Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Heterogeneous Graph-Based Facial Pain Modeling with Ordinal Boundary Analysis and Cross-Dataset Transfer

Muhamad Zulhelmi Mohd Nizam[1,♣], MaziahMohamad[1], Mohamed Hussein[1], Shaharil Mad Saad[1], Muhamad Zulhairi Mohd Nizam[1]

[1] Faculty of Mechanical Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

ARTICLE INFO

Article history:
Received 3 September 2026
Received in revised form 15 September 2026
Accepted 17 September 2026
Available online 21 September 2026

Keywords:
Facial pain assessment; Pain label granularity; Cross-dataset transfer; Heterogeneous graph neural networks; Ordinal pain modelling

ABSTRACT

Automatic facial pain assessment remains challenging due to the subtle and continuous nature of facial expressions and the substantial variability between experimental and clinical datasets. While recent deep learning approaches report strong within-dataset performance using fine-grained pain labels, it remains unclear whether such representations generalize across domains. In this work, we present a diagnostic study of facial pain representations using a structured temporal graph model based on a Heterogeneous Graph Transformer (HGT) with temporal convolution and Long Short Term Memory (LSTM) layers. Pain-relevant facial Action Units and head pose features are explicitly modelled as heterogeneous graph nodes with clinically motivated connections. Fine-grained multiclass pain modelling (levels 0–4) is first evaluated on the BioVid Heat Pain dataset and compared with state-of-the-art methods. Although competitive performance is achieved, prediction errors occur predominantly between neighbouring pain levels, while boundary-wise AUC analysis shows greater separability at higher pain intensities. Based on this analysis, three binary pain-label formulations are evaluated for cross-dataset transfer to UNBC-McMaster. The results show that pain-label grouping influences cross-dataset discrimination. The Extreme formulation achieves the highest discrimination but excludes intermediate BioVid pain levels, while the Coarse formulation retains the complete severity range with improved discrimination over the Presence formulation. Correlation analysis further shows that stronger binary discrimination does not necessarily correspond to a stronger association with the original clinical pain ratings. Overall, the findings highlight the trade-off between pain-level separability, severity coverage, and cross-dataset discrimination.

1. Introduction

Pain is a subjective experience characterized by unpleasant sensory and emotional responses linked to actual or potential tissue damage. It is commonly classified into acute and chronic pain, with the primary distinction being duration: acute pain typically resolves within three months, while chronic pain persists beyond this period and is often associated with conditions such as cancer, chronic back pain, or headaches [1].

♣ *Corresponding author*
E-mail address: mzulhelmi24@graduate.utm.my

In clinical practice, pain assessment relies on self-reporting and observational tools. Self-reporting instruments such as the Visual Analog Scale (VAS), Numerical Rating Scale (NRS), and Wong–Baker FACES scale require adequate communication and cognitive function, which may not be feasible in all patient groups. Observational tools, including the Critical-Care Pain Observation Tool (CPOT), Pain Assessment in Advanced Dementia Scale (PAINAD), and Observer Pain Rating (OPR), are time-intensive and susceptible to inter-rater variability. Prior work reported only moderate inter-rater agreement (Cohen's $\kappa = 0.61$), highlighting their limitations [2].

To address these challenges, researchers have explored automated pain recognition using input modalities such as facial expressions, body posture, vocal features, and physiological signals [3], [4], [5]. Physiological signals such as electrodermal activity (EDA) can be reliable, as shown by Othman *et al.,* [6] but they require contact-based sensors that are intrusive in many clinical scenarios [7]. By contrast, facial image analysis provides a non-contact, camera-based solution that is practical and scalable. As noted by Gkikas and Tsiknakis [8], facial features remain one of the most feasible modalities for engineering and deployment.

The effectiveness of facial pain assessment systems depends not only on model design but also on the reliability and interpretation of pain annotations. Commonly used ground-truth measures include observer-rated metrics such as the Prkachin and Solomon Pain Intensity (PSPI) score, self-reported Visual Analog Scales, and stimulus-based intensity levels as used in the BioVid dataset. Despite its widespread adoption, PSPI has been criticized for not always aligning with subjective pain experience, for example in cases where facial expressions are suppressed or absent [1], [8].

Nevertheless, several studies have shown that PSPI can reliably quantify pain-related facial activity. [9] reported moderate to high intra-class correlation between manual and automatic PSPI estimation, while more recent deep learning approaches have successfully used PSPI for ordinal or regression-based pain modeling [10], [11], [12].

Over the years, a wide range of approaches have been proposed for facial pain assessment, including handcrafted feature-based methods [13], [14], CNN-based spatial modeling, recurrent architectures for temporal dynamics [15], [16] and more recent transformer-based frameworks [17], [18], [19], [20]. These methods consistently report strong performance when trained and evaluated within the same dataset.

Despite these advances, most studies evaluate models under within-dataset settings, and external validation across datasets remains limited. Several works suggest that increasing data diversity, for example by combining multiple datasets, can improve model performance and generalization [21], [22]. However, empirical cross-dataset evaluations consistently reveal substantial performance degradation when models are applied to unseen datasets, indicating the presence of significant domain shift [22], [23], [24]. This challenge has been explicitly noted when attempting to transfer models across differing pain environments [10]

A prominent example of this domain shift arises between the UNBC-McMaster dataset, which captures spontaneous clinical pain in patients undergoing physiotherapy [5], and the BioVid Heat Pain database, which records experimentally induced thermal pain in healthy subjects [25]. Differences in pain elicitation, subject health status, and recording protocols lead to mismatched data distributions and semantic inconsistencies that hinder direct model transferability [22]. Although physiological signals are often considered less sensitive to individual variability [4], [6], recent studies indicate that domain shift can also occur between subjects within the same dataset, particularly for physiological modalities [26], further highlighting the pervasiveness of distribution mismatch.

Recent studies have continued to address the difficulty of distinguishing subtle variations across facial pain intensities through improved spatiotemporal feature learning, attention mechanisms, and strategies for class imbalance. Liu *et al.,* [27], for example, consolidated the original UNBC PSPI

intensity scale into six categories to mitigate its highly imbalanced intensity distribution, while Gao *et al.,* [28] addressed imbalance through synthetic augmentation of underrepresented pain intensities. Lu *et al.,* [29] further investigated the influence of class imbalance on video-based pain intensity estimation through different loss function and optimization strategies. However, these approaches primarily address feature representation and class imbalance. Comparatively limited attention has been given to examining where reliable separation occurs between adjacent ordinal pain levels and whether pain-label granularity itself affects generalization across datasets with different pain elicitation conditions. Consequently, it remains unclear whether fine-grained pain levels that are difficult to distinguish within an experimental dataset provide an appropriate representation for transfer to a clinically different pain dataset.

Motivated by these observations, this study conducts a diagnostic cross-dataset analysis of facial pain representation, with particular emphasis on the relationship between ordinal pain-level separability, label granularity, and cross-dataset transfer. A structured temporal–graph model based on a HGT with temporal convolution and LSTM layers is employed to represent pain-related facial Action Units (AUs), head pose, and their temporal relationships. Fine-grained pain modeling (levels 0–4) is first performed on the BioVid dataset to examine ordinal confusion and boundary separability. The analysis reveals substantial overlap between adjacent pain levels, particularly at lower intensities, despite preservation of the overall severity ordering.

Based on this diagnostic analysis, a coarser binary formulation grouping levels 0–2 and 3–4 is subsequently examined as an alternative representation for cross-dataset transfer. The binary model is then evaluated on the clinically distinct UNBC-McMaster dataset to determine how pain representations learned from controlled experimental data transfer to clinical pain expressions. Accordingly, the primary contribution of this work lies in linking ordinal separability analysis with pain-label reformulation and cross-dataset evaluation, rather than treating cross-domain degradation solely as a feature-representation problem. The specific contributions are summarized as follows:

- **Diagnostic cross-dataset analysis**: The primary contribution is a diagnostic framework that first examines fine-grained ordinal confusion and boundary separability on BioVid, uses these observations to investigate a coarse pain representation, and subsequently evaluates its transfer to the clinically distinct UNBC-McMaster dataset. This enables cross-dataset generalization to be examined from the perspective of pain-label granularity in addition to feature-domain mismatch.
- **Heterogeneous temporal graph representation**: A heterogeneous temporal graph model is employed to represent pain-related facial AUs and head pose as distinct node types and to model their spatial and temporal relationships. Combined with temporal sequence modeling, this representation provides the technical basis for investigating fine-grained and coarser pain representations within a consistent modeling framework.
- **Analysis of pain-label granularity and transferability**: The study systematically examines ordinal boundary separability and compares alternative pain formulations before cross-dataset evaluation. The resulting analysis provides evidence on how reducing pain-label granularity affects the transfer of pain representations from experimentally induced BioVid pain to clinical UNBC pain.

2. Methodology

2.1 Problem Formulation and Heterogeneous Graph Construction

Facial pain is expressed through coordinated activations of multiple facial muscles and head movements, motivating a structured representation that explicitly models relationships between facial components [30]. To capture this structure, each facial frame is represented as a heterogeneous graph, where nodes correspond to pain-relevant facial Action Units (AUs) and head pose features.

The selection of AUs is guided by the Prkachin and Solomon Pain Intensity (PSPI) definition, ensuring that each AU node corresponds to a clinically meaningful facial component associated with pain expression. Specifically, PSPI is defined as:

$$PSPI = AU4 + max(AU6, AU7) + max(AU9, AU10) + AU45 \quad (1)$$

Rather than modeling PSPI as a single scalar node, each constituent AU is treated as an independent node. This design preserves localized muscle information and allows the model to learn fine-grained interactions among facial components. In addition to AU nodes, pose nodes are included to encode global head motion information, enabling interactions between local muscle activity and global facial dynamics. This is particularly important because PSPI does not account for head movement information [1].

Formally, each facial frame is represented as heterogeneous graph $\mathrm{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{R})$, where $\mathcal{V}$ denotes the set of nodes and $\mathcal{E}$ denotes the set of edges. Each node $v \in \mathcal{V}$ is associated with a node type mapping function $\tau(v): \mathcal{V} \rightarrow \mathcal{A}$, and each edge $e \in \mathcal{E}$ is associated with an edge type mapping function $\phi(e): \mathcal{E} \rightarrow \mathcal{R}$. The sets $\mathcal{A}$ and $\mathcal{R}$ represents the node type space and relation type space, respectively.

For an edge $e = (s, t)$ connecting a source node s to a target node t, a meta-relation is defined as

$$< \tau(s), \phi(e), \tau(t) > \quad (2)$$

which explicitly encodes the semantic interaction between the source node type, the relation type, and the target node type. This formulation enables relation-specific message passing and attention mechanisms within the heterogeneous graph.

In this work, the node type set is defined as $\mathcal{A} = \{AU, Pose\}$, corresponding to AU nodes and head pose nodes. Four relation types are defined to model intra- and inter-component interactions. Accordingly, the meta-relations used in the proposed model are:

$$\langle \mathrm{AU,to,AU} \rangle_{\mathrm{intra}},$$

$$\langle \mathrm{Pose,to,Pose} \rangle_{\mathrm{intra}},$$

$$\langle \mathrm{AU,to,Pose} \rangle_{\mathrm{inter}},$$

$$\langle \mathrm{Pose,to,AU} \rangle_{\mathrm{inter}} \quad (3)$$

where the subscripts *intra* and *inter* denote interactions within the same the same facial component type and between different component types, respectively. This formulation enables relation-aware message passing between facial muscle activations and head pose features.

After defining the structural components of the heterogeneous graph, the node feature representation associated with each frame is specified. The node feature matrix at time t is defined as:

$$X(t) = \{x_v(t) \mid v \in V\} \tag{4}$$

where $x_v(t)$ represents the filtered and normalized OpenFace feature corresponding to node v. AU nodes encode their respective intensity values, while pose nodes encode three-dimensional translation and rotation parameters. These features serve as the initial node embeddings upon which relation-aware message passing is performed. The overall spatio-temporal heterogeneous graph structure across consecutive frames is illustrated in Figure 1.

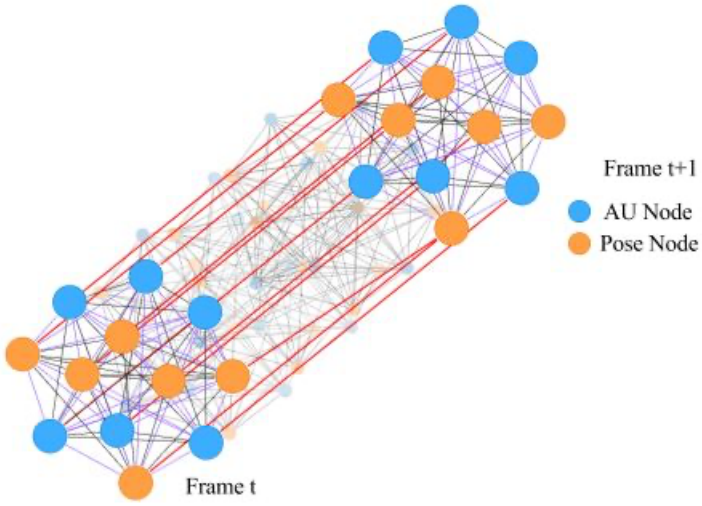


Fig. 1. Heterogeneous graph representation across consecutive frames with intra-, inter-, and temporal connections

2.2 Spatio-Temporal Modeling via HGT and BiLSTM

The overall architecture of the proposed HGT–BiLSTM framework is illustrated in Figure 2. Two layers of Heterogeneous Graph Transformer (HGT) are applied to perform type-specific, relation-aware message passing across nodes. This enables the model to learn structured dependencies among heterogeneous facial components through attention-based aggregation.

After graph propagation, node embeddings are aggregated using a learnable node-attention pooling mechanism. Instead of averaging node features, attention weights are computed separately over AU nodes and pose nodes at each time step, allowing the model to assign higher importance to facial components that contribute more strongly to pain expression. The resulting weighted AU and pose representations are concatenated and linearly projected to produce a compact frame-level embedding.

To capture temporal dynamics, the sequence of frame-level embeddings is fed into a bidirectional Long Short-Term Memory (BiLSTM) network. Let $H = [h_1, h_2, \ldots, h_T]$ denote the sequence of frame embeddings. The BiLSTM models long-range temporal dependencies across frames.

To obtain a fixed-length sequence representation, masked mean pooling is applied over valid time steps. For BioVid, where sequences have fixed length (138 frames), this operation reduces to standard mean pooling. Masked pooling is retained to support variable-length sequences during

cross-dataset evaluation on UNBC-McMaster. The final sequence-level representation is passed to either a regression head (for ordinal modelling) or a binary classification head, depending on the experimental formulation.

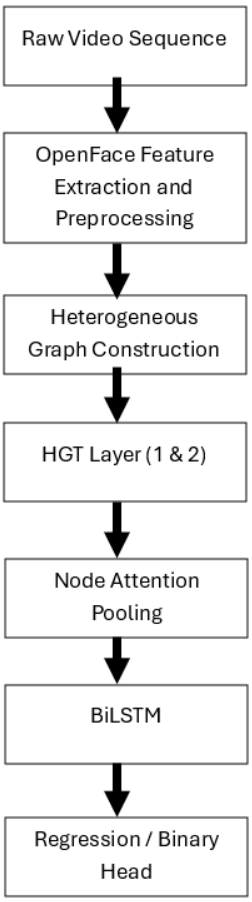


Fig. 2. Overview of the proposed HGT–BiLSTM architecture for spatio-temporal facial pain modeling

2.3 Fine-Grained Pain Formulation (Multiclass)

Fine-grained pain estimation is formulated as an ordinal regression problem to preserve the inherent ordering of pain intensity levels. Rather than predicting discrete class probabilities, the model outputs a single continuous value for each facial sequence, with ground-truth labels ranging from 0 to 4 treated as ordered numerical targets. The model is optimized using Mean Squared Error (MSE) loss to minimize the deviation between predicted and true pain levels, thereby encouraging predictions that respect ordinal relationships.

This regression-based formulation differs from conventional softmax-based multiclass classifiers, as it avoids assuming independence between adjacent pain levels and instead, models intensity as a continuous progression. The approach follows prior pain assessment studies that treat PSPI as a continuous intensity signal rather than a purely categorical label [10], [11]. During evaluation, continuous outputs are discretized by rounding and clamping to the valid label range to enable class-based analysis.

Although fine-grained ordinal regression yields competitive within-dataset performance, subtle boundaries between adjacent pain levels may be sensitive to dataset-specific characteristics, raising concerns about their reliability under cross-dataset conditions. This highlights a potential trade-off between pain-label granularity and cross-dataset transferability. This motivates the investigation of alternative pain abstractions that reduce fine-grained label resolution while retaining meaningful severity distinctions, as examined through the binary formulations described in the following section.

2.4 Coarse Pain Formulation (Binary Class)

To investigate how different levels of pain abstraction influence cross-dataset generalization, fine-grained pain estimation is reformulated as a binary classification task using the same HGT–LSTM architecture trained on the BioVid dataset. The feature extraction pipeline and backbone network remain unchanged, while only the output layer is modified to produce a single scalar logit for each facial sequence.

Specifically, the regression head is replaced with a one-dimensional linear projection that maps the sequence-level representation to a scalar output. This value represents a pain saliency score, indicating the model's confidence that the observed facial dynamics correspond to the high-pain class under a given binary abstraction. Unlike ordinal regression, the binary formulation does not attempt to predict explicit intensity levels. Instead, it captures a coarse representation of pain-related facial behavior that may be more stable under domain shift.

During training on BioVid, original pain labels (BL1, PA1, PA2, PA3 and PA4) are converted into binary targets according to predefined label mappings. These include abstractions adopted from prior studies as well as new groupings proposed in this work, each reflecting a distinct diagnostic objective. The trained binary models are subsequently evaluated on the UNBC-McMaster dataset without further fine-tuning to assess their generalization to clinically observed pain expressions.

By retaining an identical backbone architecture across ordinal and binary formulations, differences observed during cross-dataset evaluation can be attributed to the choice of training abstraction rather than architectural or optimization variations.

2.5 Binary Mapping Strategies for BioVid Training

Three binary pain mapping strategies are considered during training on the BioVid dataset, each designed to investigate a different aspect of pain representation and diagnostic utility:

- Pain Presence Model (0 vs ≥ 1): This mapping distinguishes between the absence and presence of pain-related facial behavior, regardless of severity. It reflects a screening-oriented objective and evaluates whether detecting any pain signal is sufficient for cross-dataset transfer.
- Coarse Pain Model (0–2 vs ≥ 3): The coarse mapping is motivated by diagnostic analysis performed during multiclass ordinal training on BioVid.
- Extreme Pain Model (0 vs 4): Following mapping of prior studies [18], [22], [30], [31]. This mapping focuses on distinguishing the absence of pain from severe pain.

2.6 Binary Mapping Strategies for UNBC McMaster Dataset

The UNBC-McMaster Shoulder Pain Expression Archive provides clinical pain annotations in the form of Observer Pain Ratings (OPR) on a discrete scale from 0 to 5. For cross-dataset evaluation,

OPR values are converted into binary labels using mappings that correspond directly to the binary models trained on the BioVid dataset. The following mappings are applied during evaluation:

- Pain Presence Mapping:

OPR = 0 vs. OPR ≥ 1

- Coarse Pain Mapping:

OPR ≤ 2 vs. OPR ≥ 3

- Extreme Pain Mapping:

OPR = 0 vs. OPR ≥ 4

Since BioVid includes four graded pain levels (0–4) whereas UNBC OPR spans 0–5, the extreme mapping defined as 0 vs 4 in BioVid corresponds to 0 vs ≥4 in UNBC to preserve comparable high-intensity separation across datasets. Lastly, these mappings ensure that each BioVid-trained binary model is evaluated on UNBC under a label interpretation that is consistent with its training objective, while respecting the clinical annotation protocol of the target dataset. No retraining or modification of model parameters is performed during cross-dataset evaluation.

2.7 Dataset

This study employs two widely used facial pain datasets: the UNBC-McMaster Shoulder Pain Archive [5] and the BioVid Heat Pain Database [25]. These datasets represent distinct pain settings, enabling structured cross-dataset evaluation between experimental and clinical conditions.

The UNBC-McMaster Shoulder Pain Archive consists of 200 facial video sequences collected from 25 patients experiencing shoulder pain during active movement tasks. The dataset provides frame-level annotations based on the Prkachin and Solomon Pain Intensity (PSPI) metric, as well as sequence-level observer ratings including Observer Pain Intensity (OPI), Visual Analogue Scale (VAS), Affective-Motivational Scale (AFF), and Sensory Scale (SEN). Due to its clinical setting and fine-grained facial coding, UNBC-McMaster serves as the primary benchmark for cross-dataset evaluation in this study. The BioVid Heat Pain Database contains multimodal recordings of 90 healthy participants subjected to experimentally induced heat pain at four calibrated intensity levels. The dataset is organized into five parts (A–E), corresponding to different experimental conditions.

In this study, BioVid Part A [25] is exclusively used for training both the multiclass and binary pain models. Part A provides stimulus-calibrated pain intensity labels, allowing the derivation of both fine-grained multiclass formulations and coarser binary categorizations. BioVid Part B is excluded due to partial facial occlusion caused by attached EMG sensors, which may interfere with reliable facial feature extraction. BioVid Part C is not utilized because it consists of long continuous recordings with coarse temporal annotations, where pain and neutral expressions may coexist under the same label. Such coarse labeling does not align with the proposed short-window temporal graph modeling framework and would introduce additional computational overhead without clear representational benefit.

Overall, BioVid Part A is used for model development under controlled experimental pain conditions, while UNBC-McMaster is used to evaluate cross-dataset generalization in a clinical pain setting.

2.8 Data Pre-processing

Facial features were extracted from both BioVid and UNBC-McMaster datasets using OpenFace software [32], which processes each video and outputs frame-level AU intensities and head pose parameters (translations and rotations). This approach is consistent with previous automated pain-assessment studies that have used OpenFace for facial feature extraction [6], [30], [33], [34], [35], [36].

In this study, we retained PSPI-relevant AUs (AU04, AU06, AU07, AU09, AU10, and AU45) together with head pose features, including three-dimensional translation and rotation parameters. As a result, each frame was represented by a 12-dimensional feature vector combining local facial muscle activity and global head motion. An overview of the preprocessing and graph construction procedure is shown in Figure 3.

To reduce high-frequency noise introduced by facial tracking variability, a temporal low-pass Butterworth filter was applied independently to each AU and pose channel along the time dimension. Following the facial dynamics filtering settings proposed by [37], a cutoff frequency of 1 Hz was used. This filtering step preserves slower facial dynamics associated with pain expression while suppressing frame-to-frame jitter.

Furthermore, subject-specific z-score normalization was performed. For each subject, the mean and standard deviation were computed across all filtered sequences belonging to that subject, and each sequence was normalized accordingly. This normalization mitigates inter-subject variability while preserving within-subject temporal dynamics, which is essential for cross-dataset analysis. The resulting filtered and normalized OpenFace feature sequences constitute the input to the proposed temporal graph model.

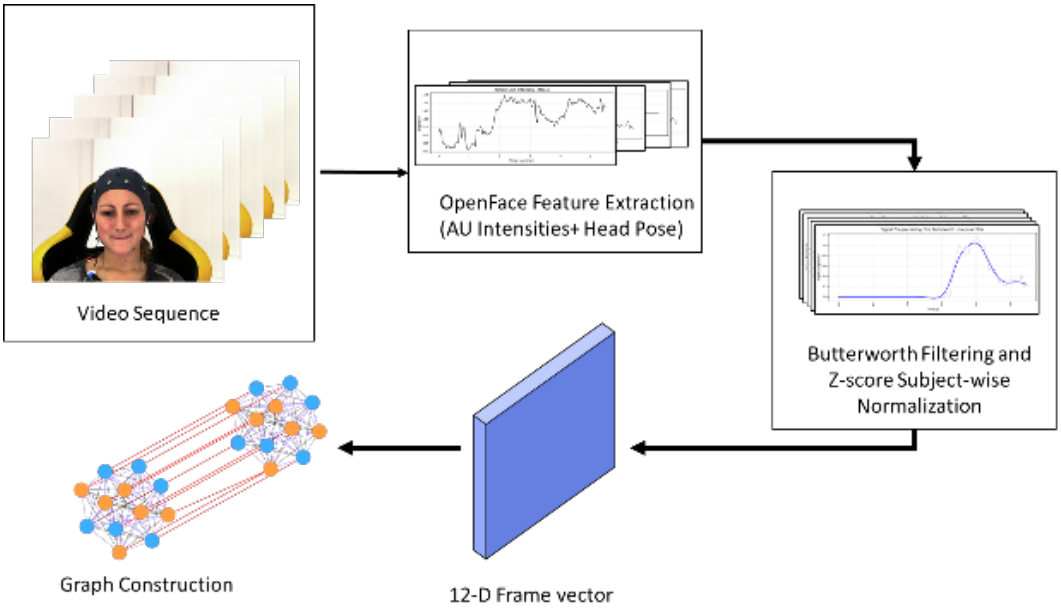


Fig. 3. Preprocessing and graph construction pipeline from facial video to 12-dimensional normalized features for graph-based modeling

2.9 BioVid Temporal Pre-processing

The BioVid dataset provides recordings with a predefined experimental protocol and a fixed temporal length of 138 frames for each sequence across all pain levels and subjects. Consequently, no additional temporal selection is required. During training and evaluation, fixed-length temporal windows of 138 frames are directly used as input to the model.

2.10 UNBC-McMaster Temporal Pre-processing

UNBC-McMaster video sequences vary substantially in duration, and recordings associated with higher observer-rated pain levels tend to be longer than those with low or no pain. In addition, pain-related facial activity in UNBC-McMaster often occurs only during short portions of the video, while much of the sequence shows neutral facial behaviour. Processing full-length sequences can therefore dilute pain-relevant information, particularly given that the binary models in this study are trained on BioVid using fixed-length temporal windows.

To address this issue, each UNBC-McMaster sequence is normalized to a fixed length of K = 128 frames. This value was selected based on preliminary analysis to provide a compact temporal representation that retains the main pain-related segment in sequences where pain is present, while limiting the contribution of long neutral intervals. Although frame-level PSPI values are provided in the dataset, they are not used as supervision targets in this work and are employed only to guide temporal selection.

For sequences labelled as no pain (OPR = 0), frames are selected using uniform temporal sampling across the entire sequence. This preserves typical baseline facial behavior and avoids over-emphasizing brief local variations in AU values that may arise from noise in OpenFace extraction. For sequences with observable pain (OPR $\geq$ 1), a contiguous temporal window of length 128 is extracted and centred on the frame with the maximum PSPI value, allowing the model to focus on the primary pain episode while maintaining temporal continuity. Sequences shorter than 128 frames are retained in full without additional sampling.

2.11 BioVid Data Subject Subset Selection and Evaluation Protocol

To improve the comparability of recognition performance, we followed the experimental recommendations provided in the BioVid Heat Pain Database documentation. Following the protocol proposed by Werner *et al.,* [38], study participants who did not react visibly to the applied pain stimuli were excluded prior to evaluation. These non-reactive subjects exhibit minimal facial expression changes across pain levels and therefore do not provide discriminative facial motion cues for facial landmark-based modelling. Hence, a total of 20 non-reactive subjects were removed, leaving 67 expressive subjects for further analysis.

As recommended by the dataset authors, performance on the BioVid Heat Pain Database is ideally evaluated using leave-one-subject-out (LOSO) cross-validation. However, due to the high computational cost of repeated deep learning training, a subject-level hold-out strategy was adopted following the alternative protocol described in the BioVid documentation. Because this study operates on a filtered expressive-subject subset, the original predefined split cannot be applied directly. An adapted subject-level split was therefore constructed while preserving subject independence and balanced pain-level representation.

Specifically, the remaining subjects were grouped into three age ranges (20–35, 36–50, and 51–65 years). Approximately 30 percent of subjects were assigned to the validation set to match the

proportion used in the original BioVid protocol. The number of validation subjects per age group was determined proportionally to their distribution in the dataset, and an approximately balanced male–female ratio was maintained within each group. To ensure transparency and reproducibility, subject selection within each age and gender subgroup was performed deterministically by sorting subject identifiers alphabetically and selecting the required number of subjects per subgroup. This procedure avoids random sampling and allows the subject split to be reproduced exactly. The resulting split consists of 20 validation subjects and 47 training subjects. The demographic composition of the validation set is summarised in Table 1.

Table 1
Demographic distribution of the validation set

Age range (years)	Validation subjects	Male	Female
20-35	9	4	5
36-50	7	3	4
51-65	4	2	2
Total	20	9	11

2.12 Training Details

The fine-grained multiclass pain model was trained in a regression-based ordinal learning setting. Optimization was performed using the Adam optimizer with an initial learning rate of 0.001 and a weight decay of 0.0001. The model was trained with a batch size of 32 for a maximum of 250 epochs. Mean Squared Error (MSE) loss was employed as loss function. Although the model produces continuous outputs, discrete ordinal pain levels were obtained during evaluation by rounding the predictions and clamping them to the valid label range of (0, 4). To prevent overfitting, early stopping was applied based on the validation Root Mean Square Error (RMSE), with a patience of 100 epochs. The model checkpoint achieving the lowest validation RMSE was retained for final evaluation.

A step-based learning rate scheduler was employed to improve convergence, where the learning rate was reduced by a factor of 0.7 every 30 epochs.

The binary models were optimized using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32. Binary Cross-Entropy with logit loss was used as the training objective. The binary model trained on the BioVid dataset was subsequently evaluated on the UNBC dataset to analyze representation transferability across pain domains. To assess variability associated with stochastic model training, each binary formulation was trained independently across repeated runs while maintaining the same subject-independent data partition, preprocessing procedure, model architecture, and training configuration. A fixed subject split was retained across runs due to the computational cost of evaluating multiple subject partitions. Performance was subsequently summarized as the mean ± standard deviation (SD) across the repeated runs.

All models were trained using the PyTorch framework on a GPU-enabled environment.

2.13 Evaluation Protocol and Metrics

For the multiclass ordinal formulation, we report Accuracy (ACC), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Spearman's rank correlation (ρ), and Quadratic Weighted Kappa (QWK). QWK is included to account for the ordered structure of pain levels by penalizing larger

disagreements more heavily than adjacent errors. Correlation-based metrics are additionally reported to evaluate monotonic relationships between predicted and ground-truth intensity levels.

For cross-dataset diagnostic analysis, Spearman's rank correlation and Pearson correlation are computed between model outputs and ground-truth pain labels to assess monotonic and linear associations, respectively.

For binary evaluation under the UNBC-McMaster dataset, Area Under the Curve (AUC) is used as the primary metric due to its robustness to class imbalance. Accuracy and macro-averaged F1-score are additionally reported to facilitate comparison with prior work. For binary experiments conducted on BioVid, AUC is reported as the principal performance measure.

To quantify prediction error under a calibrated setting, MAE and RMSE are computed after linearly rescaling model outputs to the target label range. Raw model logits are analyzed directly during cross-dataset evaluation and are not assumed to be numerically equivalent to clinical pain ratings. To further quantify uncertainty in the cross-dataset evaluation, 95% confidence intervals (CI) for AUC were estimated using subject-level bootstrap resampling on the UNBC-McMaster dataset. In each resampling, subjects were randomly selected with replacement, and all sequences belonging to each selected subject were kept together. AUC was recalculated for each resampled set of subjects, and the resulting AUC values were used to determine the 95% confidence interval.

3. Result and Discussion

3.1 Multiclass Pain Modeling on BioVid (Within Dataset Analysis)

Table 2 presents the performance of our model compared to the state-of-the-art (SOTA). The selection of comparison methods was restricted to studies evaluated on the BioVid dataset. Differences in data balancing strategies and subject-splitting protocols were not considered, as several prior works do not explicitly report these details.

As shown in Table 2, our proposed model achieves the lowest error among the compared methods, recording an RMSE of 1.25 and MAE of 1.05. This improves upon Alhamdoosh *et al.,* [30] (RMSE: 1.35, MAE: 1.09) and Rajasekhar *et al.,* [10] (MAE: 1.16). In terms of exact multiclass accuracy, Gkikas *et al.,* [39] report the highest accuracy (35.39%), followed by Alhamdoosh *et al.,* [30] (32.09%), while the proposed model achieves 28.30%. The lower exact accuracy must be interpreted in the context of the regression-based formulation adopted in this study. Standard accuracy treats all misclassifications equally and does not account for the ordinal structure of pain levels. Although the exact accuracy is modest, our additional ordinal analysis shows that the model achieves a within-1 accuracy of 74.3% and a within-2 accuracy of 96.3%, indicating that most errors occur between neighboring pain levels.

This behavior is further supported by a moderate Spearman rank correlation $(\rho = 0.43, p < 0.001)$ and a Quadratic Weighted Kappa (QWK) of 0.37, both of which account for ordinal relationships. These results suggest that the proposed model preserves relative pain severity even when exact class matching is imperfect. However, the relatively low F1-macro (25.10) and exact accuracy reflect the inherent ambiguity of fine-grained pain labels. The concentration of errors between adjacent categories raises the question of whether neighbouring pain levels are sufficiently separable using facial expression features alone. To examine this issue more closely, a boundary-wise ordinal analysis is conducted in the following section.

Table 2
SOTA performance comparison

Paper	Method	Metrics					
		ACC (%)	RMSE	MAE	F1 Score	Spearman Rank Correlation (ρ)	Quadratic Weighted Kappa (QWK)
[10]	WSDA-OR (I3DCNN+MIL)	-	-	1.16	-	-	-
[39]	Vision Transformer (ViT)	**35.39**	-	-	-	-	-
[18]	Transformer	31.52	-	-	-	-	-
[40]	RBF-SVM	27.9	-	-	-	-	-
[40]	RF	30.8	-	-	-	-	-
[30]	STAGCN-GRU	32.09	1.35	1.09	31.50	-	-
Our Model	HGT-LSTM	28.30	**1.25**	**1.05**	25.10	0.43	0.37

3.2 Ordinal Boundary Analysis

To further examine the diagnostic reliability of the proposed HGT-LSTM model, a boundary-wise Receiver Operating Characteristic Curve (ROC) analysis was conducted on the BioVid Part A dataset. Because BioVid involves experimentally induced thermal pain with precisely controlled intensity increments, evaluating performance at specific ordinal thresholds provides insight into the model's ability to capture progressive changes in pain severity. Ground-truth labels were binarized at four ordinal boundaries: 0 vs $\geq 1, \leq 1$ vs $\geq 2, \leq 2$ vs ≥ 3, and ≤ 3 vs 4.

As shown in Figure 4, model performance improves significantly as the pain threshold increases. The AUC rises from 0.643 for the lowest threshold $(0$ vs $\geq 1)$ to 0.790 for the highest intensity boundary $(\leq 3$ vs $4)$. This trend suggests that while distinguishing between baseline comfort and mild discomfort is challenging due to the subtle nature of early thermal pain responses, the model becomes increasingly sensitive as facial landmarks and Action Units become more pronounced at higher heat-induced pain levels.

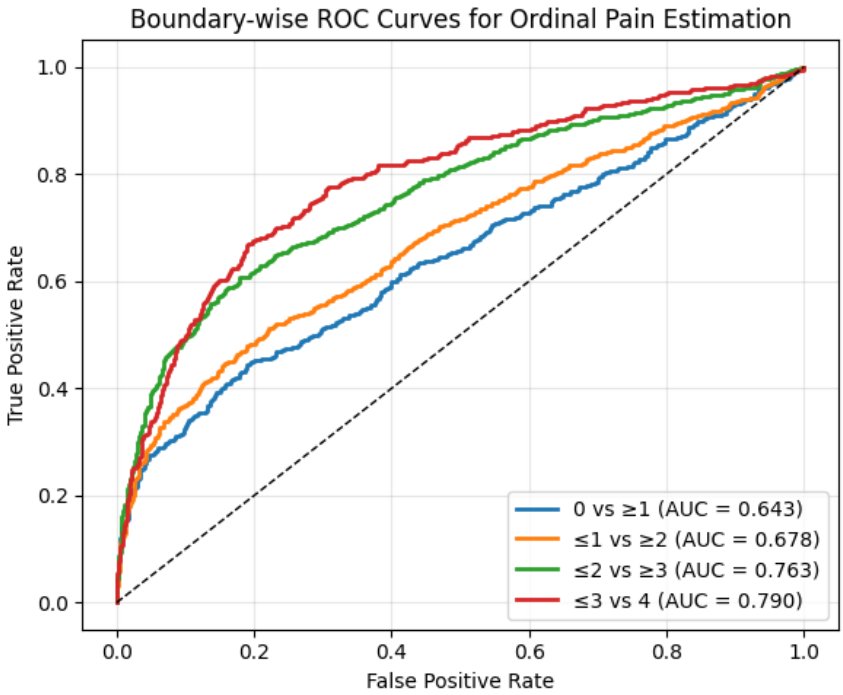


Fig. 4. Cumulative boundary ROC curves for pain estimation. Each curve represents the model's discriminative power at a specific ordinal transition point.

Notably, a significant jump in AUC is observed at the boundary ≤ 2 vs ≥ 3 (AUC = 0.763). This indicates that pain signatures become significantly more distinguishable starting between Level 2 and Level 3. Based on this diagnostic insight from the experimental data, these findings provide empirical motivation for examining a binary classification framework for the subsequent cross-dataset analysis on the UNBC-McMaster dataset. By grouping levels 0, 1, and 2 as "Low Pain" and levels 3 and 4 as "High Pain," the Coarse formulation leverages the stronger separability observed at higher intensity levels while reducing the ambiguity associated with lower-level distinctions. This grouping is examined as a coarser pain representation motivated by the observed increase in separability, rather than being selected solely on the basis of its subsequent cross-dataset performance and is evaluated alongside the Presence and Extreme formulations.

3.3 Cross-Dataset Diagnostic Analysis

To evaluate the clinical generalizability of the HGT-LSTM, and to examine whether the boundary-wise separability observed in BioVid remains consistent under domain shift, we conducted a zero-shot cross-dataset evaluation on the UNBC-McMaster shoulder pain dataset. This represents a significant domain shift from experimentally induced thermal pain in healthy volunteers to spontaneous clinical pain in patients with shoulder pain.

The same BioVid-trained model was evaluated under three binary mapping strategies, which we denote as Presence (0 vs. 1–5), Coarse (0–2 vs. 3–5), and Extreme (0 vs. 4–5).

As shown in Table 3, the relative performance pattern remained consistent across repeated training runs. The Extreme mapping (0 vs. 4–5) achieved the highest cross-dataset AUC of 0.892 ± 0.003, followed by the Coarse mapping (0–2 vs. 3–5) at 0.807 ± 0.018, while the Presence mapping (0 vs. 1–5) achieved 0.691 ± 0.009. Notably, all three formulations achieved higher mean AUC on UNBC-McMaster than on BioVid. This shows that the domain difference between the two datasets did not reduce the model's ability to separate the binary pain groups. The BioVid-trained models

were still able to distinguish the pain groups in UNBC-McMaster, although the AUC differed depending on how the pain levels were grouped. This finding further shows that how pain levels are grouped for classification can influence the results of cross-dataset transfer. The relatively small AUC standard deviations further indicate that the observed discrimination pattern was consistent across repeated training runs. However, because the repeated training runs used the same subject partition, the reported across-run variability primarily reflects stochastic training variation rather than variability arising from alternative subject splits. In comparison, the greater variability in accuracy and Macro F1 suggests it were more sensitive to stochastic training variation than AUC.

Correlation analysis provides a complementary view of the cross-dataset performance. The Presence mapping showed the strongest association with the original UNBC pain ratings, with Pearson and Spearman correlations of 0.593 ± 0.009 and 0.488 ± 0.008, respectively, despite producing the lowest binary classification AUC. The Coarse and Extreme mappings showed slightly lower correlations, as reported in Table 3. This difference reflects the distinct information captured by the evaluation metrics. AUC evaluates discrimination according to the specified binary boundary, whereas Pearson and Spearman correlations evaluate the association between the continuous model outputs and the original OPR ratings across the pain scale. Thus, lower binary discrimination does not necessarily imply a weaker association with the overall ordinal pain ratings. Similarly, although AUC differed considerably across the three formulations, their calibrated MAE values were relatively similar. This indicates that improved separation of the binary pain groups did not necessarily result in predictions that were closer to the original UNBC pain ratings.

Table 3
Performance Generalization from Experimental (BioVid) to Clinical (UNBC) Data

Model Mapping	BioVid (Training)	UNBC (Testing)					
	AUC	AUC	ACC (%)	MAE	F1 Score	Pearson Correlation (r)	Spearman Rank Correlation (ρ)
Presence (0 vs. 1-5)	0.656 ± 0.001	0.691 ± 0.009	64.000 ± 2.092	**1.047 ± 0.010**	63.416 ± 2.551	**0.593 ± 0.009**	**0.488 ± 0.008**
Coarse (0-2 vs. 3-5)	0.752 ± 0.001	0.807 ± 0.018	73.000 ± 4.257	1.086 ± 0.038	70.047 ± 3.326	0.558 ± 0.038	0.456 ± 0.032
Extreme (0 vs. 4-5)	**0.840 ± 0.001**	**0.892 ± 0.003**	**76.000 ± 4.626**	1.092 ± 0.016	**71.457 ± 3.981**	0.554 ± 0.013	0.469 ± 0.011

Subject-level bootstrap analysis showed greater uncertainty across the evaluated UNBC-McMaster subjects than across repeated model training. The 95% AUC confidence intervals ranged from 0.607 to 0.775 for Presence, 0.723 to 0.892 for Coarse, and 0.797 to 0.970 for Extreme. These ranges were considerably wider than the AUC variation observed across repeated training runs. This suggests that the cross-dataset AUC was more sensitive to differences among the evaluated clinical

subjects than to random variation during model training. Because UNBC-McMaster contains a limited number of subjects, the reported AUC values should therefore be interpreted with this uncertainty in mind. However, the repeated training runs consistently showed the same performance pattern across the three formulations.

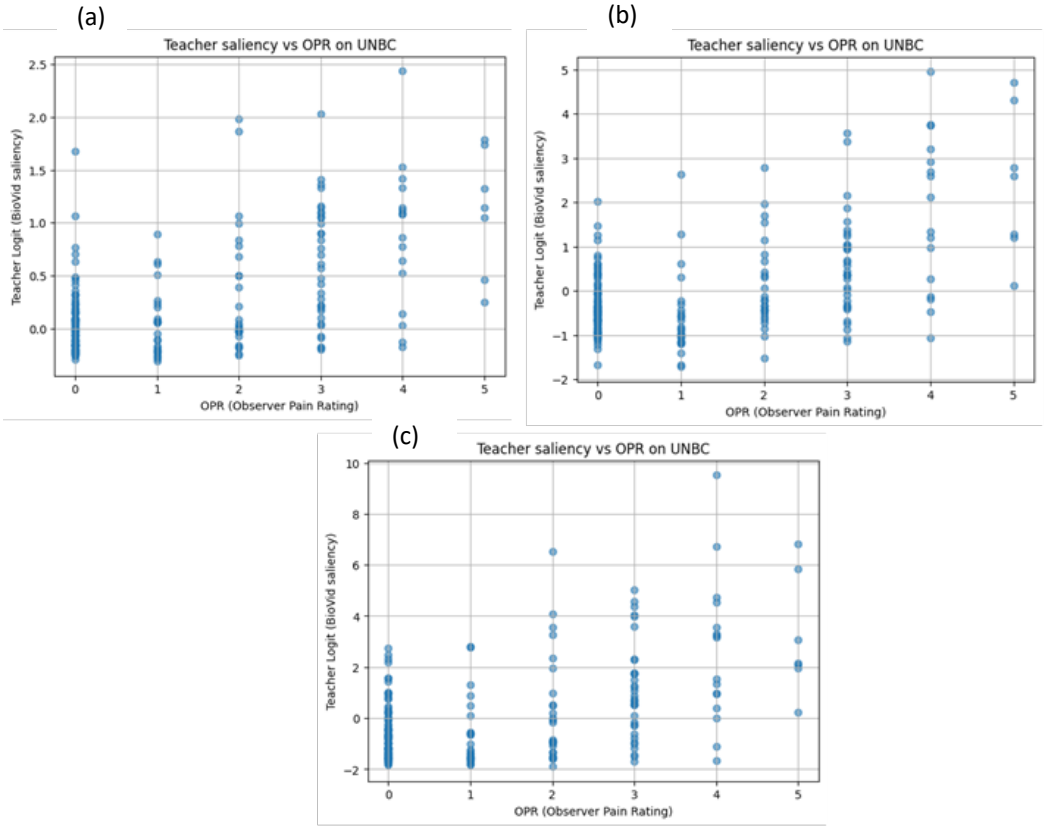


Fig. 5. (a) Presence, (b) Coarse, (c) Extreme Scatter Plots

For qualitative interpretation of the learned facial representations, the saliency distributions shown in Figure 5 were obtained from a representative trained model for each formulation and are interpreted separately from the quantitative variability assessed across repeated training runs.

Among the evaluated strategies, the Coarse pain model shows a smoothly varying saliency distribution. In this configuration, while adjacent OPR levels overlap, particularly around OPR 2–3, higher OPR levels (4–5) are associated with consistently higher scores, indicating preservation of the ordinal severity trend rather than forced discrete separation.

Conversely, the Presence model shows considerable overlap between OPR 0 and 1, which limits discrimination at the 0 vs. ≥1 boundary. Despite this overlap at the lower OPR levels, the model outputs generally increase as OPR increases, which is consistent with the stronger positive association with the original OPR ratings observed in the correlation analysis. The Extreme model similarly shows an overall increase in model outputs toward higher OPR levels, although substantial overlap remains across the intermediate pain levels. This is consistent with its positive correlation with the original OPR ratings, despite being trained to distinguish only the lowest and highest BioVid pain levels. Its higher binary discrimination therefore does not imply clear separation between all individual OPR severity levels.

The differences observed across the three formulations can be further understood by considering how the original BioVid pain levels were grouped and retained during training. The BioVid subset

contained 6,700 sequences distributed equally across the five pain levels. The Presence formulation (0 vs. ≥1) retained all sequences but resulted in a highly imbalanced 20:80 class distribution. The Coarse formulation (≤2 vs. ≥3) also retained the complete dataset while providing a more balanced 60:40 distribution. In contrast, the Extreme formulation (0 vs. 4) produced equal class proportions but excluded the intermediate pain levels (Levels 1–3), retaining only 40% of the available sequences. Therefore, although Extreme achieved the highest mean cross-dataset AUC (0.892 ± 0.003), this performance reflects discrimination between the endpoints of the BioVid pain scale. Coarse achieved a lower mean AUC (0.807 ± 0.018) but retained information across the complete severity range.

Importantly, the cross-dataset findings should also be interpreted in relation to the substantial difference in pain context between BioVid and UNBC-McMaster. BioVid represents experimentally induced thermal pain in healthy participants, where pain intensity is controlled through calibrated heat stimulation, whereas UNBC-McMaster represents clinical shoulder pain observed in patients during movement. Consequently, the pain labels in the two datasets do not represent identical pain experiences or annotation conditions. BioVid pain levels are linked to controlled experimental stimulation, while the UNBC-McMaster OPR reflects observer-rated pain behaviour in a clinical setting. The temporal characteristics also differ, as BioVid contains standardized stimulus sequences whereas UNBC-McMaster contains variable-duration movement sequences in which pain-related facial responses may occur transiently. These differences, mean that the BioVid-to-UNBC evaluation should be interpreted as a test of cross-dataset discrimination under substantially different pain conditions rather than as evidence that the two pain scales are directly equivalent. Nevertheless, the ability of the BioVid-trained models to discriminate the corresponding pain groups in UNBC-McMaster indicates that the learned facial representations retained discriminative information across these different pain conditions, with the level of discrimination varying according to the pain-label formulation.

Furthermore, demographic and inter-subject differences between BioVid and UNBC-McMaster may also contribute to cross-dataset variability. Although the BioVid subject split was stratified by age and maintained an approximately balanced male-female distribution, the present analysis focused on cross-dataset discrimination under different pain-label formulations rather than demographic-specific performance. Previous studies have reported variations in pain responses across age and gender groups [41], individual differences in facial expressiveness during experimentally induced pain [38], and demographic bias in computer-vision-based pain detection [42]. Therefore, factors such as age, sex, ethnicity, facial morphology, and individual facial expressiveness may contribute to differences between the experimental and clinical datasets. These factors were not analysed independently in the present study and should therefore be considered when interpreting the cross-dataset results. Future work using larger and demographically characterized cohorts could investigate their effects separately.

In addition to demographic and inter-subject variability, uncertainty in the facial features automatically extracted by OpenFace may also influence cross-dataset performance. Automatic facial feature estimation can be affected by differences in recording and facial conditions. Namba *et al.,* [43] reported that changes in facial viewpoint can affect the performance of automated AU detection systems. In clinical pain assessment, Yuan *et al.,* [44] noted that variations in head pose and facial occlusion can affect OpenFace AU detection confidence, while Kim *et al.,* [45] highlighted the effects of illumination variability on facial landmark detection and AU intensity measurement. These factors are particularly relevant to the present cross-dataset evaluation because BioVid and UNBC-McMaster differ in their recording conditions and facial behaviour. Consequently, variations in the automatically extracted AU and head-pose features may contribute to the observed cross-dataset differences in addition to differences in pain expression. The present study did not independently validate the

OpenFace-derived AU or head-pose estimates against manually annotated ground truth. Therefore, the extent of their influence on the cross-dataset results cannot be determined from the present experiments.

3.4 Comparative Analysis with SOTA

Direct comparison with prior SOTA methods remains challenging due to the absence of standardized benchmarks and consistent evaluation protocols in automatic pain estimation, as noted by [3], [8]. To the best of our knowledge, limited prior work evaluates direct experimental-to-clinical transfer between BioVid and UNBC, making strict benchmarking under identical protocols challenging.

To provide meaningful context, Table 4 compares the proposed model with representative SOTA approaches evaluated on either clinical or experimental datasets, regardless of stimulus type. This reflects the practical need for systems that operate reliably across diverse medical scenarios rather than being optimized for a single dataset.

As shown in Table 4, the proposed models achieved performance comparable to representative previous studies, despite being evaluated under a direct BioVid-to-UNBC cross-dataset setting. In terms of accuracy, the Extreme formulation achieved 78.30%, which was higher than the 73.28% reported in Gkikas *et al.,* [18] and the 78.14% reported in Morsali *et al.,* [31], while remaining slightly below the 79.95% reported in [46]. The Coarse formulation achieved an accuracy of 75.90%, which was also higher than the 73.28% reported in Gkikas *et al.,* [18]. For AUC, the Extreme formulation achieved 89.0%, compared with 93.3% reported in Rodriguez *et al.,* [11]. However, these numerical comparisons should be interpreted cautiously because the studies differ in training and testing datasets, pain-label definitions, and evaluation protocols. Unlike the compared studies, the present model was trained on experimentally induced thermal pain in BioVid and evaluated directly on clinical shoulder pain in UNBC-McMaster without target-domain fine-tuning. Therefore, the comparison demonstrates that the proposed model retained competitive discrimination under cross-dataset transfer rather than establishing a direct performance ranking against previous methods.

Table 4
Comparative Performance against SOTA Methods

Study	Training Data	Testing Data (External)	Metrics			
			AUC	ACC (%)	MAE	F1 Score
[18]	BioVid	-		73.28	-	72.75
[46]	AI4PAIN	BioVid	-	79.95	-	-
[31]	UNBC	BioVid	-	78.14	-	-
[11]	UNBC	-	93.3	-	0.5	-
Our Model (Presence)	BioVid	UNBC	69.0	66.50	1.05	66.4

Our Model (Coarse)	BioVid	UNBC	80.0	75.90	1.09	72.0
Our Model (Extreme)	BioVid	UNBC	89.0	78.30	1.12	74.0

4. Conclusion

This study presented a diagnostic analysis of facial pain representations using a heterogeneous graph-based spatio-temporal model. While fine-grained pain modelling on the BioVid dataset achieved competitive error performance, ordinal boundary analysis revealed substantial overlap between adjacent pain levels, particularly at lower intensities, with clearer separation observed at higher pain boundaries. These findings indicate that the individual pain levels in BioVid are not equally separable from facial representations.

Motivated by this observation, three binary pain-label formulations were evaluated to examine how pain-label granularity influences cross-dataset performance. Zero-shot evaluation from experimentally induced thermal pain in BioVid to clinical shoulder pain in UNBC-McMaster showed that cross-dataset discrimination varied considerably according to how the pain levels were grouped. The Presence formulation produced the lowest discrimination, while the Coarse formulation improved discrimination while retaining the complete BioVid severity range. The Extreme formulation achieved the highest discrimination but excluded the intermediate BioVid pain levels. These results show a trade-off between binary pain separability and representation of the complete severity range.

Correlation analysis provided a complementary finding. Although the Presence formulation achieved the lowest binary discrimination, it maintained the strongest positive association with the original UNBC OPR ratings. This indicates that stronger discrimination at a selected binary pain boundary does not necessarily correspond to a stronger relationship with the complete ordinal pain scale. Taken together, the results demonstrate that pain-label granularity influences cross-dataset facial pain assessment not only through binary discrimination, but also through the amount of severity information retained by each formulation. The Coarse formulation therefore provides a practical trade-off by retaining the complete BioVid severity range while achieving improved cross-dataset discrimination and maintaining a positive association with the original clinical pain ratings.

These findings should be considered in the context of the substantial differences between BioVid and UNBC-McMaster in pain elicitation, participant characteristics, annotation procedures, and temporal characteristics. Rather than identifying a universally optimal pain grouping, the results highlight the importance of considering pain-level separability, severity coverage, and the relationship with the original pain scale when transferring facial pain representations across different datasets. Future work may investigate domain adaptation and multimodal information to further improve transfer across heterogeneous pain settings.

Acknowledgement
The authors would like to express their gratitude to Universiti Teknologi Malaysia (UTM) for providing funding and facilities to conduct this research with Project No. Q.J130000.2724.04K18

References

[1] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. W. Picard, "Automatic Recognition Methods Supporting Pain Assessment: A Survey," IEEE Trans. Affect. Comput., vol. 13, no. 1, pp. 530–552, 2022. https://doi.org/10.1109/TAFFC.2019.2946774

[2] P. Casti et al., "Calibration of Vision-Based Measurement of Pain Intensity with Multiple Expert Observers," IEEE Trans. Instrum. Meas., vol. 68, no. 7, pp. 2442–2450, 2019. https://doi.org/10.1109/TIM.2019.2909603

[3] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, S. Gruss, and S. Walter, "Automatic vs. Human recognition of pain intensity from facial expression on the x-ite pain database," Sensors, vol. 21, no. 9, pp. 1–19, 2021. https://doi.org/10.3390/s21093273

[4] M. Kächele et al., "Multimodal data fusion for person-independent, continuous estimation of pain intensity," Communications in Computer and Information Science, vol. 517, pp. 275–285, 2015. https://doi.org/10.1007/978-3-319-23983-5_26

[5] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011, pp. 57–64, 2011. https://doi.org/10.1109/FG.2011.5771462

[6] E. Othman, P. Werner, F. Saxen, M.-A. Fiedler, and A. Al-Hamadi, "An Automatic System for Continuous Pain Intensity Monitoring Based on Analyzing Data from Uni-, Bi-, and Multi-Modality," 2022. https://doi.org/10.3390/s22134992

[7] T. Hassan et al., "Automatic Detection of Pain from Facial Expressions: A Survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 6, pp. 1815–1831, 2021. https://doi.org/10.1109/TPAMI.2019.2958341

[8] S. Gkikas and M. Tsiknakis, "Automatic assessment of pain based on deep learning methods: A systematic review," Comput. Methods Programs Biomed., vol. 231, 2023. https://doi.org/10.1016/j.cmpb.2023.107365

[9] Z. Hammal and J. F. Cohn, "Automatic detection of pain intensity," in Proceedings of the 14th ACM international conference on Multimodal interaction, New York, NY, USA: ACM, Oct. 2012, pp. 47–52. https://doi.org/10.1145/2388676.2388688

[10] G. P. Rajasekhar, E. Granger, and P. Cardinal, "Deep domain adaptation with ordinal regression for pain assessment using weakly-labeled videos," Image Vis. Comput., vol. 110, p. 104167, 2021. https://doi.org/10.1016/j.imavis.2021.104167

[11] P. Rodriguez et al., "Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification," IEEE Trans. Cybern., vol. 52, no. 5, pp. 3314–3324, 2022. https://doi.org/10.1109/TCYB.2017.2662199

[12] M. T. Vu and M. Beurton-Aimar, "Learning to focus on region-of-interests for pain intensity estimation," 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition, FG 2023, pp. 1–6, 2023. https://doi.org/10.1109/FG57933.2023.10042583

[13] C. Florea, L. Florea, and C. Vertan, "Learning pain from emotion: Transferred HoT data representation for pain intensity estimation," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8927, no. February 2017, pp. 778–790, 2015. https://doi.org/10.1007/978-3-319-16199-0_54

[14] S. Kaltwang, S. Todorovic, and M. Pantic, "Doubly Sparse Relevance Vector Machine for Continuous Facial Behavior Estimation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 9, pp. 1748–1761, 2016. https://doi.org/10.1109/TPAMI.2015.2501824

[15] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent Convolutional Neural Network Regression for Continuous Pain Intensity Estimation in Video," IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1535–1543, 2016. https://doi.org/10.1109/CVPRW.2016.191

[16] M. Tavakolian and A. Hadid, "Deep binary representation of facial expressions: A novel framework for automatic pain intensity recognition," Proceedings - International Conference on Image Processing, ICIP, pp. 1952–1956, 2018. https://doi.org/10.1109/ICIP.2018.8451681

[17] S. El Morabit and A. Rivenq, "Pain Detection From Facial Expressions Based on Transformers and Distillation," 11th International Symposium on Signal, Image, Video and Communications, ISIVC 2022 - Conference Proceedings, pp. 1–5, 2022. https://doi.org/10.1109/ISIVC54825.2022.9800746

[18] S. Gkikas and M. Tsiknakis, "A Full Transformer-based Framework for Automatic Pain Estimation using Videos," Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp. 1–6, 2023. https://doi.org/10.1109/EMBC40787.2023.10340872

[19] X. Yuan, S. Zhang, C. Zhao, X. He, B. Ouyang, and S. Yang, "Pain Intensity Recognition from Masked Facial Expressions using Swin-Transformer," 2022 IEEE International Conference on Robotics and Biomimetics, ROBIO 2022, pp. 723–728, 2022. https://doi.org/10.1109/ROBIO55434.2022.10011731

[20] L. Wang, Z. Wang, A. Xu, and S. Liu, "A Generative Adversarial Network-Based Approach for Facial Pain Assessment," Proceedings - 2024 8th International Conference on Biomedical Engineering and Applications, ICBEA 2024, pp. 44–49, 2024. https://doi.org/10.1109/ICBEA62825.2024.00017

[21] J. Farmani, G. Bargshady, S. Gkikas, M. Tsiknakis, and R. F. Rojas, "A CrossMod-Transformer deep learning framework for multi-modal pain detection through EDA and ECG fusion," Sci. Rep., vol. 15, no. 1, p. 29467, Aug. 2025. https://doi.org/10.1038/s41598-025-14238-y

[22] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, and S. Walter, "Cross-Database Evaluation of Pain Recognition from Facial Video," in 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), IEEE, Sep. 2019, pp. 181–186. https://doi.org/10.1109/ISPA.2019.8868562

[23] P. Prajod, T. Huber, and E. André, "Using Explainable AI to Identify Differences Between Clinical and Experimental Pain Detection Models Based on Facial Expressions," 2022, pp. 311–322. https://doi.org/10.1007/978-3-030-98358-1_25

[24] J. F. Cohn, I. O. Ertugrul, W. S. Chu, J. M. Girard, L. A. Jeni, and Z. Hammal, "Affective facial computing: Generalizability across domains," in Multimodal Behavior Analysis in the Wild: Advances and Challenges., Elsevier, 2018, pp. 407–441. https://doi.org/10.1016/B978-0-12-814601-9.00026-2

[25] S. Walter et al., "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), IEEE, Jun. 2013, pp. 128–131. https://doi.org/10.1109/CYBConf.2013.6617456

[26] R. Fang et al., "ATLAS: An Adaptive Transfer Learning Based Pain Assessment System: A Real Life Unsupervised Pain Assessment Solution," Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2022-July, pp. 1331–1337, 2022. https://doi.org/10.1109/EMBC48229.2022.9871536

[27] H. Liu, H. Xu, J. Qiu, S. Wu, and M. Liu, "Hierarchical global and local transformer for pain estimation with facial expression videos," Pattern Analysis and Applications, vol. 27, no. 3, p. 85, Sep. 2024. https://doi.org/10.1007/s10044-024-01302-y

[28] F. Gao, L. Qing, L. Li, G. Yang, R. Xu, and L. Gao, "Enhancing pain intensity evaluation via an attention-driven channel–spatial fusion network," Visual Computer, vol. 41, no. 15, pp. 12477–12496, 2025. https://doi.org/10.1007/s00371-025-04167-2

[29] Z. Lu, F. A. B. Zulkifle, A. Z. U. S. B. M. Japeri, M. R. Bin Shamsuddin, and A. B. M. Jasin, "Handling Class Imbalance in Video-Based Pain Intensity Estimation Using ResNetBiLSTM: A Study on Loss Functions and Optimizers," International Journal of Electrical and Electronics Research, vol. 13, no. 4, pp. 820–829, 2025. https://doi.org/10.37391/IJEER.130422

[30] F. Alhamdoosh, P. Pala, and S. Berretti, "Pain Level Estimation From Videos by Analyzing the Dynamics of Facial Landmarks With a Spatio-Temporal Graph Neural Network," IEEE Trans. Biom. Behav. Identity Sci., vol. 7, no. 4, pp. 610–619, 2025. https://doi.org/10.1109/TBIOM.2025.3592836

[31] M. Morsali and A. Ghaffari, "Enhanced residual attention-based subject-specific network (ErAS-Net): facial expression-based pain classification with multiple attention mechanisms," Sci. Rep., vol. 15, no. 1, p. 19425, Jun. 2025. https://doi.org/10.1038/s41598-025-04552-w

[32] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, May 2018, pp. 59–66. https://doi.org/10.1109/FG.2018.00019

[33] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, "Automatic Pain Assessment with Facial Activity Descriptors," IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 286–299, 2017. https://doi.org/10.1109/TAFFC.2016.2537327

[34] Y. Zhao et al., "Pose-invariant and occlusion-robust neonatal facial pain assessment," Comput. Biol. Med., vol. 165, no. September, p. 107462, 2023. https://doi.org/10.1016/j.compbiomed.2023.107462

[35] M. Cascella et al., "AI-based cancer pain assessment through speech emotion recognition and video facial expressions classification," Signa Vitae, vol. 20, no. 12, pp. 28–38, 2024. https://doi.org/10.22514/sv.2024.153

[36] D. Bourou, A. Pampouchidou, M. Tsiknakis, K. Marias, and P. Simos, "Video-based Pain Level Assessment: Feature Selection and Inter-Subject Variability Modeling," 2018 41st International Conference on Telecommunications and Signal Processing, TSP 2018, pp. 1–6, 2018. https://doi.org/10.1109/TSP.2018.8441252

[37] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue, "Automatic Pain Recognition from Video and Biomedical Signals," in 2014 22nd International Conference on Pattern Recognition, IEEE, Aug. 2014, pp. 4582–4587. https://doi.org/10.1109/ICPR.2014.784

[38] P. Werner, A. Al-Hamadi, and S. Walter, "Analysis of facial expressiveness during experimentally induced heat pain," in 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), IEEE, Oct. 2017, pp. 176–180. https://doi.org/10.1109/ACIIW.2017.8272610

[39] S. Gkikas et al., "Multimodal automatic assessment of acute pain through facial videos and heart rate signals utilizing transformer-based architectures," Frontiers in Pain Research, vol. 5, Mar. 2024. https://doi.org/10.3389/fpain.2024.1372814

[40] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, "Automatic Pain Assessment with Facial Activity Descriptors," IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 286–299, 2017. https://doi.org/10.1109/TAFFC.2016.2537327

[41] D. Albahdal, W. Aljebreen, and D. M. Ibrahim, "PainMeter: Automatic Assessment of Pain Intensity Levels from Multiple Physiological Signals Using Machine Learning," IEEE Access, vol. 12, no. April, pp. 48349–48365, 2024. https://doi.org/10.1109/ACCESS.2024.3384359

[42] S. Fabi, X. Xu, and V. R. de Sa, "Exploring the Racial Bias in Pain Detection with a Computer Vision Model," Proceedings of the 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022, pp. 358–365, 2022.

[43] S. Namba, W. Sato, and S. Yoshikawa, "Viewpoint robustness of automated facial action unit detection systems," Applied Sciences (Switzerland), vol. 11, no. 23, 2021. https://doi.org/10.3390/app112311171

[44] X. Yuan et al., "Occluded Facial Pain Assessment in the ICU Using Action Units Guided Network," IEEE J. Biomed. Health Inform., vol. 28, no. 1, pp. 438–449, 2024. https://doi.org/10.1109/JBHI.2023.3336157

[45] J. H. Kim et al., "An exploratory study of headache pain intensity using facial expressions and APEX frames," NPJ Digit. Med., vol. 9, no. 1, pp. 1–13, 2026. https://doi.org/10.1038/s41746-026-02617-2

[46] G. Bargshady, C. Joseph, N. Hirachan, R. Goecke, and R. F. Rojas, "Acute Pain Recognition from Facial Expression Videos using Vision Transformers," 2024 46th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2–5, 2024. https://doi.org/10.1109/EMBC53108.2024.10781616