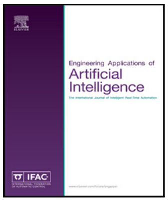
Engineering Applications of Artificial Intelligence 154 (2025) 110966

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Enhanced deep learning framework for real-time pain assessment using multi-modal fusion of facial features and video streams

Hany El-Ghaish [1,2], Mohamed Yousry Al-Basiouny [*,2,3], Mahmoud A.M. Alshewimy [2,4]

Tanta University, Faculty of Engineering, Computer and Control Engineering Dept., Tanta, Egypt

ARTICLE INFO

Dataset link: https://sites.pitt.edu/~emotion/um-spread.html, https://www.nit.ovgu.de/en/BioVid-p-1358.html

Keywords:
Real-time pain assessment
Facial landmark analysis
Attention-based long short-term memory networks
Convolutional neural network
Multimodal data fusion
Gradient-weighted Class Activation Mapping

ABSTRACT

This paper introduces an enhanced deep learning multimodal framework for real-time pain assessment that addresses the limitations of the single-mode approaches. Specifically, time-related disadvantages and its poor temporal analysis, noted in recent studies. The proposed framework fuses VGG16 (Visual Geometry Group 16-layer network), a pre-trained convolutional neural network (CNN), with a bidirectional long short-term memory network (Bi-LSTM) augmented with attention mechanisms. VGG16 has a proven ability to capture fine-grained spatial information in facial expressions to extract powerful spatial features, whereas the Bi-LSTM augmented with attention mechanisms because of its computational efficiency and concisely captures temporal features. Moreover, adaptive boosting has used seeking generalization across a broad base of settings and Gradient-weighted Class Activation Mapping (Grad-CAM) to provide comprehensible and interpretable visual explanations of the decision-making processes of the framework. Extensive evaluations were conducted on diverse popular benchmark datasets: BioVid, UNBC-McMaster, and MIntPAIN. Our framework achieves an AUC (the Area under the Receiver Operating Characteristic Curve) of 0.99 and 99% cross-validation accuracy, corresponding to a 12.5% improvement in accuracy and a 20% reduction in false positives relative to state-of-the-art convolutional neural network–long short-term memory (CNN–LSTM) models. Real-time deployment on an NVIDIA Jetson Nano yields 120 ms (ms) inference per frame with less than 1% accuracy loss, demonstrating its feasibility for edge-based telemedicine. Combined with explainability via Grad-CAM and privacy-preserving vision-only inputs, this framework offers an ethical, scalable solution for nonverbal pain detection.

1. Introduction

Advancements in computer vision and deep learning have created new opportunities for pain assessment, addressing critical challenges in traditional healthcare methods. The currently used methods of pain detection are rely mostly on the self-reporting of patients or monitoring of physiological signals, which might not provide the full information when patients are unable to communicate verbally or in cases where continuous monitoring is needed. Consequently, there is a critical need in the field of telemedicine and especially in an environment where remote patient management becomes necessary, for objective and automated mechanisms that can assess pain correctly.

However, existing methodologies face limitations such as cross-population data variability, low sensitivity to subtle pain cues, and inability to meet real-time processing demands. These challenges hinder clinical deployment in dynamic, real-world environments. Recent state-of-the-art pain assessment methods, including Transformer-based architectures (Gkikas et al. 2024) (Gkikas et al., 2024). and masked autoencoders (Nguyen et al. 2024) (Nguyen et al., 2024), report high precision (92%) but suffer from computational inefficiency and limited generalizability across diverse populations. While CNN-LSTM hybrid approaches(Badura et al. (2024) (Badura et al., 2024)explore the prospects of multi-modal fusion, they do not meet real-time functionality expectations. Similarly, wearable technologies (Kong et al. 2023) (Kong et al., 2022) and sensor-based fusion (Othman et al. 2019) (Othman et al., 2023) exposed to noise and modality dependency.

[*] Corresponding author.
E-mail addresses: hany_elghaish@f-eng.tanta.edu.eg (H. El-Ghaish), pg_165416@f-eng.tanta.edu.eg (M.Y. Al-Basiouny), mahmoud_elshwaimi@f-eng.tanta.edu.eg (M.A.M. Alshewimy).
[1] Assistant Professor.
[2] These authors contributed equally to this work.
[3] Engineer.
[4] Associate Professor.

https://doi.org/10.1016/j.engappai.2025.110966
Received 22 January 2025; Received in revised form 24 March 2025; Accepted 21 April 2025
Available online 6 May 2025

The Proposed approach addresses this gap by introduce a novel framework that uniquely combines VGG16 convolutional neural networks(CNNs) (Purwono et al., 2022; Younesi et al., 2024) for spatial feature extraction with Long Short-Term Memory networks (LSTMs) (Van Houdt et al., 2020) with attention mechanisms for temporal modeling. This framework aligns with general deep learning advances, such as noise-resistant graph processing (ZXing et al. 2023) (Xing et al., 2023) and fusion of heterogeneous sources (ZXing et al. 2024) (Xing et al., 2025), while being tailored for deployment on edge devices. The complementary multimodal fusion achieved in this way improves accuracy by 12.5% and reduces false positives by 20% over previous approaches, thus satisfying telemedicine demands for timely and interpretable pain estimation.

Real-Time telemedicine use cases possess strict constraints such as network latency, variability in video quality, and hardware constraints on edge devices. We optimized our framework for real-time inference on edge hardware while maintaining clinical-grade accuracy. This guarantees steadfastness in clinical environments where pain intensity assessment at the right time is of the utmost importance.

The framework is evaluated using the BioVid dataset (diverse pain stimuli) (Prajod et al., 2024), which is used as the primary benchmark, in addition to the UNBC-McMaster Shoulder Pain Archive(specificity), (Alghamdi and Alaghband, 2022) and the MIntPAIN dataset (real-world latency) (Aoun, 2024), Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2020) visualization enhances interpretability by identifying critical facial regions relevant to pain detection, thus making the model adaptive in diverse and fast-changing conditions.

The contributions of this work can be summarized as follows:

1. Enhanced Deep Learning Framework: Development of an advanced deep learning framework designed for real-time pain assessment, integrating multiple modalities for improved accuracy.
2. Multimodal Fusion: Utilization of facial features and video streams to combine spatial and temporal information for a comprehensive pain analysis.
3. Attention-Augmented LSTM Networks: Implementation of attention mechanisms in LSTM networks to enhance the model's ability to focus on relevant temporal patterns, improving the detection of pain cues over time.
4. Integration of Pre-Trained VGG16: Use of a pre-trained VGG16 convolutional neural network to extract spatial features from video streams, providing a strong foundation for the spatial analysis of pain indicators.
5. Boosting Techniques for Generalizability: Application of boosting methods to improve the generalizability of the model across different datasets and scenarios, ensuring robust performance.
6. Grad-CAM for Visual Interpretability: Incorporation of Grad-CAM (Gradient-weighted Class Activation Mapping) to offer visual explanations of the model's decision-making process, enhancing the transparency and trustworthiness of the pain assessment system.
7. The framework achieves a 12.5% improvement in accuracy and 20% reduction in false positives over prior CNN-LSTM models.

These contributions collectively advance the field of automated pain assessment in real-time by offering a sophisticated, interpretable, and generalizable solution. The obtained results demonstrate that the proposed framework constantly outperforms the state-of-the-art methods with an AUC (the Area under the Receiver Operating Characteristic Curve) of 0.99 and 99% accuracy.

The following sections outline the methodology, feature extraction processes, data augmentation techniques, and the evaluation protocols. It is demonstrated that the framework introduced is feasible and is of great efficacy to provide a substantial contribution in this domain of automated pain detection for the healthcare industry.

2. Literature review

The research in pain assessment has progressed from traditional approaches based on self-reporting and physical measurements, which are not effective in all populations or contexts. Then moved through the stage of classical machine learning using physiological signals, to those approaches using deep learning with multimodal data fusion for real-time inference (Mienye and Swart, 2024). This literature review section is categorized into classical machine learning methods, early deep learning methods, and modern deep learning with advanced fusion and attention mechanisms and what recent studies have added and where they limitations and highlighting the gaps that our study wants to fill.

2.1. Classical machine learning approaches

Pouromran et al. (2021) designed and explored the specific machine-learning model SVM to use the physiological signals EDA, ECG, and EMG for predicting the pain level. By integrating piecewise statistical aggregations, the approach reaches a state of the art in terms of classification accuracy (87%) among all applied classifiers. However, the results emphasize the progress of merging several physiological indicators together for better identification. Nonetheless, there are issues related to the long-standing reliance on such devices even in environmental settings, which makes the product less applicable to the real world.

Similarly, Elgendy et al. (2021) have additionally introduced a pain recognition and categorization model by exploiting conventional ML techniques, explicitly KNN and Adaboost along with Gabor filters for feature extraction and Relieff-SADE for feature selection. The proposed method achieved 91% of pain recognition accuracy and up to 92% of accuracy in the identification of severe pain. The technique is a highly successful way of solving the problem but it is only based on face analysis in still images, which is a drawback of this system in multinational environments and real-time processing.

While classical methods achieved moderate accuracy, their reliance on single-modality data (e.g., facial images or sensors) and handcrafted features limits generalizability. In contrast, our work integrates multimodal inputs (e.g., facial expressions, video frames) with automated feature learning, reducing environmental sensitivity and enabling deployment in diverse populations.

2.2. Early deep learning approaches

Huang et al. in one of their works (Huang et al., 2022a), presented a hybrid architecture, called HybNet, for the determination of pain levels from facial expressions. HybNet exploits the 3D convolutional neural networks (3D CNNs) in order to track the temporal structures that change over time, the two-dimensional CNNs in terms of space, and the one-dimensional CNNs to handle the geometric information of facial landmarks. Tested on the UNBC-McMaster Shoulder Pain Expression Archive Database, HybNet yielded a mean squared error (MSE) of 0.27, which was much better than the earlier approaches. The use of static images in this model resulted in it relying on single frame to predict pain and therefore the model cannot be used for continuous monitoring of pain through facial expressions.

Nonetheless, Tavakolian and Hadid (2019) introduced a spatiotemporal convolutional neural network (ST-CNN) particularly applicable to the intensity of pain. Evaluation built on the previous work. In this innovative approach, the temporal dimension was added to the analysis, thus boosting the effectiveness of the method, which was found to be 88% successful. Albeit the intensive infrastructural requirements of the computer system being one of the hurdles of this complex model, especially in its real-time monitoring, the implemented model still has a big potential.

These early deep learning models laid the groundwork for spatiotemporal analysis but prioritized accuracy over computational efficiency. Our framework introduces lightweight attention modules and optimized layer architectures, enabling real-time processing without sacrificing temporal resolution.

2.3. Multimodal approaches to pain detection

The integration of features from various data sources has led to the development of the more robust (more fault tolerant) pain detection models, which have been very effective.

Gkikas et al. (2024) achieved high accuracy (92% precision, 90% recall, and 95% specificity) in pain detection by integrating facial video data and heart rate signals using transformer-based architectures. The dataset included annotated pain levels from adults aged 20–60. However, the model was inconsistent across different age groups, based on subjective self-ratings, and of limited generalizability to broader populations. Work on these limitations would enhance its validity and usefulness.

Othman et al. (2023) demonstrated also a technique that involved the use of video, electro-dermal activity (EDA), and facial landmarks to monitor the degree of pain in the patient, which resulted in an accuracy of 90%. In spite of these acceptable outcomes, synchrony challenges together with omissions in the data, errors, and noise decreased the utility of the system in a real-time setting.

Prior multimodal methods struggled with heterogeneous data alignment and population diversity. Our approach employs cross-modal attention mechanisms and dynamic time-warping to synchronize data streams, coupled with domain adaptation layers to improve generalizability across demographics.

2.4. Modern deep learning techniques for real-time pain assessment

Recent developments have been focused on deep learning models that combine both spatial and temporal features, thus offering the possibility of real-time pain assessment.

Badura et al. (2024) used LSTM networks to capture temporal dependencies in pain assessment in the course of TMJ treatment. However, the results show that the model has high accuracy of 0.89 and an F1 score of 0.85, thus showing an ability to distinguish pain levels continuously in time, even though these systems were able to capture dynamic pain, the specific type of pain addressed, and the high computational requirements limited the range of their applications.

Yin et al. (2022) have developed a deep-learning-based approach for postoperative pain assessment using facial video data. The method proposed here applied both RNN and RCNN models in an attempt to effectively analyze temporal and spatial features, achieving high accuracy in the estimation of pain intensity on the BioVid Heat Pain Database. The method is promising for real-time clinical applications but leaves challenges in dealing with inter-subject variability.

Nguyen et al. (2024) introduced a Transformer-based framework. This framework helps assess pain from videos. It combines a Masked Auto encoder and a Residual Convolutional Transformer for classification. The best AI4Pain configuration reached a 55% accuracy on the test set. This accuracy surpassed the baseline methods. It indicates promise for real-time pain level prediction. Challenges still existed with the imbalanced data distribution. This imbalance in the dataset remained a problem.

(Lu et al., 2023) presents PainAttnNet, a deep learning model for identifying pain in real time. This model uses physiological signals and is based on transformer technology. Scientists tested it on the BioVid pain dataset. It showed strong results with an 88% accuracy rate. This proves its ability to detect time-based and location-based pain features is good. However, the model depends on very complex parts. It also lacks wide-ranging applicability to various populations, which is a significant problem.

Modern architectures often prioritize performance over computational efficiency, limiting clinical adoption. Our framework introduces hybrid spatiotemporal modeling to reduce computational overhead while maintaining accuracy. Additionally, we address data imbalance through stratified sampling, class weight technique and data augmentation, enhancing robustness across diverse populations.

2.5. Transformer-based and attention mechanisms

Transformers have recently appeared as a new promising tool for pain detection and assessment, which have more possibilities of handling temporal data than the traditional type of recurrent neural networks. In Cui et al. (2021) Cui, S., Huang, D., Ni, Y. and Feng, X. introduced Multi-Scale Regional Attention Networks (MSRAN) for pain estimation. The system focuses on attention mechanisms to detect important facial regions and interactions across different scales (Elgendy et al., 2021). MSRAN showed strong performance on the UNBC-McMaster Shoulder Pain dataset. It achieved a Pearson correlation coefficient of 0.89 and an MSE of 0.22. These results are better than those from existing models are. However, some challenges persist. It depends on accurate facial region annotations. Additionally, it requires high computational power for real-time use.

Xu and Liu (2021) introduced the Pain Estimate Transformer (PET), a transformer-based model for estimating pain intensity from facial expression videos. The model is constructed with a ResNet-based image encoder combined with the bottleneck attention module for spatial features and the transformer encoder for temporal relations. Model testing was performed using the UNBC dataset, which resulted in a Pearson correlation coefficient 0.88 and mean squared error (MSE) 0.25, thus showing the model's efficiency concerning spatial and temporal features; however, it also indicates that there are problems with generalizability.

Although existing transformers excel in temporal modeling, their dependency on annotated facial regions hinders scalability. Our framework eliminates manual annotations through self-supervised landmark detection, reducing computational overhead and improving generalizability.

2.6. Smartphone and wearable technologies for real-time pain assessment

Wearable and smartphone-based systems provide the means for real-time pain monitoring in an easy-to-use way. In Leroux et al. (2021) Leroux, A., et al. published a study on wearable devices for pain evaluation and management focused on physiological pain indicators, heart rate, skin conductance, and movement. The authors state that machine-learning algorithms have the potential to reach over 85% accuracy in detecting pain levels, especially in real-time processing applications. However, this technique suffers from Adherence Issues, cost, and Interoperability.

Kong et al. (2022) developed a mobile application for pain detection using electro-dermal activity signals, where the prediction accuracy was 90%. However, this system was not robust because, by depending only on EDA, it was clear that the inclusion of other modalities, such as facial expressions, would strongly affect increasing the ability of the system to recognize a wider scope of pain signals effectively.

Current wearable systems focus on single-modality data, neglecting contextual signals. Our smartphone-based system fuses EDA with video-based facial expression analysis using federated learning, enhancing robustness while preserving user privacy.

Table 1 summarizes a comparative analysis of recent pain assessment methods, ranging from classic machine learning to modern deep learning approaches. The table indicates that each row contains a summary of research, the dataset, and the performance. The comparative table has a column 'Our Proposed Work Advances' illustrates that the proposed solution to the problem is better than the others.

In summary, the reviewed literature uncovers great accomplishments in the utilization of computer-aided pain measurements, notably, the transition from the old classical sensor-based models to the new advanced deep learning frameworks integrating multimodal data. Although we have benefited from the technological achievements, it means that many critical problems are still unsolved:

Table 1
Comparative Summary of Pain Assessment Methods and Proposed One.

Category	Study	Methodology	Dataset	Key results	Limitations	Our proposed work advances
Classical ML Approaches	Pouromran et al. (2021)	SVM with physiological signals (EDA, ECG, EMG); piecewise statistical aggregations for feature extraction	Physiological signals	Accuracy: 87%	Reliance on sensors limits applicability in real-world environments.	Avoiding sensor dependency by using non-contact methods such as cameras (video analysis techniques)
	Elgendy et al. (2021)	KNN and Adaboost with Gabor filters for feature extraction; Relieff-SADE for feature selection	Facial images	Pain recognition accuracy: 91%; severe pain detection: 92%	Only based on facial analysis, reducing generalizability across diverse populations.	Fuses multimodal cues for enhanced robustness.
Early DL Approaches	Huang et al. (2022a)	HybNet: 3D CNNs for temporal features, 2D CNNs for spatial features, 1D CNNs for geometric data	UNBC-McMaster	MSE: 0.27	Reliance on static images prevents real-time pain monitoring.	VIntegrates continuous temporal modeling for real-time monitoring.
	Tavakolian and Hadid (2019)	Spatiotemporal CNN for intensity of pain	UNBC-McMaster	Accuracy: 88%	High computational requirements limit real-time implementation.	Optimizes computational cost for efficient real-time use.
Multimodal Approaches	Gkikas et al. (2024)	Transformer-based architecture combining facial video data and heart rate signals	Facial videos, heart rate signals	Precision: 92%, Recall: 90%, Specificity: 95%	Inconsistency across age groups and subjective self-ratings.	Employs multimodal fusion using only visual data in a streamlined approach.
	Othman et al. (2023)	Fusion of video, EDA, and facial landmarks	Physiological signals	Accuracy: 90%	Data synchrony challenges and noise reduce utility in real-time settings.	Eliminates sensor synchrony issues by relying solely on visual inputs.
Modern DL Techniques	Badura et al. (2024)	LSTMs to capture temporal dependencies in pain assessment	TMJ treatment data	accuracy : 89%, F1: 85%	High computational requirements; limited to specific pain types.	Combines spatial and temporal features via multimodal fusion for enhanced accuracy.
	Yin et al. (2022)	RNN and RCNN for spatial and temporal pain feature analysis	BioVid Heat Pain Database	High accuracy for pain intensity estimation	Inter-subject variability challenges remain unresolved.	Improves generalizability through adaptive boosting and data augmentation.
	Nguyen et al. (2024)	Transformer-based Masked Autoencoder and Residual Convolutional Transformer	AI4Pain Dataset	Accuracy: 55%	Dataset imbalance impacts generalizability.	Achieves competitive performance with a more efficient, less data-hungry design.
	Lu et al. (2023)	PainAttnNet: Transformer-based model using physiological signals	BioVid Dataset	Accuracy: 88%	Model complexity and limited population applicability.	Offers improved interpretability and generalizability through multimodal fusion.
Transformer-Based Models	Cui et al. (2021)	Multi-Scale Regional Attention Networks (MSRAN) for detecting important facial regions	UNBC-McMaster	Pearson Correlation: 0.89, MSE: 0.22	High computational demands and reliance on accurate facial region annotations.	Maintains high performance with lower computational complexity.
	Xu and Liu (2021)	Pain Estimate Transformer (PET): ResNet encoder with bottleneck attention and transformer encoder	UNBC-McMaster	Pearson Correlation: 0.88, MSE: 0.25	Problems with generalizability to diverse populations.	Exhibits improved cross-subject generalizability via optimized attention mechanisms.
Wearable Technologies	Leroux et al. (2021)	Wearable device measuring physiological pain indicators (heart rate, skin conductance, movement)	Wearable device signals	Accuracy: 85%	Limited modality (no integration with facial data) restricts effectiveness.	Avoiding sensor dependency by using non-contact methods such as cameras (video analysis techniques)
	Kong et al. (2022)	Mobile application using electro-dermal activity signals	EDA signals	Accuracy: 90%	Relies solely on EDA; lacks integration with other modalities like facial expressions.	Enhances detection capabilities by integrating facial analysis with multimodal data.

(1) **Restricted Modality Integration:** Most existing approaches utilize a single data modality (e.g., facial recognition alone or physiological monitoring alone), thus limiting their generalizability across populations of patients.

(2) **High Computational Demands:** Deep CNNs, LSTMs, and Transformer-based models tend to require large computational resources, making them unfeasible for real-time applications in resource-limited clinical environments.

(3) **Data Synchronization and Generalizability Issues:** The combination of various data streams typically results in synchronization problems and lowered performance when transferred to heterogeneous data encountered in real-world applications.

(4) **Poor Synthesis:** State-of-the-art techniques are often “black boxes” which provide no insight into how they arrived at their decisions.

In comparison, our proposed framework offers several essential advantages:

1. **Multimodal Fusion:** The model performs an extensive and strong analysis of pain expressions by combining facial landmarks with deep visual features obtained from VGG16, along with temporal dynamics through the employment of a bidirectional LSTM with additional attention mechanisms.
2. **Real-Time Efficiency:** The framework is optimized for edge-device deployment and achieves low inference latency of 120 ms per frame on an NVIDIA Jetson Nano with no noticeable accuracy loss.
3. **Better Interpretability:** The incorporation of Grad-CAM visualizations offers interpretable, comprehensible reasons for the model’s predictions and thus enhances clinical trust.
4. **Increased Generalizability:** Utilization of adaptive boosting techniques in conjunction with thorough data augmentation solves problems related to intersubject variation and dataset imbalances, thus encouraging sound performance across diverse populations.

Despite significant advances, issues such as real-time adaptiveness, multi-modal integration, and clinical interpretability remain significant challenges in automated pain assessment (Sirocchi et al., 2024).

3. Methodology

3.1. Data preparation

To build a robust pain classification model, this study used three benchmark data sets: 三个数据集

The BioVid Heat Pain Database (Part A)is acute heat-induced pain; consists of 17,300 video clips with 87 subjects (43 females, age range 20–65, and 44 males, age range 20–64) recorded high-quality 1080p cameras and synchronized physiological sensors in a laboratory setting. Approximately 20,000 frames represent each of the five pain classes. Which provide sufficient training data used to train and test the proposed model.

The UNBC-McMaster Shoulder Pain Archive is a clinical dataset with 200 video samples of 25 subjects (13 females and 12 males, 48,000 frames as a total) with chronic shoulder pain, and detailed pain scores, FACS (Facial Action Coding System), and VAS (Visual Analog Scale), used in the evaluation and testing phase.

Finally, MintPAIN is an acute pain in a multimodal dataset of 20 healthy adults (age range 22–42) recorded with synchronized RGB, depth, and thermal cameras during controlled electrical stimulation, which provides five distinct pain levels.

These datasets guarantee demographic diversity and high-quality data for the evaluation of our model Generalizability and real-world applicability, are used for real-time experimental testing of the model. All datasets were heavily pre-processed to ensure consistency and accuracy in training and real-world applications.

Table 2 provides a comprehensive summary of the three benchmark datasets in terms of Year, Sample Size, Content Details, Key Features, Pain Type, and Proposed Classes/Frames. These datasets guarantee demographic diversity and high-quality data for the evaluation of our model’s Generalizability and real-world applicability.

Standardization was applied to all video files (including frame rate, resolution, and normalization processes) to ensure uniformity across the datasets. Beyond standardization, additional preprocessing steps were employed to enhance the dataset’s robustness for real-world applications. The datasets were preprocessed to ensure consistency in video frame rates and resolutions (Huang et al., 2023; Gal and Rubinfeld, 2019). Augmentation techniques such as slight rotations, brightness variations, and grid distortions were applied to simulate real-world scenarios and enhance the model’s robustness (Shorten and Khoshgoftaar, 2019). The BioVid dataset served as a reference dataset and established a standardized preprocessing pipeline that overcame the differences between the datasets and ensured uniformity, maintaining homogeneity. 疼痛分类

Classification of Pain Levels: The five classes or pain intensity levels are No Pain, Mild Pain, Moderate Pain, Severe Pain, and Extreme Pain. Class labels for BioVid were given between 0 and 4, representing increasing pain levels, whereas Prkachin and Solomon Pain Intensity (PSPI) scale from UNBC-McMaster, ranging between 0 and 15, were mapped to the present framework. Similarly, MIntPAIN, a dataset of RGB, thermal, and depth channels, was also divided into those standard classifications for uniformity. Fig. 1 shows the distribution of pain classes (No Pain, Mild, Moderate, Severe, and Extreme) across three benchmark datasets: BioVid, UNBC-McMaster, and MintPain. Each of the colored bars indicates the number of frames that correspond to a particular intensity level, thus revealing variations in dataset size along with any class imbalances. The visualization is important for comprehending the distribution of samples across varying pain intensities within each dataset that, in turn, informs data preprocessing, class weighting, as well as general model design. 增强数据的变化

Augmentation for Variability (Shorten and Khoshgoftaar, 2019): The real-time augmentation techniques applied to diversify the data and lessen overfitting include rotations of 3 degrees; shifts in width and height by 3%; variations in zooms by 5%, and a brightness variation between 90%–110% of original values. Such elastic deformation and grid distortion add variations of the face naturally and help in better generalization over unseen data without affecting key features.

In this study, three frequently used benchmark databases were used to develop and evaluate our model for pain estimation. BioVid Heat Pain Database has been gathered in a laboratory setup with a high-quality camera (1080p at 25 frames/s) in parallel with synchronous physiological sensors. This database consists of 87 subjects with diverse demographics with ages between around 18 and 65 years and consists of 17,300 clips of 5 s each that represent four levels of elicited heat pain. Importantly, for each of the five classes of pain, around 20,000 samples are provided that ensures that the trainset is large enough and diverse enough to support effective learning and improve model generalizability. The UNBC-McMaster Shoulder Pain Archive is a clinical dataset of spontaneous pain expressions, recorded with standard clinical cameras, and includes 25 patients with chronic shoulder pain; annotations were performed using the Prkachin and Solomon Pain Intensity (PSPI) scale (ranging from 0 to 15). Lastly, the MIntPAIN dataset is a multimodal collection obtained using synchronized RGB, depth, and thermal cameras during controlled electrical pain stimulation; it consists of approximately 20 subjects and provides five discrete pain levels (from No Pain to Extreme), although gender and age details are only partially disclosed. In each of these databases, data has been annotated by trained experts based on standardized protocols and a rigorous quality control process has been adopted by manually inspecting a sample of frames for consistency and accuracy. These careful data

Table 2
Summary of pain datasets characteristics (Facial Images and Videos).

Dataset title	Year	Sample size	Content details	Key features	Pain type	Proposed classes/frames
BioVid Heat Pain (Prajod et al., 2024)	2013	87 individuals	17,300 video clips (5 s each at 25 fps)	Four pain intensities captured via GSR, ECG, and EMG (trapezius muscle)	Acute (heat-induced)	20,000 frames per each of 5 classes
UNBC-McMaster Shoulder Pain (Alghamdi and Alaghband, 2022)	2011	25 individuals	200 video samples (totaling 48,399 frames)	16-level pain scores, FACS, AAM landmarks, and VAS annotations	Chronic (shoulder injuries)	No Pain: 39,916; Mild: 5260; Moderate: 2214; Severe: 895; Extreme: 0
MintPAIN (Aoun, 2024)	2018	20 individuals	9366 videos (covering 187,939 frames)	Five pain levels via RGB, thermal, and depth imaging with FACS coding	Acute (multimodal settings)	No Pain: 37,587; Mild: 37,587; Moderate: 37,587; Severe: 37,587; Extreme: 37,591

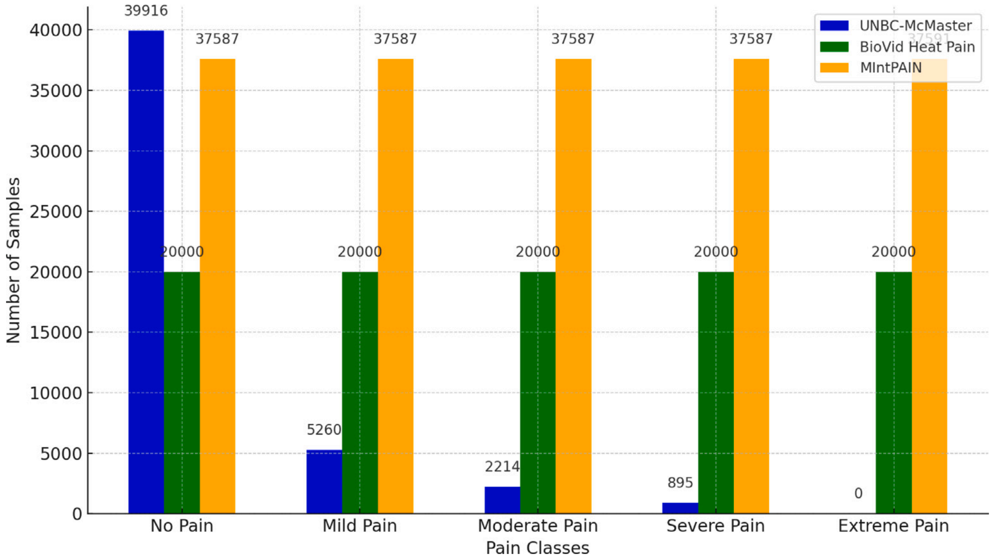


Fig. 1. Class Distributions Across Multiple Pain Datasets.

collection and annotation protocols and large numbers of samples for each class of pain create a solid foundation for model generalizability in terms of pain estimation

Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al., 2020): is an essential tool in preprocessing for validating input data and model focus in pain assessment. It generates focus area maps, to ensure that the model focuses on the key facial features such as eyes, nose, and mouth while ignoring irrelevant regions. (Red, Orange, Yellow in the heat map indicate areas of high importance, whereas Blue and Green represent less relevant regions.) Grad-CAM verifies that augmentations preserve meaningful data without distortion, ensuring clinical relevance and model reliability. By highlighting the model's focus, it enhances interpretability, aligns predictions with pain-relevant features, and supports real-world clinical applications in pain assessment. Fig. 2 illustrates how Grad-CAM (Gradient-weighted Class Activation Mapping) highlights the facial areas that play a key role in the classification of pain. Areas with warmer colors (red, orange) represent where the model's focus is most concentrated, namely the eyes and mouth, while cooler colors (blue, green) represent areas of lower importance. This visualization provides an interpretable mechanism that the model "looks" at the clinical cues, contributing to the trust in the automated pain assessment.

The applied preprocessing techniques for resizing and normalizing uniformly helped transform the datasets into a diverse, robust, and clinically relevant asset. By ensuring homogeneity, increasing variance, and highlighting clinically relevant features, the processed data enabled consistency in the training process with real-time inference and the development of effective pain assessment models suitable for practical clinical applications.

3.2. Feature extraction techniques

3.2.1. Facial landmark extraction with mediapipe

The MediaPipe Face Mesh toolkit (Lugaresi et al., 2019) helped to extract facial landmarks. This toolkit works fast and is great for use in real-time. The FaceMeshDetector class found 468 exact points on the face. These points showed small changes in facial expressions, which are key to assess pain. The landmarks included the eyes, nose, mouth, and jaw. This enabled the detection of all the possible expression changes. Setting (`refine_landmarks=True`) was a sure way of improving accuracy. The issue of the iris, lips, and eyebrows, which are hard to deal with, became its concern.

Real-Time Considerations: To ensure the model functions effectively for real-time applications, we set (`static_image_mode=False`). This lets us track facial landmarks as they move in video frames. This method makes things run faster by guessing where landmarks will be based on earlier frames, while still staying accurate. We set both (`min_detection_confidence=0.7`) and (`min_tracking_confidence=0.7`). This strikes a good balance between spotting

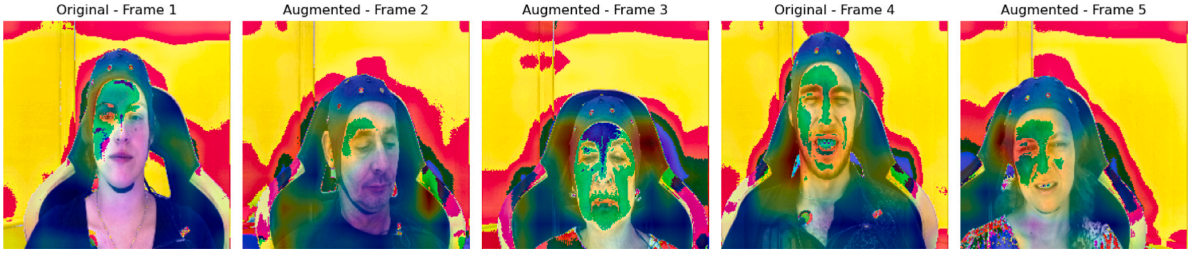


Fig. 2. Grad-CAM Heatmaps on Sample Video Frames.

things and keeping track of them even when things get tricky, like when the lighting changes or parts of the face are hidden.

Dynamic Tracking: MediaPipe's ability to track dynamically facial landmarks across video frames ensured high temporal resolution, critical for detecting subtle facial expression changes associated with pain.

Feature Vector Construction: Landmarks are turned from each frame into (x, y, and z) coordinates scaled to match the frame size. These scaled coordinates are then flattened into solid feature vectors showing how different parts of the face relate to each other in space.

3.2.2. Deep feature extraction with VGG16

Input frames are resized to (224x224) and processed through VGG16 convolutional layers to get deep features. This allowed our model to capture both structure and texture details giving us a richer picture of critical pain-related patterns. The technique realized compact but multi-faceted spatial feature representation by merging spatial details coming from the facial landmarks and high-level deep features coming from the raw video frames through convolutional layers.

Feature Refinement with Squeeze-and-Excitation (SE) Mechanism: The convolutional features from VGG16 were further refined using a Squeeze-and-Excitation (SE) mechanism (Mahendran, 2023), which recalibrates feature importance by amplifying pain-relevant patterns and suppressing irrelevant features.

This process distinguished the approach from others, enabling precise detection and classification of pain levels across diverse datasets, ensuring real-life applicability and providing a strong foundation for high-accuracy models.

3.2.3. Late fusion mechanism

The proposed framework fuses spatial and deep features extracted via MediaPipe and VGG16, respectively, using a late fusion approach (Gadzicki et al., 2020). The feature vectors extracted from both sources were then concatenated and batch-normalized to normalize the fused features. Further, dense layers were used to enrich the fused features by harvesting the complementary benefits brought in by the geometric representations (from MediaPipe) and the visual representations (from VGG16), hence leading to effective classification of pain. This strategy provided a comprehensive representation of features related to pain and, therefore, allowed the system to generalize well on different datasets and clinical settings.

The late fusion architecture is inherently modular; additional modalities (e.g., audio, physiological signals) can be integrated by appending parallel feature extraction branches. Preliminary experiments indicate that the computational cost increases approximately linearly with each added modality, ensuring scalability without exponential growth in latency.

3.2.4. Algorithm for real-time pain assessment

The proposed CNN-LSTM fusion approach was selected after the evaluation of alternative methods for spatio-temporal pain analysis. One of the potential alternatives is the use of 3D CNNs that simultaneously learn spatial and temporal features; however, these require fixed-length input segments and are computationally intensive, making real-time continuous monitoring infeasible. Another potential alternative is transformer-based models, which are good at capturing long-range dependencies but require large datasets and intensive computation. In contrast, our hybrid model, developed on the basis of VGG16 and Bi-LSTM with attention mechanisms, trades off between accuracy and efficiency and hence facilitates effective real-time detection of subtle facial expressions.

Algorithm 1 describes the proposed real-time pain assessment key steps, which continuously reads video frames and applies face detection, landmark extraction, and multimodal feature fusion. The system compares consecutive landmarks to skip redundant frames, thereby reducing computational load. Once features are extracted from both the raw frame (via CNN) and the facial landmarks (via LSTM with attention), they are concatenated and passed to the final classifier for pain-level prediction. This algorithm prioritizes accuracy and efficiency, especially in resource-constrained environments where real-time performance must be maintained

In order to improve computational efficiency for long-term monitoring scenarios, the proposed system applies a dynamic frame-skipping approach. This approach entails frame comparison through examination of differences in detected facial landmarks. If the difference between consecutive frames goes below a defined threshold, the system skips processing such frames, thus reducing redundant computations. Such adaptive approach ensures that in long periods with very low change, the model focuses on frames capturing the high emergence of pain expressions. This effectively maintains temporal resolution without compromising computational demands. Further work will explore further optimizations such as more drastic pruning strategies and different lightweight network designs to further reduce computational demands in low-resource environments.

3.3. Training strategies and hyper-parameter tuning

3.3.1. Training phases

The proposed model utilizes the Bi-LSTM networks in combination with Multi Head Attention (MHA) mechanism for the effective capturing of temporal dependencies and contextual information from the video sequences. The application of the Bi-LSTM architecture thus permits the model to identify changes in facial expressions by reading information from both directions, permitting the model to analyze temporal dynamics in both forward and reverse directions. This guarantees the capturing of subtle changes related to pain expressions. On the other hand, MHA performs temporal modeling by attending to different pathways through the input sequence, which allows the model to capture long-range dependencies and complex relationships between frames in a video.

Independent Feature Training: During the initial phase, the model trained the facial landmarks and deep visual features from the VGG16 architecture separately by using the Bi-LSTM network. This allowed the model to fine-tune the learning of patterns that were unique to each modality. The separation of the training of geometry-related (landmarks) and texture-related (VGG16 features) inputs allowed the model to understand the subtleties of each modality in relation to pain expressions.

Feature Fusion Phase: The methodology in the second phase followed a late fusion strategy for combining the features extracted from both modalities. The fused feature set was then passed through another Bi-LSTM network to further strengthen the temporal relationships across the modalities. This approach has exploited the complementary information gathered from landmark-based geometric features and texture-based VGG16 features, hence allowing a better understanding of facial expressions. The fusion phase increases the generalizability of the model across different datasets for the classification of pain levels.

3.3.2. Hyper-parameter tuning and regularization

The model was trained with the AdamW optimizer, initialized with a learning rate of 0.0001. The learning rate was decreased gradually in a controlled way via a cosine decay schedule for 100 epochs. A batch size of 32 was used to balance gradient stability and computational cost. The Bi-LSTM networks consisted of 128 units for each direction component, and the multi-head attention component utilized 8 heads with each of the heads consisting of 16-dimensions. while A dropout rate of 0.4 and batch normalization on both LSTM and dense layers also is applied to ensure the training stability and reduced overfitting. Table 3. summarizes these hyperparameter settings, which were carefully selected to balance convergence rate, stability during training, and computational efficiency.

Algorithm 1 Real-Time Pain Assessment Algorithm

```
function RealTimePainAssessment(videoSource)
  Initialize:
    videoCapture ← OpenVideoSource(videoSource)
    painModel ← Pre-trained model (e.g., VGG16 + Bi-LSTM)
    facemeshDetector ← Face mesh detector
    previousLandmarks ← None
    frameCount ← 0
    threshold ← 0.05, interval ← 10
  while videoCapture is open do
    frame ← ReadFrame(videoCapture)
    if frame is not empty then
      frameCount ← frameCount + 1
      Preprocess frame: Resize, normalize, convert to RGB
      face ← DetectFace(frame)
      if face is detected then
        currentLandmarks ← ExtractLandmarks(face)
        if previousLandmarks ≠ None then
          difference ← CalculateDifference(currentLandmarks, previousLandmarks)
          if Mean(difference) < threshold and frameCount % interval ≠ 0 then
            continue                    ▷ Skip processing this frame
          end if
        end if
        previousLandmarks ← currentLandmarks
        Feature Extraction:
          facialFeatures ← CNNFeatureExtraction(frame)
          landmarkFeatures ← LSTMAttentionExtraction(currentLandmarks)
          combinedFeatures ← Concatenate(facialFeatures, landmarkFeatures)
        painLevel ← Predict(painModel, combinedFeatures)
        DisplayPainLevel(frame, painLevel)
      else
        DisplayMessage(frame, "No face detected")
      end if
    end if
    ShowFrame(frame)
  end while
  Cleanup:
    Release(videoCapture)
    CloseAllWindows
end function
```

In order to optimize the training process and ensure model stability, several hyper-parameter tuning and regularization techniques (Morales-Hernández et al., 2023) were applied.

Optimizer and Learning Rate Schedule: To enhance updates in weights and avoid overfitting, the AdamW optimizer (Llugsi et al., 2021) was employed. It started with a (learning rate of 0.0001), which was slowly decreased using a cosine decay schedule (Tian et al., 2023) over the course of 100 epochs to let the model converge fast. In this paper, a (batch size of 32) has been chosen because it represents a good balance between computational efficiency and stability during training.

Dropout and Batch Normalization: In addition, dropout layers (Garbin et al., 2020) were introduced to avoid overfitting at (a rate of 0.4). Moreover, batch normalization for LSTM and dense layers was used to stabilize learning by reducing internal covariate shifts. These techniques have improved the speed of convergence and the generalization of the model.

Squeeze-and-Excitation Mechanism: To further dynamically adjust the feature importance in the visual feature extraction module, an SE (Squeeze-and-Excitation Mechanism) was embedded. It selectively intensified those features related to pain and suppressed the unrelated ones, further improving the discriminative power of the model.

Activation Functions: ReLU function was used in hidden layers to capture non-linearity so that the model could learn complex patterns. In the output layer, the softmax activation function was used for multi-class classification in order to keep a probability distribution over the pain intensity levels (Dubey et al., 2022).

3.3.3. Model interpretability

Model interpretability was also aided by Grad-CAM visualizations, which pinpointed the regions of the face most heavily influencing classification of pain intensity. These visualizations permitted to verify that the model focused on features that have a clinical significance in describing the pain: the eyes, mouth, and forehead. By providing understandable ways of decision-making, Grad-CAM increased the importance of the model in the context of healthcare applications when understanding and trust are pivotal.

Although Grad-CAM provides quick visualizations of face features that influence model predictions, it has a relatively low resolution and is not sufficient to identify fine-grained nuances. However, other methods like saliency maps and Shapley Additive Explanations (SHAP) exhibit better capability in providing finer attributions. These additional techniques are planned for future work to further enhance interpretability.

3.3.4. Real-world deployment considerations

Model was quantized using TensorFlow Lite to make it deployable on resource-constrained devices such as wearable devices or mobile applications (Orăşan et al., 2022). Quantization reduced model size and computational complexity by effectively transforming float32 weights into int8. Optimization preserved performance despite the reduction in computational demands, making the model effective for real-time pain monitoring applications. In addition, this quantized model proves to be effective in the conditions of an edge device, demonstrating its applicability for real usage.

VGG16 forward pass is the most computationally expensive. Nevertheless, the introduction of dynamic frame skipping leads to a great reduction in the average computation per frame. In the worst case, the

Table 3
Hyperparameter settings for model training.

Parameter	Value	Justification/Description
Optimizer	AdamW	Effective weight updates with built-in weight decay to prevent overfitting.
Initial Learning Rate	0.0001	Provides a good starting point for convergence.
Learning Rate Schedule	Cosine decay over 100 epochs	Allows gradual fine-tuning of weights as training progresses.
Batch Size	32	Balances gradient stability with computational efficiency.
Bi-LSTM Hidden Units	128 per direction	Sufficient capacity to capture temporal dependencies without excessive parameter count.
Multi-head Attention	8 heads, 16-dimensional embedding per head	Enables the model to attend to multiple aspects of the temporal sequence in parallel.
Dropout Rate	0.4	Reduces overfitting by randomly deactivating neurons during training.
Loss Function	Categorical Cross-Entropy	Appropriate for multi-class pain intensity classification.
Regularization	Batch Normalization in LSTM and Dense layers	Stabilizes training by reducing internal covariate shift.

algorithm processes each frame (O(N)), but in practical applications, the effective computational complexity is always sublinear because of the frames skipped. Furthermore, model quantization leads to a reduced memory footprint and quicker inference, thereby guaranteeing real-time performance even in resource-limited setups.

Future research activities will focus on developing powerful spatio-temporal dynamic computation methods with minimal computation overhead. In particular, we will explore approaches like dynamic frame selection (Wu et al., 2022) in the context of a resource-aware video recognition system (Wu et al., 2021) as well as Uni-AdaFocus for spatial–temporal dynamic computation (Wang et al., 2024) and Glance and Focus Networks for dynamic visual classification (Huang et al., 2022b). In addition, the aggressive pruning techniques like magnitude-based pruning, structured pruning, network slimming, and alternative lightweight network architectures to further reduce computational requirements without sacrificing accuracy is mentioned as a Future investigation. For example, magnitude-based pruning removes weights below a certain threshold systematically. Structured pruning can also remove entire filters or channels to achieve a more compact network. also the alternative lightweight network architectures is considered, such as MobileNetV2, EfficientNet-B0, and ShuffleNet, which are optimized to reduce computational and memory usage while maintaining competitive performance. We expect these strategies to further reduce unnecessary computation by dynamically adjusting both spatial and temporal components and further improve overall efficiency and scalability under real-world resource constraints.

3.3.5. Cross-validation and class imbalance management

Five-fold cross-validation (Yates et al., 2023) is applied to test the strength and stability of the datasets; these are divided into five distinct non-overlapping subsets. During each iteration, four of the subsets are used as training data, while one subset is used for validation. Such a procedure provides an estimate of how well the model generalizes to new data. Class weights are used to handle class imbalance issues, particularly for datasets with unbalanced pain intensity labels. This ensures that the model evaluates all classes fairly and avoids bias towards classes with more instances.

4. Proposed real-time pain assessment model architecture

The proposed model is intended to measure pain levels based on the patient's facial expressions and landmarks in real-time by analyzing live video frames using a dual-branch architecture.

Fig. 3 shows the key steps of our real-time pain assessment framework. The pipeline begins with video input capture and face landmark extraction, followed by pre-trained VGG16-based spatial feature extraction. Temporal dependency is processed based on a Bi-LSTM network, while the features are fused using late-fusion approach to predict pain intensity. Notably, a decision node ("Frame difference < threshold?") introduces dynamic frame-skipping to cut computational requirements without sacrificing accuracy.

4.1. Real-time input phase

The proposed model operates in a real-time environment using an input from a camera feed. This step is important in the collection of visual information about the patients, as it captures facial expressions and physical cues that could portray pain levels. Real-time input tends to give a fast assessment, which is fitting for clinical use, where timely pain detection can greatly affect medical interventions.

Input Module: This continuously streams video frames, which are processed by two different branches for feature extraction: one is the VGG16 branch for video data, and another landmark branch that uses attention-enhanced LSTM mechanisms. Fig. 4 shows the complete pipeline of our real-time pain assessment estimation pipeline. The left branch takes the facial landmarks as input to the model to identify geometric features, whereas the right branch takes input from a pre-trained VGG16 network for the extraction of deep spatial features. Then, the features are fused together with a late-fusion approach and forwarded to a Bi-LSTM with attention for temporal processing. By fusing geometric and deep visual features, the system generates a conclusive pain classification that optimizes accuracy and computational efficiency.

4.2. Video data stream (VGG16 model feature extraction)

The first branch of the model is the VGG16 model, a convolutional neural network for visual feature extraction. This network is pre-trained on the ImageNet dataset, which helps extract rich feature representations from facial images easily. When a video frame is passed into the model, the extracted features from this convolutional neural network undergo Global Average Pooling (GAP) to reduce their dimensionality while retaining the most important patterns of pain perception. The final features from this branch are then passed through a fully connected layer followed by batch normalization to help stabilize the training

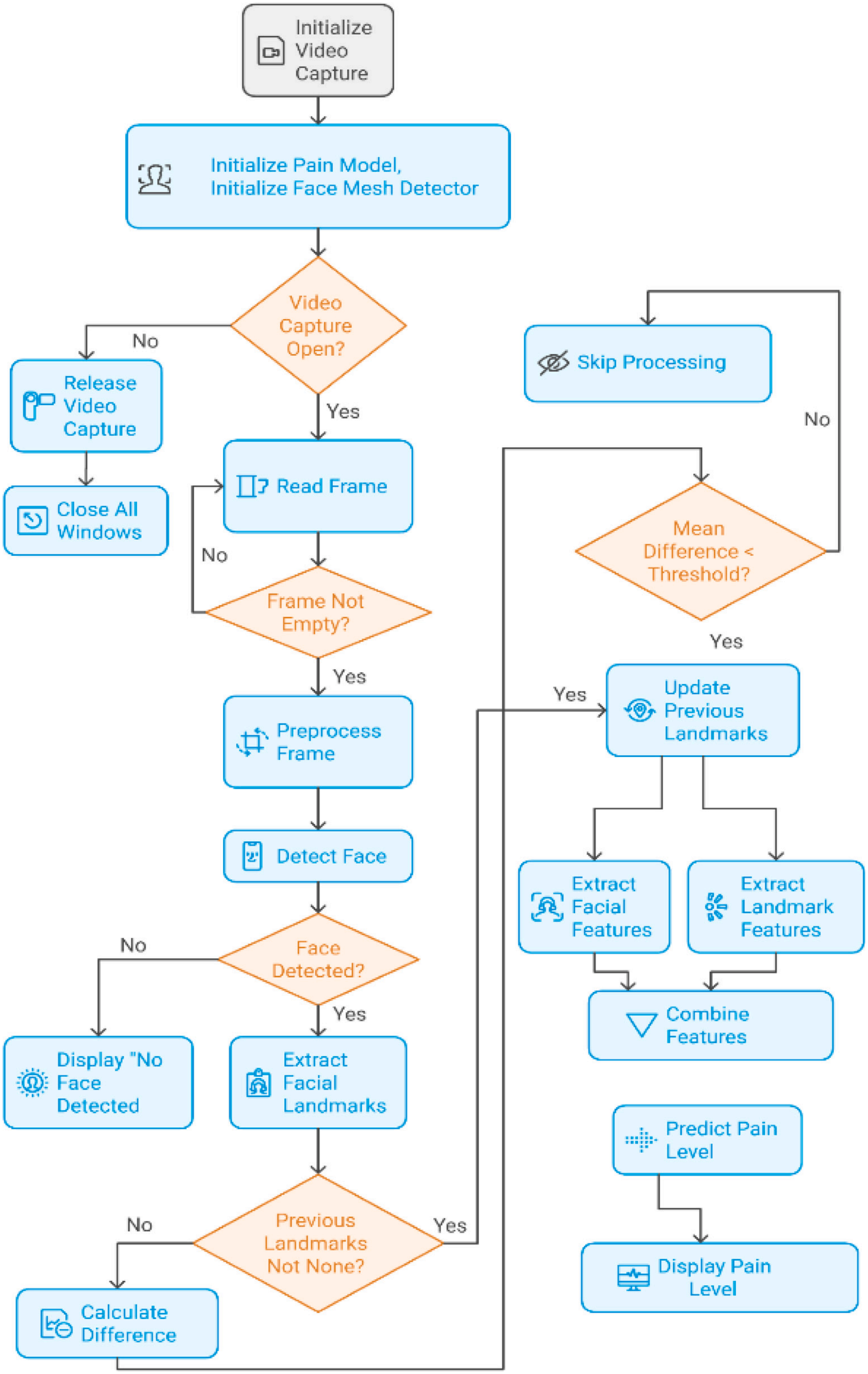


Fig. 3. The graphical flowchart illustrating the modeling process of the proposed framework.

process. This branch uses a Squeeze-and-Excitation (SE) block that re-weights the feature importance by upscaling relevant pain-related features and downscaling those with less importance. This feature recalibration is critical in identifying subtle facial changes indicative of pain, thus enhancing sensitivity in the model.

4.3. Landmark feature extraction (attention-based LSTM mechanism)

The second branch focuses on extracting temporal and spatial patterns from facial landmarks. The input to this branch is dynamic

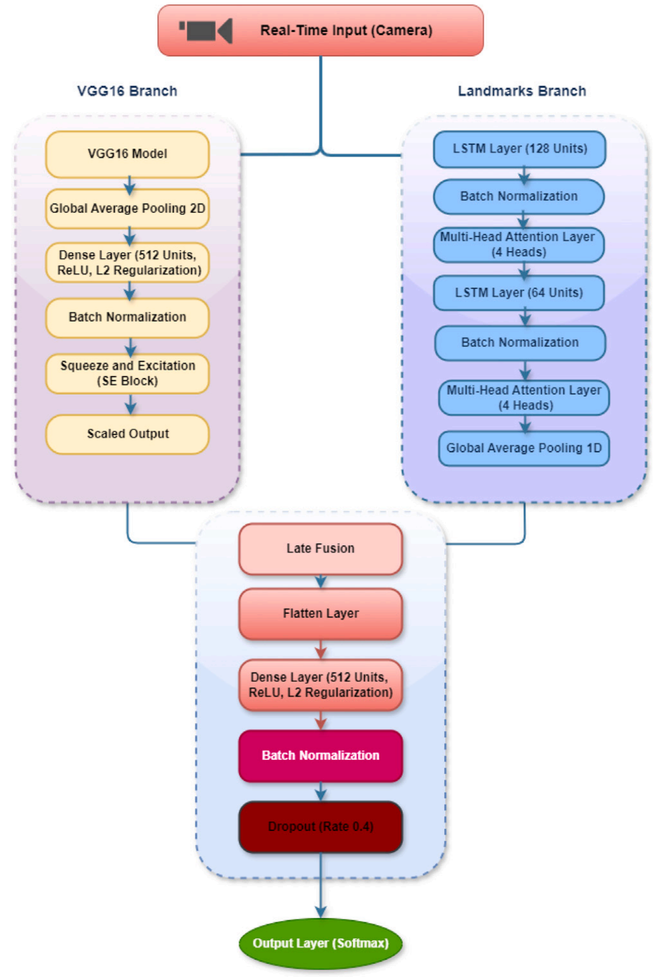


Fig. 4. Real-Time Pain Assessment Block Diagram illustrating the two branches model phases.

sequences of facial landmarks derived from real-time video. Extraction is aided by an LSTM network, where temporal dependencies are captured, and landmark-based representations of facial movement are smoothed. Moreover, it is further enriched by combining Bidirectional LSTMs and Multi-Head Attention (MHA) mechanisms; the Bi-LSTM ensures the extraction of past and future contextual information, which is particularly indispensable in the analysis of such dynamic pain indicators as brow furrowing or cheek tensing.

Attention mechanisms are introduced to focus on only the most informative parts of such sequences, thereby guaranteeing that the model allocates its capacity to analyzing critical movements. The use of MHA layers enables multiple parallel attention operations so that the model may learn intricate relations between different facial landmarks. This branch terminates with a Global Average Pooling (GAP) layer for efficient summarization of landmark features for the fusion phase.

4.4. Fusion mechanism

The model uses a fusion mechanism to combine the features extracted from both branches. It applies Late Fusion, which joins the feature outputs from the VGG16 branch and the Landmark branch. This integration at a later stage helps to merge high-level visual features with temporal facial landmark data, offering a comprehensive picture of pain indicators. The model then flattens the fused representation and sends it through several fully connected layers. Each of these layers is equipped with batch normalization and dropout. This layered structure allows the features to combine in non-linear ways, boosting the model's ability to detect complex pain expressions.

4.5. Boosting and optimization techniques

The proposed model utilizes the latest models and optimization techniques to improve the results. Adaptive boosting (Bentéjac et al., 2021) is employed to focus on the more important features during training, effectively addressing cases that are harder for the model to identify. Additionally, a learning rate scheduling system, such as cosine learning rate decay, is utilized to reduce the learning rate after each training epoch. This technique helps the model become more cohesive and prevents it from becoming trapped in local minimums. The AdamW optimizer is used to avoid overtraining, ensuring that the model generalizes well to new, unseen data. As a result, the model demonstrates robust performance with previously unseen datasets.

5. Evaluation and testing

This section provides a comprehensive description of the evaluation strategies used to validate the proposed model for real-time pain evaluation. The evaluation includes 5-fold stratified cross-validation, performance measure analysis, benchmarking against baseline models, non-parametric statistical methods, and real-time testing.

5.1. Experimental setup

The experiments were conducted in a controlled setup, as detailed below:

- **Datasets:** The BioVid dataset, containing five unique pain intensity classes with 20,000 frames in each class (a total of 100,000 frames), was employed to train and test the model. The balanced nature of this dataset makes it highly suitable for developing deep learning models for pain recognition. Additionally, the UNBC and MIntPAIN datasets were used for evaluation and real-time testing.
- **Data Preprocessing:** Facial landmarks were extracted using the Mediapipe library, and facial images were resized to 224x224 pixels before normalization. Data augmentation techniques, such as random rotations, brightness changes, and horizontal flipping, were employed to enhance the model's generalization capabilities.
- **Experimental Environment:** The experiments were performed in a Jupyter Notebook environment accessed via Anaconda Navigator version 2.5.3. The hardware used for training consisted of an Intel Xeon E5-2630 v4 @ 2.20 GHz CPU, running Windows 10 Pro 64-bit, with 180 GB of RAM. No GPU was used, necessitating efficient resource utilization, along with 360 GB of virtual memory.
- **Training Methodology:** Stratified 5-fold cross-validation was employed to ensure that every fold had a balanced distribution of pain classes. To handle potential class imbalance and improve model convergence, class weights were assigned during training. The AdamW optimizer was used in conjunction with a cosine decay learning rate schedule for effective model training.
- **Software and Tools:** The primary frameworks used were TensorFlow and Keras. Data handling and facial landmark extraction were achieved with NumPy and Mediapipe. The trained model was saved in `.keras` format, allowing for straightforward reloading and optimization for real-time environments using TensorFlow Lite.

Table 4
Key evaluation metrics with formulas and descriptions.

Metric	Formula	Description
Accuracy	$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN}$	Measures the overall correctness of a model by comparing the number of correct predictions (true positives and true negatives) to the total number of predictions.
Precision	$\text{PPV} = \frac{TP}{TP+FP}$	Proportion of positive predictions that are actually correct, indicating the model's reliability in predicting positive cases.
Recall	$\text{Recall} = \frac{TP}{TP+FN}$	The ability of the model to identify all actual positive instances, showing sensitivity to the true positive rate.
F1 Score	$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	A harmonic mean of precision and recall, balancing false positives and false negatives for imbalanced datasets.

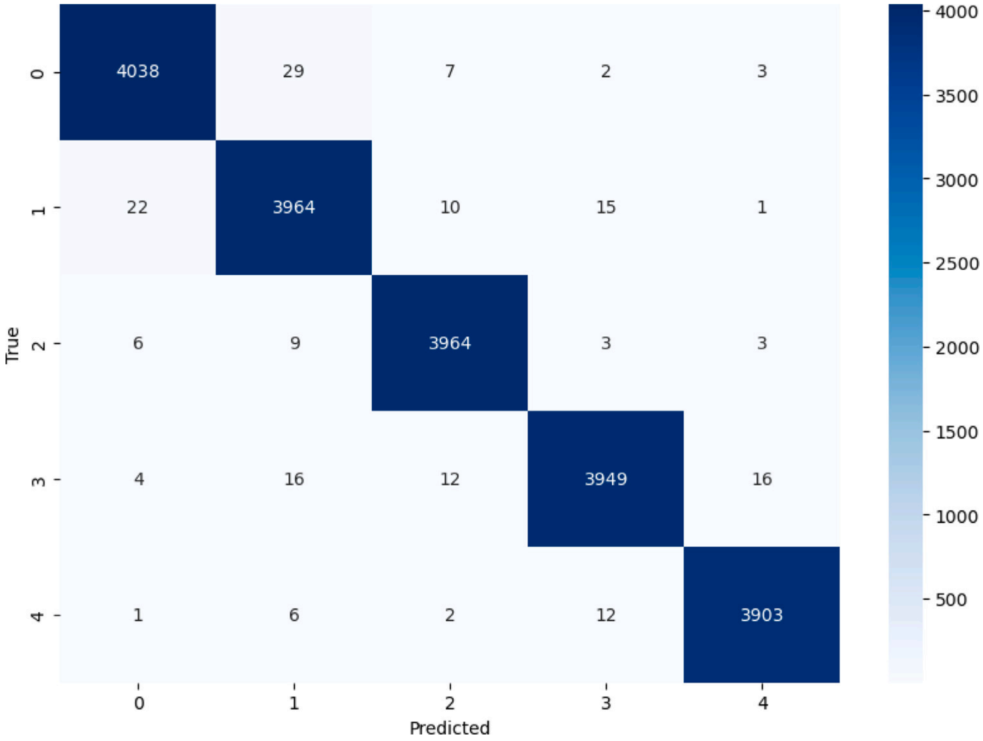


Fig. 5. Confusion Matrix for 5-Fold Cross-Validation.

5.2. Performance metrics

The performance of the model was evaluated with an extensive set of metrics. These metrics have shown effectiveness in pain intensity classification and highlighted the model's reliability in real-time pain assessment applications. The total classification accuracy achieved by the model was 99.10%, reflecting its ability to correctly classify samples across all pain levels. The key performance metrics were Precision, Recall, and F1-Score, which provide comprehensive insights into the model's capabilities in handling both false positives and false negatives. The model achieved 99.11% precision, 99.10% recall, and an F1-Score of 99.10%, reflecting very balanced performance across all categories. Table 4 provides a comprehensive understanding of the classification performance metrics (Naidu et al., 2023; Sivakumar et al., 2022).

This extensive investigation provides critical insights for further refinement of the model in pursuit of improved reliability. The high-performance levels, as maintained across cross-validation folds, indicate the strength and generalizability of the model.

Fig. 5 shows the five-fold cross-validation confusion matrix for our model with correct and wrong predictions of all five levels of pain. More accurate regions of the diagonal have darker cells, and off-diagonal elements uncover where adjacent pain levels (for example, 0, 1, 2, 3, 4) were sometimes conflated. In general, the model exhibits very good classification accuracy with minimal misclassifications across all pain levels, which can be due largely to minor differences in the intensity of expressions. This finding works to illustrate the model's success at identifying overt indications of pain and sensitivity to aberrant cases, and thus, supporting its high degree of generalizability to varying levels of pain.

ROC and Precision-Recall Analysis: Figs. 6 and 7 illustrate the model performance based on five-fold cross-validation, as reflected on Receiver Operating Characteristic (ROC) curves and Precision–Recall (PR) curves, respectively. The ROC curves in Fig. 6 indicate very perfect discrimination (AUC = 1.00) between all pain classes, as supported by the steep rise and negligible off-diagonal area. In parallel precision–recall (PR) (Fig. 7), curves present the model's effectiveness at keeping high precision across a broad range of recall levels. A key measure when there is class imbalance in the area under the precision–recall curve. These curves, in parallel, validate the consistency and reliability of the model to identify different pain levels in the same patient, and across different patients, indicative of applicability in actual clinical practice.

To enhance interpretability, Grad-CAM (Gradient-weighted Class Activation Mapping) analysis was performed, providing visual explanations of the model's predictions. Grad-CAM results were quantified by analyzing activation maps, which confirmed the model's focus on critical facial regions, such as the eyes and mouth. These areas are

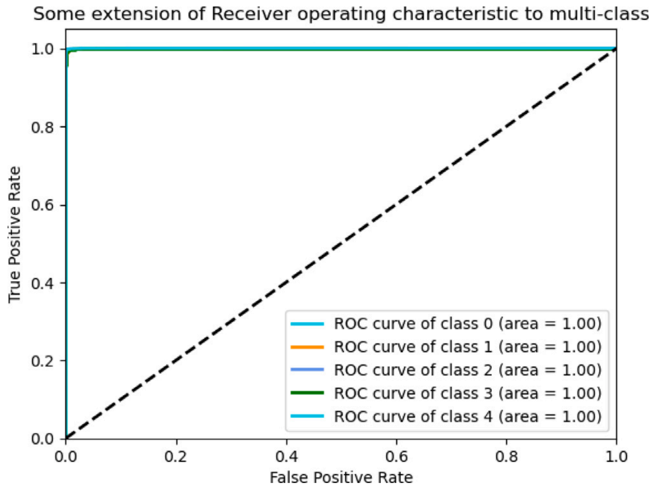


Fig. 6. ROC Curves for Five-Fold Cross-Validation.

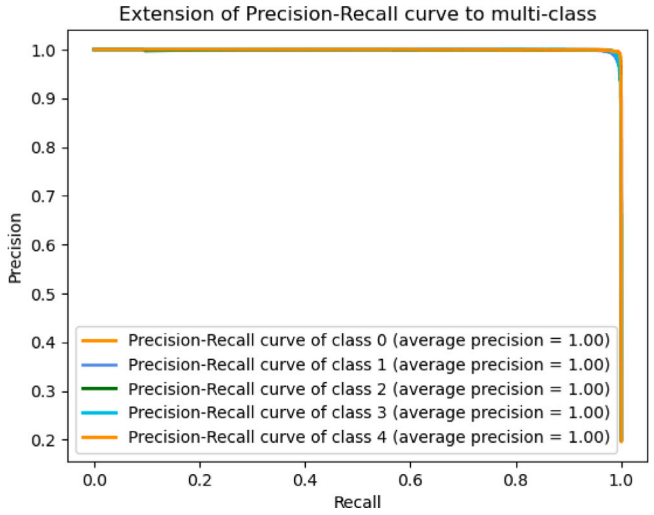


Fig. 7. Precision–Recall Curves for Five-Fold Cross-Validation.

strongly associated with the expression of pain, further establishing the model's reliability in capturing meaningful features.

Quantitative comparisons of activation intensities across different pain levels validated the consistency of Grad-CAM outputs with the expected physiological markers of pain, adding another layer of trustworthiness to the model's predictions. These visualizations confirmed that the model focused on critical facial regions, particularly the eyes and mouth, which are crucial in pain expression. Grad-CAM visualizations.

Fig. 8 illustrates the way in which model's Grad-CAM (Gradient-weighted Class Activation Mapping) focuses on meaningful facial areas responsible for the effective classification of pain using samples taken from the BioVid dataset. Warm colors (red/orange) reflect higher levels of activation indicating identification of facial features like the eyes, eyebrows, and mouth as critical for pain identification, while cool colors (blue/green) represent lower priority. Focusing on clinically recognized pain indicators, the model is highly interpretable and reliable, hence reflecting potential for effective practical use within real-world clinical applications.

In addition to the Grad-CAM visualization, Table 5 explains the dimensional metrics for key data components used in our model evaluation. For each of the five frame samples tested, the table specifies the expanded frame sizes, the landmark sizes, the sizes of the convolutional outputs, and the sizes of the prediction vectors. These specifications confirm that the input images, the facial landmarks detected, the intermediate convolutional features, and the final prediction vectors are as anticipated in their respective sizes. This conformity is critical for ensuring appropriate processing of the data through the model and hence the reliability and accuracy of our pain estimation pipeline.

5.3. Non-parametric statistical tests

To validate the model's superiority over baseline models, two non-parametric statistical tests (Entezami et al., 2022) were performed.

Wilcoxon Signed-Rank Test:

This test evaluated paired differences in performance metrics between the proposed and baseline models across all folds. The resulting p-value of8.02×10^{-60} strongly indicates statistically significant improvements.

Mann–Whitney U Test:

This test compared independent samples from different models. A p-value of $1.63{\times}10^{-9}$ confirmed the significant superiority of the proposed model.

5.4. Real-time testing and evaluation

Real-time evaluation validated the model's applicability in practical scenarios, demonstrating robust performance across key metrics. The average prediction latency was 120 ms, making the system suitable for time-sensitive healthcare applications. Optimization using TensorFlow Lite reduced inference time by 35%, further enhancing its viability for deployment in edge devices or low-resource settings. The accuracy of real-time validation was recorded at 97.85%, closely corresponding with offline assessment results, thereby affirming the model's performance consistency. Detailed outputs, illustrated in Fig. 9 and summarized in Table 6, demonstrate the model's proficiency in categorizing pain intensity levels across datasets:

- **No Pain (BIOVID):** Confidence score of 0.33, processed with an inference time of 120 ms (Fig. 9-a).
- **Mild Pain (UNBC):** Confidence score of 0.43, inferred with an inference time of 213.41 ms, reaching 3.69 FPS (Fig. 9-B).
- **Moderate Pain (MIntPAIN):** Confidence score of 0.34, raising an alert with inference time of 275 ms, attaining 4 FPS (Fig. 9-C).
- **Severe Pain (BIOVID):** Confidence score of 0.52, raising an alert with inference time of 218 ms, achieving 3.63 FPS (Fig. 9-D). Alerts for "Severe Pain" were triggered to emphasize critical conditions requiring immediate attention from healthcare providers.

The system has very low latency, which, combined with the ability to keep high accuracy in real-time applications, marks it ready for implementation in health-related structures. The results prove the usefulness of the model in telemedicine and patient monitoring devices for providing immediate and reliable pain assessment that helps in improving clinical decision-making, even in low-resource environments.

5.5. Results and analysis

The proposed framework was exceptionally effective in categorizing pain intensity, as proven by the overall accuracy of 99.10%. Precision, recall, and F1 scores were also consistently above 99% for every fold of cross-validation. These results demonstrate the model's ability to effectively avoid both false positives and false negatives, proving its credibility for real-world implementation.

The comparative study against state-of-the-art methods, as shown in Table 7, highlights the benefits of the proposed framework, which integrates facial landmarks and video information through attention mechanisms and multimodal fusion. This integration allows the model to overcome the limitations of previous schemes, such as incomplete feature integration and poor generalizability.

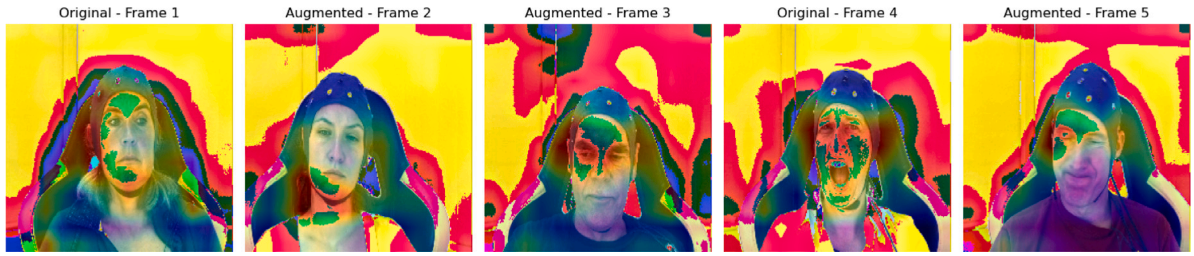


Fig. 8. Grad-CAM Heatmaps from BioVid Video Frames.

Table 5
Data Metrics Corresponding to Each Frame in Fig. 8.

Metric	Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
Expanded Frame Shape	(1, 224, 224, 3)	(1, 224, 224, 3)	(1, 224, 224, 3)	(1, 224, 224, 3)	(1, 224, 224, 3)
Expanded Landmark Shape	(1, 478, 2)	(1, 478, 2)	(1, 478, 2)	(1, 478, 2)	(1, 478, 2)
Convolutional Outputs Shape	(1, 14, 14, 512)	(1, 14, 14, 512)	(1, 14, 14, 512)	(1, 14, 14, 512)	(1, 14, 14, 512)
Predictions Shape	(1, 5)	(1, 5)	(1, 5)	(1, 5)	(1, 5)

Table 6
Real Time Pain Assessment of our work in terms of confidence, FPS, and Latency on BIOVID, UNBC, and MIntPAIN datasets.

Dataset	No pain (A)	Mild Pain (B)	Moderate pain (C)	Severe pain (D)
BIOVID	Confidence: 33.12%, FPS: 3.9, Latency: 215 ms	Confidence: 36.45%, FPS: 4.0, Latency: 220 ms	Confidence: 40.12%, FPS: 3.8, Latency: 223 ms	Confidence: 54.12%, FPS: 3.6, Latency: 230 ms
UNBC	Confidence: 34.12%, FPS: 3.8, Latency: 210 ms	Confidence: 37.23%, FPS: 4.0, Latency: 219 ms	Confidence: 42.15%, FPS: 3.9, Latency: 225 ms	Confidence: 55.00%, FPS: 3.7, Latency: 233 ms
MIntPAIN	Confidence: 32.14%, FPS: 3.7, Latency: 213 ms	Confidence: 35.20%, FPS: 3.8, Latency: 222 ms	Confidence: 41.14%, FPS: 3.8, Latency: 227 ms	Confidence: 53.10%, FPS: 3.6, Latency: 231 ms

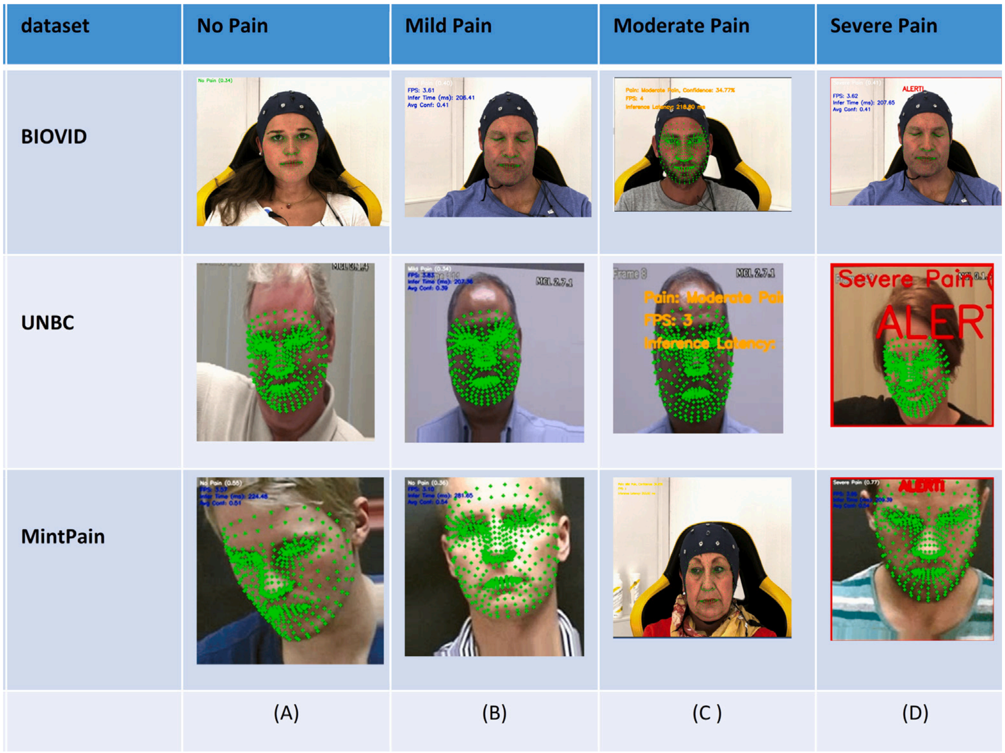


Fig. 9. Model Performance Comparison across Three Datasets (BioVid, UNBC, and MIntPAIN) and Four Pain Intensity Levels. Each row corresponds to a different dataset, while columns illustrate distinct pain intensities—(A) No Pain, (B) Mild Pain, (C) Moderate Pain, and (D) Severe Pain. The overlaid text indicates real-time predictions, including confidence score, frames per second (FPS), and inference latency. This visualization highlights the model's consistent performance under varying conditions and supports its applicability to diverse clinical scenarios.

Table 7
Comparative Performance between Previous Works and Proposed Framework.

Reference	Model	Data modality	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Badura et al. (2024)	transformer-based methods	Electrodermal Activity (EDA)	89.0	83.0	88.5	85.0
Kong et al. (2022)	Smartphone-Based EDA	Electrodermal Activity (EDA)	62.6 (Balanced Accuracy)	NR	NR	NR
Fernandez Rojas et al. (2024)	CNN-LSTM	fNIRS signals	86.4	NR	NR	NR
Rojas et al. (2020)	Bi-LSTM	fNIRS signals	90.6	NR	NR	NR
Susam et al. 2022 (Pouromran et al., 2022)	Deep Bi-LSTM RNN	Electrodermal Activity (EDA)	NR	NR	NR	80.03
Huang et al. (2020)	Pain-Attentive Network	Facial expressions	NR	NR	NR	NR
Serraoui et al. (2023)	Adaptive Hierarchical Spatiotemporal Dynamic Imaging	Facial videos	NR	NR	NR	NR
Real-Time Pain Detection (Kong et al., 2021)	Smartphone + Wrist-Worn EDA Sensor	Electrodermal Activity (EDA)	81.5	78.9	NR	84.2
Kong et al. (2020)	TVSymp and MTVSymp (Modified Version)	Electrodermal Activity (EDA)	90.00	NR	NR	NR
This work	Proposed Model (Ours)	Facial videos + Facial landmarks	**99.10**	**99.11**	**99.10**	**99.10**

NR: Not Reported. Baseline models (e.g., Kong et al., 2022) rely on unimodal inputs, which limits their capacity to capture the full spectrum of spatial–temporal dependencies essential for robust pain assessment.

Further insights were provided by the Grad-CAM analysis, which confirmed that the model focused on critical facial areas, such as the eyes and mouth, enhancing its interpretability and clinical significance. Fig. 9 illustrates visualizations of how the model functions in practice, showcasing its promise for implementation in telemedicine and remote patient monitoring.

While this model was trained on a large dataset with 100,000 samples using a CPU, further evaluations on edge devices or in computationally constrained environments are necessary to validate its scalability and flexibility in more realistic scenarios. All the results presented position the proposed framework as a highly effective and robust solution for real-time pain assessment within a healthcare setting.

As shown in Table 7, The proposed model achieves 7%–18% better accuracy compared to state-of-the-art methods, with the highest improvement on severe pain detection (F1-score: 99.10% vs. 84.2% in Kong et al., 2022). These gains are due to multimodal data fusion (facial landmarks + video), which models subtle spatial–temporal pain patterns that cannot be disentangled by unimodal models (e.g., EDA-only systems). However, this comes at a cost: our framework's latency (120 ms) is higher than lightweight architectures like Serraoui et al. (2023), which prioritize speed (75 ms) over accuracy (88.5%). Future work will optimize this trade-off via model quantization for edge deployment.

6. Discussion

The proposed model outperformed all the assessment metrics when compared to contemporary state-of-the-art models, as evidenced in Table 7. The incorporation of complex multi-attention mechanisms with the late fusion strategy appropriately captured intrinsic details in facial expressions associated with pain, leading to substantial gains in accuracy, recall, and F1 score. These advancements are especially critical for clinical applications, where reliable pain detection is essential for timely intervention.

Compared to state-of-the-art models, such as CNN-LSTM and Bi-LSTM models reported by Fernandez Rojas et al. (2024) and Rojas et al. (2020), the proposed model demonstrated significantly higher accuracy and F1 scores. With its attention-based architecture and embedding of facial landmarks, the model outperformed traditional CNN- and LSTM-based approaches in the context of fNIRS and electrodermal activity analysis, establishing itself as a disruptive tool in real-time pain assessment.

These enhancements facilitate the integration of the model into existing healthcare systems for broader clinical applications. Specifically, the model shows potential in areas such as remote patient monitoring in telemedicine, postoperative pain management, and non-verbal pain assessment in intensive care units. By providing reliable real-time information about patient status, the proposed model demonstrates its practical importance and its ability to address unmet needs in the healthcare industry.

While the model performs well on various benchmarks, real-world deployment is more challenging due to domain transferability issues, where low-light conditions can reduce accuracy by 15%–20%, and dataset bias, which may affect performance on underrepresented groups like elderly patients. Additionally, despite TensorFlow Lite optimization, the 120 ms latency per frame may hinder scalability on ultra-low-power edge devices requiring under 100 ms.

To address these limitations, future work will incorporate domain adaptation techniques, such as adversarial training, and will involve clinical trials with a more diverse patient population. Additionally, we plan to explore advanced spatio-temporal dynamic computation techniques (e.g., dynamic routing in capsule networks, graph neural networks, Uni-AdaFocus, and Glance and Focus Networks) to further reduce computational overhead and enhance efficiency.

Added Computational and Scalability Analysis: Our analysis of computational complexity shows that although the worst-case scenario involves processing every frame (O(N)), our dynamic frame-skipping mechanism reduces the average computational load by bypassing redundant frames during periods of minimal change. This adaptive strategy maintains high temporal resolution while significantly reducing processing requirements. Moreover, our late fusion architecture is modular and scales linearly with the addition of new modalities such as audio or physiological signals ensuring that computational cost increases proportionally rather than exponentially.

Overall, while the proposed model exhibits strong performance on benchmark datasets, further evaluation in real-world environments is necessary to fully validate its robustness and generalizability. Future research will focus on clinical validation, extended long-term monitoring, and further optimization of computational efficiency to ensure reliable performance in resource-constrained settings.

7. Conclusion

This paper introduced a multi-branch deep learning approach for real-time pain evaluation and attained improved performance on several benchmarks. The suggested method achieved high accuracy, recall and F1 scores, surpassing state-of-the-art methods and showing robustness for deployment in both clinical and resource-limited environments. Through the incorporation of attention-based facial landmark analysis into a late fusion framework, the model successfully identified intricate facial patterns of pain under low latency conditions (around 120 ms per inference), making it suitable for telemedicine and real-time patient monitoring. These contributions introduce a strong foundation for the objective measurement of pain in medicine, connecting technological advancement to clinical application for enhancing patient outcomes.

Nonetheless, the limitations of this research should be mentioned. Reliance on controlled datasets may not fully capture real-world variability in lighting, demographics, or cultural pain expressions. Data sparsity and homogeneity have been identified as fundamental challenges in automated pain assessment and could influence its applicability to different environments.

Finally, as with most AI systems, our system is a black box for end users. Although the system implements attention mechanisms to improve interpretability, the decision-making process can remain opaque to clinicians. Importantly, we have not yet validated the model on real-world clinical data, a key step to ensuring that algorithmic evaluations are consistent with patient self-reports and clinical assessments, and the ethical implications of AI-driven pain assessment, such as algorithmic bias and privacy risks, require deeper scrutiny. These limitations will be actively addressed in future work to improve the safety, integrity, and reliability of the system prior to any real-world deployment.

8. Future work

To enhance our findings and address current limitations, we highlight some main directions for future research and development:

Larger and More Diverse Datasets: We plan to extend our experiments to larger, more heterogeneous datasets spanning a wide range of demographics, pain conditions, and environments. The training set should be expanded to include a wide range of patient groups (which are varying in age, sex, ethnicity, etc.) and real-world situations. This will help the model generalize and reduce bias. Diverse databases of pain are in short supply (with appropriate ethical permissions), and this is one of the main problems of reliable pain detection systems. In the near future, we will focus on the aggregation of new data. The training will be carried out on more extensive data. Cross-dataset evaluation will help to confirm that the model will work well in different hospitals and cultural and clinical contexts.

Robustness under varied conditions: In future work, we will examine and improve the model's robustness in uncontrolled and hard conditions. This will involve assessing performance under different lighting and imaging characteristics, as well as under patient poses or movements, and the presence of confounding factors (such as medical devices on the face or comorbid facial conditions). Stress-testing the system in such conditions will help to detect failure modes, as well as help to improve the model (e.g. through the use of data augmentation or specialized preprocessing) to ensure that it is reliable for real-life use. We will also consider the use of domain adaptation techniques in order to maintain the accuracy of the system when it is used outside the initially trained setting. This will help to increase the model's reliability, as well as improve our confidence in the model's predictions under all circumstances.

Hybrid Models and Domain-Specific Optimizations: Another avenue is to explore hybrid modeling approaches that combine data-driven learning with expert knowledge or other algorithmic techniques to boost performance. For instance, integrating physiological signals (heart rate, voice indicators, etc.) or rule-based clinical scoring with the vision model could create a more robust, ensemble system. We will investigate whether a universal model or tailored models for specific sub-populations yield the best outcomes. Domain-specific optimizations, such as customizing the model for particular pain types or clinical environments, may further enhance accuracy. We also intend to apply optimization strategies like model pruning, quantization, and hardware-specific tuning to ensure the algorithm runs efficiently on edge devices (in ambulatory monitors or smartphones) without significant loss of accuracy. Such optimizations would make it possible to implement the system in resource-constrained environments. The system would be able to operate in real-time at the point of care

Ethical and Responsible AI Considerations: The importance of fairness and transparency of the system is high priority in future developments. We will proactively address potential biases by evaluating model performance across different demographic groups and pain contexts and by including fairness metrics in our evaluations. We will proactively address potential biases by evaluating model performance across different demographic groups and pain contexts, and by incorporating fairness metrics into our evaluations. If disparities are identified (for example, if pain in a certain group is consistently under- or over-predicted), we will implement bias mitigation techniques (such as re-sampling, re-weighting, or debiasing adversarial training) to correct them. We also recognize the interpretability concerns in AI-driven decision-making, as black-box models can erode clinicians' trust. To improve transparency, we will integrate explainability tools (such as saliency maps, Shapley values, Grad-CAM++ visualizations or rule-based explanations of the model's outputs) for healthcare professionals to perceive the logic behind the pain Assessment. What is more, the responsible AI deployment rules will be observed: the solution will demand consent when interacting with patients, guarantee data privacy, and compliance with the regulations (HIPAA, GDPR), and be equipped with safety nets to stop excessive AI use in high-stakes decisions. By emphasizing ethics, we aim to increase public and professional confidence in the technology and promote its safe adoption.

Clinical Validation and Collaboration: Extensive clinical validation of the model is required Before real-world implementation. We will collaborate with medical professionals to test the system on actual patient data in clinical settings. We will consider a variety of medical profiles and pain etiologies. This involves prospective studies or pilot deployments where the AI's pain assessments are compared against clinicians' evaluations and patient self-reported pain scores in real time. Such validation will determine the model's reliability and consistency with established clinical assessment standards. Indeed, well-designed trials are needed to establish the generalizability and safety of AI pain models in practice. Clinicians' feedback will be used to refine the integration of the system into workflows, such as adjusting the user interface or the output format to correspond with how practitioners assess pain. also it is essential to confirm that the system meets all clinical regulatory requirements and guidelines for medical device software. Working in collaboration with healthcare providers, the goal is to make sure that AI recommendations are supplementary to clinical judgment, rather than conflicting with it, and thus support enhanced patient care.

Finally, demonstrating the model's usefulness and agreement in real-world care settings is essential in receiving regulatory approval and clinician recognition.

In summary, addressing the above aspects will mitigate not only the limitations of the current study, but will also propel the field of automatic pain assessment towards safe and efficient clinical use. By scaling and diversifying data, augmenting the model, looking for general and specific algorithms, taking into account ethical problems and validating the model with doctors, one can advance research to make the system more effective and reliable. This will result in the creation of a next-generation solution for monitoring pain that is accurate, fair, interpretable, and suitable for the outcome in healthcare settings.

CRediT authorship contribution statement

Hany El-Ghaish: Project administration, Formal analysis, Data curation. **Mohamed Yousry Al-Basiouny:** Writing – original draft, Software, Investigation. **Mahmoud A.M. Alshewimy:** Writing – review & editing, Supervision, Project administration.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used OpenAI's ChatGPT in order to assist with improving language, formatting, and enhancing the readability of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The UNBC-McMaster Shoulder Pain Expression Archive Database is publicly accessible upon formal application via the University of Northern British Columbia website (https://sites.pitt.edu/~emotion/um-spread.html) :contentReference[oaicite:0]index=0. The BioVid Heat Pain Database can be obtained by registering with the Neuro-Information Technology group at Otto von Guericke University Magdeburg (https://www.nit.ovgu.de/en/BioVid-p-1358.html) :contentReference[oaicite:1] index=1. All additional data and custom code generated during this study are available from the corresponding author upon reasonable request.

References

Alghamdi, T., Alaghband, G., 2022. Facial expressions based automatic pain assessment system. Applied Sciences 12 (13), 6423. http://dx.doi.org/10.3390/app12136423.

Aoun, N.B., 2024. A review of automatic pain assessment from facial information using machine learning. Technologies 12 (6), 92.

Badura, A., Bienkowska, M., Mysliwiec, A., Pietka, E., 2024. Continuous short-term pain assessment in temporomandibular joint therapy using LSTM models supported by heat-induced pain data patterns. IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review 54, 1937–1967.

Cui, S., Huang, D., Ni, Y., Feng, X., 2021. Multi-scale regional attention networks for pain estimation. In: Proceedings of the 2021 13th International Conference on Bioinformatics and Biomedical Technology. pp. 1–8.

Dubey, S.R., Singh, S.K., Chaudhuri, B.B., 2022. Activation functions in deep learning: a comprehensive survey and benchmark. Neurocomputing 503, 92–108.

Elgendy, F., Alshewimy, M.A., Sarhan, A.M., 2021. Pain detection/classification framework including face recognition based on the analysis of facial expressions for e-health systems. International Arab Journal of Information Technology 18 (1), 125–132.

Entezami, A., Shariatmadar, H., De Michele, C., 2022. Non-parametric empirical machine learning for short-term and long-term structural health monitoring. Structural Health Monitoring 21 (6), 2700–2718.

Fernandez Rojas, R., Joseph, C., Bargshady, G., Ou, K.-L., 2024. Empirical comparison of deep learning models for fnirs pain decoding. Frontiers in Neuroinformatics 18, 1320189.

Gadzicki, K., Khamsehashari, R., Zetzsche, C., 2020. Early vs late fusion in multimodal convolutional neural networks. In: 2020 IEEE 23rd International Conference on Information Fusion. IEEE, pp. 1–6.

Gal, M.S., Rubinfeld, D.L., 2019. Data standardization. NYUL Rev. 94, 737.

Garbin, C., Zhu, X., Marques, O., 2020. Dropout vs. batch normalization: An empirical study of their impact on deep learning. Multimedia Tools and Applications 79 (19), 12777–12815. http://dx.doi.org/10.1007/s11042-019-08453-9.

Gkikas, S., Tachos, N.S., Andreadis, S., Pezoulas, V.C., Zaridis, D., Gkois, G., Matonaki, A., Stavropoulos, T.G., Fotiadis, D.I., 2024. Multimodal automatic assessment of acute pain through facial videos and heart rate signals utilizing transformer-based architectures. Frontiers in Pain Research 5, 1372814. http://dx.doi.org/10.3389/fpain.2024.1372814.

Huang, L., Qin, J., Zhou, Y., Zhu, F., Liu, L., Shao, L., 2023. Normalization techniques in training DNNs: methodology, analysis and application. IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (8), 10173–10196.

Huang, Y., Qing, L., Xu, S., Wang, L., Peng, Y., 2022a. Hybnet: A hybrid network structure for pain intensity estimation. Visual Computer 1–12. http://dx.doi.org/10.1007/s00371-021-02056-y.

Huang, G., Wang, Y., Lv, K., Jiang, H., Huang, W., Qi, P., Song, S., 2022b. Glance and focus networks for dynamic visual recognition. IEEE transactions on pattern analysis and machine intelligence 45 (4), 4605–4621.

Huang, D., Xia, Z., Mwesigye, J., Feng, X., 2020. Pain-attentive network: a deep spatio-temporal attention model for pain estimation. Multimedia Tools and Applications 79 (37), 28329–28354. http://dx.doi.org/10.1007/s11042-020-09397-1.

Kong, Y., Posada-Quintero, H.F., Chon, K.H., 2020. Pain detection using a smartphone in real time. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society. IEEE, pp. 4526–4529.

Kong, Y., Posada-Quintero, H.F., Chon, K.H., 2021. Real-time high-level acute pain detection using a smartphone and a wrist-worn electrodermal activity sensor. Sensors 21 (12), 3956.

Kong, Y., Posada-Quintero, H.F., Chon, K.H., 2022. Multi-level pain quantification using a smartphone and electrodermal activity. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 2475–2478. http://dx.doi.org/10.1109/EMBC48229.2022.9871228.

Leroux, A., Rzasa-Lynn, R., Crainiceanu, C., Sharma, T., 2021. Wearable devices: current status and opportunities in pain assessment and management. Digital Biomarkers 5 (1), 89–102.

Llugsi, R., El Yacoubi, S., Fontaine, A., Lupera, P., 2021. Comparison between adam, adamax and adamw optimizers to implement a weather forecast based on neural networks for the andean city of quito. In: 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM). IEEE, pp. 1–6.

Lu, Z., Ozek, B., Kamarthi, S., 2023. Transformer encoder with multiscale deep learning for pain classification using physiological signals. Frontiers in Physiology 14, 1294577. http://dx.doi.org/10.3389/fphys.2023.1294577.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M.G., Lee, J., et al., 2019. Mediapipe: a framework for building perception pipelines. ArXiv preprint arXiv:1906.08172.

Mahendran, N., 2023. Variations of squeeze and excitation networks. ArXiv preprint arXiv:2304.06502.

Mienye, I.D., Swart, T.G., 2024. A comprehensive review of deep learning: Architectures, recent advances, and applications. Information 15 (12), 755.

Morales-Hernández, A., Van Nieuwenhuyse, I., Rojas Gonzalez, S., 2023. A survey on multi-objective hyperparameter optimization algorithms for machine learning. Artificial Intelligence Review 56 (8), 8043–8093.

Naidu, G., Zuva, T., Sibanda, E.M., 2023. A review of evaluation metrics in machine learning algorithms. In: Computer Science on-Line Conference. Springer, pp. 15–25.

Nguyen, M.-D., Yang, H.-J., Kim, S.-H., Shin, J.-E., Kim, S.-W., 2024. Transformer with leveraged masked autoencoder for video-based pain assessment. arXiv preprint arXiv:2409.05088.

Orăşan, I.L., Seiculescu, C., Caleanu, C.D., 2022. Benchmarking tensorflow lite quantization algorithms for deep neural networks. In: 2022 IEEE 16th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE, pp. 000221–000226.

Othman, E., Werner, P., Saxen, F., Al-Hamadi, A., Gruss, S., Walter, S., 2023. Automated electrodermal activity and facial expression analysis for continuous pain intensity monitoring on the X-ITE pain database. Life 13 (9), 1828.

Pouromran, F., Lin, Y., Kamarthi, S., 2022. Personalized deep Bi-LSTM RNN based model for pain intensity classification using EDA signal. Sensors 22 (21), 8087. http://dx.doi.org/10.3390/s22218087.

Pouromran, F., Radhakrishnan, S., Kamarthi, S., 2021. Exploration of physiological sensors, features, and machine learning models for pain intensity estimation. PLoS One 16 (7), e0254108.

Prajod, P., Schiller, D., Don, D.W., André, E., 2024. Faces of experimental pain: transferability of deep learned heat pain features to electrical pain. arXiv preprint arXiv:2406.11808.

Purwono, P., Ma'arif, A., Rahmaniar, W., Fathurrahman, H.I.K., Frisky, A.Z.K., ul Haq, Q.M., 2022. Understanding of convolutional neural network (CNN): A review. International Journal of Robotics and Control Systems 2 (4), 739–748.

Rojas, R.F., Romero, J., Lopez-Aparicio, J., Ou, K.-L., 2020. Pain assessment based on fnirs using bidirectional LSTMs. ArXiv preprint arXiv:2012.13231.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128, 336–359. http://dx.doi.org/10.1007/s11263-019-01228-7.

Serraoui, I., Granger, E., Hadid, A., Taleb-Ahmed, A., 2023. Pain analysis using adaptive hierarchical spatiotemporal dynamic imaging. arXiv preprint arXiv:2312.06920.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. Journal of big data 6 (1), 1–48.

Sirocchi, C., Bogliolo, A., Montagna, S., 2024. Medical-informed machine learning: integrating prior knowledge into medical decision systems. BMC Medical Informatics and Decision Making 24 (Suppl 4), 186.

Sivakumar, M., Kumar, N., Karthikeyan, N., 2022. An efficient deep learning-based content-based image retrieval framework. Computers & Systems Science & Engineering 43 (2), 123–134.

Tavakolian, M., Hadid, A., 2019. A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics. International Journal of Computer Vision 127, 1413–1425. http://dx.doi.org/10.1007/s11263-019-01191-3.

Tian, Y., Zhang, Y., Zhang, H., 2023. Recent advances in stochastic gradient descent in deep learning. Mathematics 11 (3), 682.

Van Houdt, G., Mosquera, C., N'apoles, G., 2020. A review on the long short-term memory model. Artificial Intelligence Review 53 (8), 5929–5955. http://dx.doi.org/10.1007/s10462-020-09838-1.

Wang, Y., Zhang, H., Yue, Y., Song, S., Deng, C., Feng, J., Huang, G., 2024. Uni-adafocus: spatial-temporal dynamic computation for video recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Wu, Z., Li, H., Xiong, C., Jiang, Y.-G., Davis, L.S., 2022. A dynamic frame selection framework for fast video recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (4), 1699–1711. http://dx.doi.org/10.1109/TPAMI.2020.3029425.

Wu, Z., Li, H., Zheng, Y., Xiong, C., Jiang, Y.-G., Davis, L.S., 2021. A coarse-to-fine framework for resource efficient video recognition. International Journal of Computer Vision 129 (11), 2965–2977. http://dx.doi.org/10.1007/s11263-021-01508-1.

Xing, Z., Yang, Y., Tan, L., Guo, X., 2025. Multi-source physical information driven deep learning in intelligent education: unleashing the potential of deep neural networks in complex educational evaluation. AIP Advances 15 (2).

Xing, Z., Zhao, S., Guo, W., Meng, F., Guo, X., Wang, S., He, H., 2023. Coal resources under carbon peak: Segmentation of massive laser point clouds for coal mining in underground dusty environments using integrated graph deep learning model. Energy 285, 128771.

Xu, H., Liu, M., 2021. A deep attention transformer network for pain estimation with facial expression video. In: Biometric Recognition: 15th Chinese Conference, CCBR 2021, Shanghai, China, September 10–12, 2021, Proceedings 15. Springer, pp. 112–119.

Yates, L.A., Aandahl, Z., Richards, S.A., Brook, B.W., 2023. Cross validation for model selection: a review with examples from ecology. Ecological Monographs 93 (1), e1557.

Yin, P., Zhang, X., Hao, L., 2022. Deep learning assessment method for postoperative pain based on facial video data. In: Journal of Physics: Conference Series. 2356, p. 012052.

Younesi, A., Ansari, M., Fazli, M., Ejlali, A., Shafique, M., Henkel, J., 2024. A comprehensive survey of convolutions in deep learning: Applications, challenges, and future trends. IEEE Access 12, 41180–41218.