M.-D.; Zheng, K.; Dutt, N.; Liljeberg, P.; Salanterä, S.; Nelson, A. M.; and Rahmani, A. M. 2021. Pain assessment tool with electrodermal activity for postoperative patients: method validation study. *JMIR mHealth and uHealth*, 9(5): e25258.

Cao, W.; Mirjalili, V.; and Raschka, S. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140: 325–331.

Chae, Y.; Park, H.-J.; and Lee, I.-S. 2022. Pain modalities in the body and brain: current knowledge and future perspectives. *Neuroscience & Biobehavioral Reviews*, 104744.

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

Jiang, M.; Mieronkoski, R.; Rahmani, A. M.; Hagelberg, N.; Salanterä, S.; and Liljeberg, P. 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)*, 1025–1032. IEEE.

Kächele, M.; Amirian, M.; Thiam, P.; Werner, P.; Walter, S.; Palm, G.; and Schwenker, F. 2017. Adaptive confidence learning for the personalization of pain intensity estimation systems. *Evolving Systems*, 8: 71–83.

Kamper-Fuhrmann, E.; Winkler, A.; Hahn, A.; and Hermann, C. 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.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Kendall, A.; Gal, Y.; and Cipolla, R. 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*, 7482–7491.

Kong, Y.; Posada-Quintero, H. F.; and Chon, K. H. 2021. Sensitive physiological indices of pain based on differential characteristics of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, 68(10): 3122–3130.

Lundberg, A.; Fraschini, N.; and Aliani, R. 2022. What is subjectivity? Scholarly perspectives on the elephant in the room. *Quality & Quantity*, 1–21.

Madden, V. J.; Kamerman, P. R.; Catley, M. J.; Bellan, V.; Russek, L. N.; Camfferman, D.; and Moseley, G. L. 2021. Variability in experimental pain studies: nuisance or opportunity? *British journal of anaesthesia*, 126(2): e61–e64.

Mosley, G. L.; and Butler, D. S. 2017. *Explain pain supercharged*. NOI.

Pouromran, F.; Radhakrishnan, S.; and Kamarthi, S. 2021. Exploration of physiological sensors, features, and machine learning models for pain intensity estimation. *Plos one*, 16(7): e0254108.

Rajasekhar, G. P.; Granger, E.; and Cardinal, P. 2021. Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos. *Image and Vision Computing*, 110: 104167.

Sugimine, S.; Saito, S.; and Takazawa, T. 2020. Normalized skin conductance level could differentiate physical pain stimuli from other sympathetic stimuli. *Scientific reports*, 10(1): 10950.

Susam, B. T.; Riek, N. T.; Akcakaya, M.; Xu, X.; de Sa, V. R.; Nezamfar, H.; Diaz, D.; Craig, K. D.; Goodwin, M. S.; and Huang, J. S. 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, B.; Daoudi, M.; Berretti, S.; Pala, P.; Del Bimbo, A.; and Hammal, Z. 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, P.; Hihn, H.; Braun, D. A.; Kestler, H. A.; and Schwenker, F. 2021. Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective. *Frontiers in Physiology*, 12: 720464.

Thought Technology. 2023. eVu TPS Package - T4500. https://thoughttechnology.com/evu-tps-package-t4500/. Accessed: 2023-08-07.

Tracey, I.; Woolf, C. J.; and Andrews, N. A. 2019. Composite pain biomarker signatures for objective assessment and effective treatment. *Neuron*, 101(5): 783–800.

Van Der Miesen, M. M.; Lindquist, M. A.; and Wager, T. D. 2019. Neuroimaging-based biomarkers for pain: state of the field and current directions. *Pain reports*, 4(4).

von Hehn, C. A.; Baron, R.; and Woolf, C. J. 2012. Deconstructing the neuropathic pain phenotype to reveal neural mechanisms. *Neuron*, 73(4): 638–652.

Walter, S.; Gruss, S.; Ehleiter, H.; Tan, J.; Traue, H. C.; Werner, P.; Al-Hamadi, A.; Crawcour, S.; Andrade, A. O.; and da Silva, G. M. 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)*, 128–131. IEEE.

Wang, R.; Wang, S.; Duan, N.; and Wang, Q. 2020. From patient-controlled analgesia to artificial intelligence-assisted patient-controlled analgesia: Practices and perspectives. *Frontiers in Medicine*, 7: 145.

Williamson, A.; and Hoggart, B. 2005. Pain: a review of three commonly used pain rating scales. *Journal of clinical nursing*, 14(7): 798–804.

Winslow, B. D.; Kwasinski, R.; Whirlow, K.; Mills, E.; Hullfish, J.; and Carroll, M. 2022. Automatic detection of pain using machine learning. *Frontiers in Pain Research*, 3: 1044518.

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, X.; and de Sa, V. R. 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)*, 4163–4168. IEEE.

Zaman, J.; Van Oudenhove, L.; and Vlaeyen, J. W. 2021. Uncertainty in a context of pain: disliked but also more painful? *Pain*, 162(4): 995–998.