Dean&Francis
ISSN 2959-6157

A Comprehensive Investigate of Deep Learning in Facial Pain Prediction

Minhao Li

School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, China
liwt10009@gmail.com

Abstract:

Traditional approaches to pain assessment, such as patient self-reporting and clinician observation, have inherent limitations, particularly for non-verbal and vulnerable populations, due to their subjective nature and lack of objectivity. This study delves into the potential of deep learning for facial pain prediction. Five primary model categories are examined: Convolutional Neural Networks (CNNs) like the improved EfficientNet B4S demonstrate 99.7% accuracy in detecting high-intensity pain, while ResNet101 combined with LSTM achieves 86.13% accuracy in binary classification. Spatio-temporal models, exemplified by AHDI, surpass current state-of-the-art methods. Additionally, Transformer-based architectures, point cloud/Graph Neural Networks (GNNs), and multimodal fusion models exhibit promising results. However, challenges persist, including model interpretability issues, limited clinical generalizability, and data annotation bottlenecks. Future research will emphasize explainable AI (XAI), domain adaptation, and lightweight, privacy-preserving model deployment to facilitate the transition from laboratory settings to clinical practice, ultimately benefiting non-verbal patients, and achieving for more equitable, reliable, and ethically responsible pain assessment solutions.

Keywords: Deep learning, facial pain prediction, XAI

1. Introduction

Physical pain is a ubiquitous experience encountered in various clinical settings, including post-surgical recovery, chronic conditions like arthritis, and traumatic injuries. Traditional methods for assessing pain, primarily relying on patient self-reporting and clinician observation, possess significant limitations. Self-reporting is impractical for non-verbal patients such as infants or those under intubation, or for individuals with cognitive impairments. Even among verbal patients, the timing of pain reporting can be delayed, particularly in post-operative scenarios, and self-reported scales often overlook subtle facial expressions indicative of pain. Furthermore, clinician assessments are inherently subjective and prone to inconsistencies, impacted by varying levels of training and potential inattention during observations. These assessments frequently miss delicate signs of

pain, such as muscle micro-movements, specific Action Units (AUs), and fleeting grimaces, all of which are critical indicators of a patient's discomfort.

Deep learning has emerged as a powerful tool in bridging the gap between subtle, imperceptible features and their accurate interpretation, particularly in the context of facial pain assessment. Sophisticated integrated frameworks that harmonize Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) have demonstrated remarkable success. These frameworks have achieved an impressive accuracy rate exceeding 89% in classifying pain levels solely from facial expressions, surpassing human-centered evaluations in the process [1-4]. In clinical environments, the application of deep learning for facial pain prediction has proven invaluable, especially in critical areas such as postoperative and trauma care. Persistent postsurgical pain affects a staggering 57% of patients [5], often due to inadequate early assessments that lead to treatment delays. To address this challenge, the UNBC-McMaster Shoulder Pain Expression Database has provided a crucial resource. This database comprises 200 video sequences from 25 patients suffering from shoulder pain, each meticulously annotated with pain intensity scores ranging from 0 to 10 points and corresponding facial action units, such as frowning and eyelid contraction [6]. Deep learning models, when trained on this dataset, exhibit a remarkable ability to extract spatial features of facial muscle contractions. Convolutional neural networks perform very well in recognizing fine-grained features. For instance, they can accurately identify the specific depth of the crease between the eyebrows and the actual degree of lip drooping. Concurrently, RNNs enable temporal modeling, capturing the dynamic evolution of pain expressions—transitioning from mild discomfort to overt grimacing. The combined prowess of these models has led to an accuracy rate exceeding 85% in grading pain during rehabilitation, outperforming subjective assessments using the Visual Analogue Scale (VAS) by a significant margin. Thus, deep learning not only amplifies the capacity to recognize and interpret facial cues of pain but also enhances the precision and reliability of pain assessments in clinical settings, ultimately paving the way for more timely and effective treatment interventions.

The intensive care unit (ICU) poses unique and intricate challenges, particularly in assessing and managing pain in patients. Studies reveal that 3.3% of Critical-Care Pain Observation Tool (CPOT) assessments indicate severe pain in ICU patients [7]. However, these patients, often intubated or sedated, are unable to communicate their discomfort. Nurses, conducting manual hourly assessments, often miss fleeting episodes of pain.To tackle this pressing issue, deep learning systems have been introduced, leveraging real-time facial image capture via bedside cameras. These systems are designed to operate even when facial features are partially obscured by medical devices like oxygen tubes. By focusing on crucial regions such as the brow ridge and eyelids, and utilizing recurrent neural networks (RNNs) to track microexpression changes within 10 seconds (for instance, transient eyelid tightness), these systems demonstrate remarkable accuracy.A study conducted by Wu et al. highlights the system's efficacy, with a sensitivity of 89% in identifying severe pain (CPOT score 2). This represents a significant improvement of 23% over manual assessments. Such advanced technology offers immediate insights, guiding healthcare professionals to promptly adjust analgesic medications and improve patient care.

At the algorithmic frontier, two significant technological advancements stand out. The integration of Hybrid CNN-RNN architectures has proven pivotal in addressing the intricacies of recognizing dynamic pain expressions. By employing a "spatial-temporal" dual-dimensional modeling approach, these architectures offer a comprehensive solution. Li et al.'s research utilized ResNet34 as the convolutional neural network (CNN) backbone, enabling the extraction of intricate, pain-related facial features. This includes granular details such as pixel-level changes in orbicularis oculi contraction. Complementing this spatial analysis, the study employed Bidirectional Long Short-Term Memory (BiLSTM) to process temporal information across 50 frames per video. This allowed the model to capture the evolutionary patterns of pain expressions, such as the progressive muscle movement from calmness to tension. When tested on a clinical dataset comprising 63 critically ill patients, the hybrid model demonstrated remarkable performance, achieving an accuracy of 89.2% in classifying pain grades (ranging from 0 to 2 points). Notably, it exhibited 91% specificity in identifying grade 2 pain, which necessitates urgent intervention. These results significantly surpass those of traditional machine learning models, like Support Vector Machines (SVM), which typically achieve around 76% accuracy.

Multimodal fusion technology has significantly advanced beyond the constraints of singular data sources, benefiting greatly from the comprehensive support of the MINT database. This repository houses a treasure trove of diverse information, including RGB facial images, depth data reflecting 3D facial structures, thermal imaging that captures facial blood flow variations, and synchronized physiological indicators such as heart rate and skin conductance from 100 participants. The database encompasses various pain stimuli, ranging from pressure to temperature. Models trained on this multifaceted data employ an advanced "early feature concatenation" technique, which

ISSN 2959-6157

smartly amalgamates facial texture features extracted through convolutional neural networks (CNNs) with the temporal dynamics of physiological signals—for instance, the instantaneous surge in heart rate during pain. These integrated models have achieved an impressive 91% accuracy in estimating pain intensity, marking a notable 9-percentage-point improvement over models relying solely on RGB images. This remarkable outcome underscores the pivotal importance of incorporating multi-dimensional data in enhancing the robustness of pain assessment systems.

This paper delves into the realm of facial pain prediction, focusing on the utilization of deep learning techniques. It seeks to elucidate the technical underpinnings, clinical implications, and future trajectories of this application. Section 2 delves into the core deep learning models, highlighting the innovative aspects and operational workflows of models like Convolutional Neural Networks (CNNs), Vision Transformers (ViT), and Long Short-Term Memory networks (LSTM). Section 3 examines the existing challenges, such as the interpretability of models and their clinical viability, while also offering insights into potential future developments, notably domain adaptation. Finally, Section 4 concludes by summarizing the pivotal findings and the significant contributions of this research.

2. Current Model Research Progress

2.1 Convolutional Neural Networks (CNNs) and Their Variants

CNNs serve as a cornerstone in image processing, particularly for facial pain prediction. Classic architectures like VGG, ResNet, and MobileNet leverage multi-layer convolution and pooling to extract facial features hierarchically, laying a solid groundwork for pain classification tasks.

2.1.1 Application of EfficientNet in Pain Expression Evaluation

EfficientNet, a pioneering convolutional neural network architecture, excels at balancing depth, width, and resolution via a sophisticated compound scaling technique, enabling it to achieve high classification accuracy with minimal parameters. In 2023, Chen and colleagues leveraged this efficient framework in their study titled „Study on Cancer Pain Facial Expression Evaluation Method Based on EfficientNet“ [8]. They opted for the EfficientNet B4 model, renowned for its optimal blend of accuracy and speed, as the foundation for their research. The study introduced two pivotal enhancements to the basic architecture. Firstly, they substituted the Swish activation function with Mish, finding that Mish more effectively sustains accuracy in deep networks and enhances information flow. Secondly, they replaced the MobileNet module with the Inception v4 module, a strategic move designed to mitigate overfitting and lighten the computational burden. These modifications collectively contributed to a refined and more effective method for evaluating cancer pain facial expressions.

The EfficientNet B4S model demonstrated remarkable performance across various pain levels in the test set [8]. Specifically, when individuals reported minimal pain (0–1), the model achieved an accuracy rate of 96.3%. As pain intensified to the mild-to-moderate range (2–3), accuracy dipped to 82.1%. However, for moderate pain (level 4), accuracy rebounded to 94.6%. Notably, for severe pain levels (5–6), the model excelled with an accuracy rate of 99.7%. With extreme pain levels (7–8), accuracy declined slightly but remained high at 96.8%.

This study demonstrates that the EfficientNet architecture exhibits high accuracy and stability in facial pain prediction, particularly in identifying high-intensity pain [8].

2.1.2 Residual Network (ResNet)-Based Pain Evaluation Model

In 2023, Wu Jiang and colleagues presented a groundbreaking model for facial pain prediction in their study, leveraging the powerful capabilities of ResNet and its advancements. Specifically, they introduced a dynamic video pain evaluation model that seamlessly integrates ResNet101 with Long Short-Term Memory (LSTM) networks. The model‘s architecture is meticulously designed: Firstly, it employs a pre-trained ResNet101 to meticulously extract spatial features from each individual frame of the video. Subsequently, these features are fed into an LSTM network, which effectively captures the temporal correlations between consecutive frames. Ultimately, a fully connected layer processes this comprehensive information to automate the pain evaluation process, showcasing the model‘s robustness and precision in detecting facial pain cues from video data.

To bolster the feature representation in pain evaluation, the researchers innovatively introduced a Dynamic Fusion Module (DFNB). This module employs a Non-local framework as its backbone. By intricately merging features derived from the fourth and fifth blocks of the ResNet101 network, the DFNB notably enhanced performance. Notably, experimental outcomes revealed that the model attained an impressive 86.13% accuracy in binary classification on the UNBC-McMaster Shoulder Pain Expression Dataset, surpassing conventional methods by a significant margin.

2.2 Spatio-Temporal Feature Fusion Models

Pain intensity evaluation hinges on understanding facial expressions‘ dynamic nature. Studies thus prioritize developing deep learning models to seamlessly integrate temporal changes in these expressions.

2.2.1 Video Pain Evaluation Model Based on Dynamic Fusion Module

In 2023, Wu et al. introduced an innovative method for evaluating pain in facial dynamic videos, focusing on the utilization of a dynamic fusion module. Their groundbreaking work, titled „Facial Dynamic Video Pain Evaluation Method Based on Dynamic Fusion Module,“ outlined the key steps involved in this process. Initially, they addressed the imbalance in the UNBC-McMaster Shoulder Pain Expression Dataset by performing data balancing, ensuring an equitable distribution of painful and non-painful frames. To enhance feature extraction, they improved the pre-trained ResNet101 network using a Non-local framework, thereby creating a dynamic fusion module. This module formed the foundation of their image spatial feature extraction model. Subsequently, they employed this model to extract spatial feature information from each frame of the video. These features were then fed into an LSTM network to capture temporal dynamics. Through this two-step process—spatial feature extraction followed by temporal analysis—Wu et al. achieved highly accurate facial pain recognition and automatic evaluation.

The novelty of this approach resides in its dynamic fusion module, strategically integrating features from the fourth and fifth blocks of the ResNet101 network. This fusion effectively captures the subtleties of facial expression changes, addressing clinical demands in both offline analysis and real-time applications. Consequently, it offers a viable solution for prolonged pain monitoring.

2.2.2 Adaptive Hierarchical Spatio-Temporal Dynamic Imaging (AHDI) Technology

In 2023, Issam and colleagues introduced a groundbreaking technology named Adaptive Hierarchical Spatio-temporal Dynamic Image (AHDI) for pain assessment. AHDI encodes the spatial and temporal variations in facial videos into a single, comprehensive RGB image. This innovation streamlines video processing significantly, allowing for the application of simpler 2D deep learning models. Core attributes of AHDI include its ability to transform videos into a solitary dynamic image, simplifying workflows. It utilizes residual networks to extract facial features, enhancing the accuracy of pain intensity estimation and differentiating between genuine and simulated pain expressions. Additionally, AHDI minimizes the dependence on labeled data, which decreases the time and expense associated with data collection.

AHDI technology has demonstrated superior performance in pain detection compared to existing techniques. On the UNBC dataset, its Mean Squared Error (MSE) was 0.27, surpassing the previous State-of-the-Art (SOTA) by 0.13. Similarly, on the BioVid dataset, AHDI achieved an accuracy of 89.76%, marking a 5.37% improvement over the SOTA. Notably, in distinguishing between real and simulated pain, AHDI exhibited an accuracy of 94.03%, a significant 8.98% higher than the previous benchmark.

This study provides an efficient and accurate new method for video-based pain assessment, with significant application value in real clinical environments [9].

2.3 Application of Transformer Architecture in Pain Prediction

Due to its strong global modeling capability and ability to capture long-range dependencies, the Transformer architecture has received extensive attention in recent studies on facial pain prediction.

2.3.1 Pain Automatic Recognition System Combining ResNet and Transformer

In 2023, Wei Xinyi and colleagues introduced an innovative AI system for recognizing pain through facial expressions in their paper. They designed a hybrid neural network architecture that merges ResNet18 with a Transformer. This system‘s core strengths lie in its ability to extract detailed facial features using ResNet18 and then model intricate relationships between these features via the Transformer network. To assess the model‘s efficacy, rigorous evaluation methods such as K-fold cross-validation and external validation were employed, ensuring robust performance in recognizing pain from facial expressions.

Experimental results revealed that the hybrid model excelled in pain classification tasks across various diseases. With an accuracy of 79%, sensitivity of 80%, and specificity of 93%, it demonstrated robust performance. The model also showed a positive predictive value of 82%, a negative predictive value of 94%, and an F1 score of 80%. External validation showed a slight improvement in system performance. Accuracy reached 82%, Sensitivity 78%, Specificity 93%, PPV 80%, NPV 94%, and F1 Score 78%.

The research team created AI systems for automatically recognizing pain in specific conditions, such as herpes zoster, scapulohumeral periarthritis, and myofasciitis. Notably, the AI for scapulohumeral periarthritis excelled, achieving 78% accuracy and a 79% F1 score.

To enhance model interpretability, researchers employed Grad-CAM to visualize AI focus during facial classifi-

ISSN 2959-6157

cation. Findings revealed the model precisely targeted pain-related facial areas, like the periorbital and nasal regions.

2.3.2 Transformer-Based Pain Analysis Method Using Spatio-Temporal Information

In 2023, Ye and colleagues introduced a novel approach for analyzing facial pain expressions by leveraging spatio-temporal information in their research titled „Study on Facial Pain Expression Analysis Method Based on Spatio-Temporal Information“ [10]. The core of this method lies in the Transformer architecture, which is employed to accurately capture the intricate spatial and temporal characteristics of facial expressions. Here‘s an in-depth look at the methodology:Initially, the input facial expression video is methodically sampled using a sliding window technique. This results in multiple video segments, each providing a comprehensive view of the facial movements. Subsequently, each frame within these segments undergoes Discrete Cosine Transform (DCT), a process designed to enhance fine details in the facial features, thus facilitating better feature extraction.Next, these processed video frames are fed into the Transformer model. Within this model, a serial cross-approach is utilized to methodically extract both temporal and spatial features. This approach ensures that the model not only understands the individual facial configurations but also how they evolve over time.Finally, the pain analysis result is derived by aggregating the pain levels detected across all sampled videos. This comprehensive approach ensures a robust and reliable pain expression analysis, leveraging the rich information embedded in the facial movements over time.

The innovative approach leverages a serial cross method within the Transformer model to adeptly extract both temporal and spatial features from facial expression videos. This technique efficiently captures dynamic shifts in facial expressions, enhancing the extraction of feature information. Consequently, it boosts the accuracy of pain evaluations and offers robust temporality, ideal for extended pain monitoring scenarios.

2.4 Pain Prediction Based on Point Clouds and Graph Neural Networks (GNNs)

In recent years, point cloud processing and Graph Neural Networks (GNNs) have shown unique advantages in facial pain prediction, particularly in capturing facial geometric features and dynamic changes.

2.4.1 Facial Feature Point Cloud Pain Diagnosis Model Based on Spatio-Temporal Distribution

In 2025, Li Zhipeng and colleagues introduced a groundbreaking pain diagnosis model leveraging facial feature point clouds and their spatio-temporal distribution. Their study, titled „Study on Facial Feature Point Cloud Pain Diagnosis Model Based on Spatio-Temporal Distribution,“ outlined a method utilizing facial videos to diagnose pain. The process begins by utilizing a sophisticated facial feature point extraction model to create a dynamic 3D point set of facial landmarks, which undergoes normalization to standardize the data. Subsequently, a point cloud classification model is employed to derive feature vectors from this normalized information. These feature vectors encapsulate the intricate changes in facial expressions over time associated with pain. Finally, a classifier is utilized to accomplish two tasks: binary classification, distinguishing between painful and non-painful expressions, and a more nuanced five-level pain classification. This innovative approach holds promise for advancing pain assessment in medical diagnostics.

The model was rigorously trained and evaluated on the BioVid Heat Pain Dataset, yielding impressive results: 84.98% accuracy in binary pain recognition and 37.65% in five-level classification. Notably, it excelled in identifying level 4 pain with 72% precision.

Experimental results showed that the model has unique advantages in capturing dynamic changes and geometric features of facial expressions, providing a new technical path for pain diagnosis [11].

2.4.2 Application of Weighted Graph Neural Network (WGNN) in Pain Evaluation

In 2025, researchers introduced an innovative approach to evaluating sheep pain by leveraging Weighted Graph Neural Networks (WGNNs) in their paper titled „Study on Sheep Pain Evaluation Method Based on Weighted Graph Neural Network.“ The core of this method revolves around correlating facial landmarks with pain levels. To facilitate this, they compiled a novel dataset of sheep facial landmarks that aligns with the parameters of the Sheep Pain Facial Expression Scale (SPFES). Utilizing the YOLOv8n detector, the researchers accurately identified these landmarks, achieving an average precision of 59.30%. By employing the WGNN model, they were able to establish connections between the detected facial landmarks and define corresponding pain levels. Notably, their method attained an impressive accuracy rate of 92.71%, demonstrating its effectiveness in tracking subtle expression changes across multiple facial regions. This breakthrough offers a promising tool for assessing sheep pain with enhanced precision and reliability.

This study on sheep pain evaluation offers significant insights for human facial pain prediction. Its methods and concepts provide new perspectives on analyzing dynamic facial landmark changes, enhancing the understanding of

their correlation with pain levels.

2.5 Multimodal Fusion Models

Multimodal fusion models combine facial expressions with other physiological signals (e.g., heart rate, electrodermal activity) to provide more comprehensive pain information and improve evaluation accuracy.

2.5.1 Pain Evaluation by Fusing Facial Videos and Physiological Signals

In 2024, Gkikas and colleagues presented an innovative multimodal framework for pain assessment leveraging facial videos and heart rate signals, grounded in Transformer architecture. Their study, titled „A Transformer-based Multimodal Framework for Pain Assessment," emphasized the synergy between behavioral and physiological indicators. The facial video component extracted data from 30 frames per clip, while functional Near-Infrared Spectroscopy (fNIRS) measured oxyhemoglobin and deoxyhemoglobin across 24 channels. This combined approach significantly bolstered the accuracy of pain estimation, showcasing the potential of multimodal integration in enhancing diagnostic capabilities.

This multimodal approach excels in amalgamating pain-related data from diverse sources, offering a holistic view of pain characteristics. This integration enhances the precision and dependability of pain evaluations in clinical settings. Specifically, multimodal fusion models adeptly manage complex and varied pain presentations, demonstrating significant benefits in differentiating subtle pain intensities and pinpointing unique pain types.

2.5.2 Pain Recognition by Fusing Facial Expressions and Non-Contact Physiological Signals

In 2024, researchers introduced an innovative non-contact pain recognition system leveraging video-based data analysis. This system integrates two key branches: one focused on facial expressions and another on non-contact physiological signals. By merging insights from both visual and physiological cues, the multimodal approach enhances the system's capability to capture a wide range of pain-related features, ultimately boosting its accuracy in detecting pain states.

This system offers a significant advantage by eliminating direct patient contact, thus minimizing discomfort and the risk of cross-infection. It enables real-time monitoring of patients' pain responses, equipping doctors with timely data to devise more precise pain management plans.

3. Discussion

3.1 Challenges and Limitations

3.1.1 Limited Model Interpretability

Many AI models in healthcare operate as black boxes, eroding clinical trust—a crucial factor for their successful implementation. Although Wei et al. employed Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize pain-related facial areas, such as the periorbital and nasal regions, this technique merely reveals where the model focuses. It fails to elucidate how or why specific features, like eyebrow furrowing or lip trembling, correspond to particular pain intensities. For AI models with more intricate designs, this limitation becomes even more pronounced.

Transformer models rely on global attention mechanisms, but the interpretation of attention weights (e.g., how temporal dependencies between frames drive pain classification) remains unaddressed;

GNN and point cloud models transform facial landmarks into graphs or 3D point sets, but the relationship between node/geometric alterations and pain intensity remains unclear.

Lack of interpretable decision-making hinders clinicians from aligning model outputs with their expertise, impeding use in critical cases like pediatric or impaired patients unable to communicate pain.

3.1.2 Poor Applicability to Diverse Clinical Scenarios

Current models suffer from narrow generalizability due to constraints in data sources and application scenarios:

Dataset homogeneity presents a critical challenge in pain-focused AI research. Many studies utilize specialized datasets tailored to specific pain conditions, such as UNBC-McMaster for shoulder pain or BioVid for heat pain, often restricted to narrow age, etiology, or ethnic demographics. Consequently, systems trained on these homogeneous datasets, like an AI system achieving 78% accuracy in scapulohumeral periarthritis, may demonstrate limited generalizability. Clinical pain exhibits significant heterogeneity; for instance, elderly patients may display subtle facial cues, whereas children might exhibit exaggerated movements. Artificial intelligence models trained on a unified dataset may encounter obstacles when adapting to and accurately evaluating various pain manifestations in actual clinical environments.

Multimodal models necessitate specialized, costly equipment like 24-channel fNIRS, limiting their application to hospitals with extensive resources. Non-contact multimodal systems alleviate patient discomfort but are suscep-

ISSN 2959-6157

tible to environmental factors such as lighting and motion artifacts, rendering their effectiveness uncertain in busy, noisy clinical environments like emergency rooms.

The WGNN model, while validated on sheep, faces challenges due to significant anatomical and expressive differences between sheep and humans, necessitating substantial retraining for direct application to humans. Similarly, models built on adult data may not adequately generalize to pediatric patients or those with facial paralysis.

3.1.3 Data Dependency and Annotation Bottlenecks

High-performance models rely heavily on large-scale, high-quality labeled data, which is scarce and costly to obtain in pain research:

Pain levels are often annotated by patient self-reports or clinician evaluations, which can lead to inconsistencies. For instance, a patient might rate their pain as a 4/10, while a clinician assesses it as a 3/10. This subjectivity poses a challenge, particularly in cases where patients cannot communicate, such as infants or coma patients, introducing noise into the training data.

Label scarcity remains a significant challenge. Though Issam et al. introduced AHDI to minimize labeled data dependence, mainstream models like ResNet101-LSTM and point cloud models still necessitate vast annotated frames. Additionally, clinical data gathering faces privacy hurdles like HIPAA, hindering access to diverse datasets.

Class imbalance in datasets is a persistent issue, particularly when it comes to pain intensity levels. High-intensity pain (levels 7–8) is less frequently represented than low-intensity pain (levels 0–1). This disparity can skew model training towards majority classes, reducing their efficacy in detecting severe pain—a critical factor for prompt clinical intervention. Wu et al. attempted to address this imbalance through data balancing, highlighting the necessity of addressing this issue to ensure accurate pain detection [12].

3.2 . Future Prospects

3.2.1 Enhancing Interpretability with Explainable AI (XAI)

To build clinical trust, future models should integrate XAI techniques tailored to pain assessment:

Fine-tune feature interpretation by integrating Grad-CAM with SHAP or LIME. These methods quantify facial feature contributions, such as eyebrow depression accounting for 35% in pain level 5 classification.

Interpretable Module Design enhances model understanding. For Transformers, visualize temporal attention weights to illustrate how frame changes influence pain level updates. For GNNs, emphasize key landmark links, such as the correlation between outer canthus-eyebrow distance and 40% of pain level 4 predictions.

Create clinician-friendly dashboards displaying XAI results via intuitive formats like heatmaps on facial images and feature contribution bar charts for swift validation.

3.2.2 Domain Adaptation for Enhanced Generalizability

Domain adaptation (DA) is essential for overcoming dataset homogeneity by transferring knowledge. Models can leverage labeled data from a source domain, like laboratory datasets, to unlabeled or partially labeled real-world clinical data. Key approaches involve adapting these models to different hospitals or populations effectively.

Develop cross-disease semi-supervised domain adaptation techniques to generalize single-pain models, like those for cancer pain, to multiple pain etiologies, by aligning feature distributions across diverse conditions.

Cross-device and cross-scenario domain adaptation (DA) mitigates data variability from diverse cameras or physiological sensors. By employing adversarial DA, a discriminator network minimizes discrepancies between source and target device data, enhancing accuracy.

Tailor DA frameworks to protect vulnerable groups like children and the elderly. By integrating age/ethnicity meta-features, this paper adjusts model parameters for equitable performance across demographics.

3.2.3 Deepening Multimodal Fusion with Advanced Strategies

Current multimodal models primarily use simple feature concatenation, which fails to prioritize informative modalities. Future work should focus on:

Modality-Aware Attention Mechanisms dynamically adjust focus on different pain indicators. For severe pain, they prioritize physiological signals like heart rate variability, as facial expressions may be concealed. For mild pain, facial features are emphasized.

Enhance pain assessment by integrating diverse data: voice groans, guarded movements, and EEG signals, forming a comprehensive pain fingerprint.

Optimizing non-contact modality enhances physiological signal accuracy, such as heart rate via remote photoplethysmography, by minimizing environmental noise through adaptive filtering, suiting home or limited-resource clinical environments.

3.2.4 Model Lightweighting and Clinical Deployment

To bridge the gap between laboratory performance and real-world use, models must be optimized for portability and efficiency:

Lightweight architecture design involves techniques such as knowledge distillation, which condenses large models

like ResNet101 into more compact ResNet18 versions, and neural architecture search (NAS) to cut down parameters without sacrificing accuracy. The adaptive approach of efficient networks can generate particularly lightweight models that are especially suitable for mobile devices. This capability has significantly enhanced some applications, such as bed monitoring systems based on tablet computers.

Federated Learning ensures data privacy in medical training by allowing hospitals to train local models. Only model parameters are shared, preventing the need for sensitive data transmission across institutions.

Conduct extensive multi-center clinical trials across diverse settings like hospitals and nursing homes to validate models and secure regulatory approvals, such as FDA clearance, adhering to healthcare standards.

4. Conclusion

Facial pain assessment aids vulnerable patients, prompting AI integration in pain management. This study explores deep learning for facial pain prediction, merging AI with clinical evaluations.

The core contributions encompass five distinct model categories, each tailored for specific tasks. Advanced CNN variations, notably the refined EfficientNet B4S integrated with Mish activation and Inception v4, attain a remarkable 99.7% accuracy in detecting high-intensity pain. Meanwhile, the ResNet101-LSTM combination excels in binary classification, demonstrating its prowess. Spatio-temporal fusion models, exemplified by AHDI, streamline video processing and surpass previous methodologies in effectiveness. Transformer-based frameworks enhance global feature representation, with hybrid models yielding accuracy rates between 79% and 82% in mixed-disease classifications. Point cloud and Graph Neural Network (GNN) models harness geometric properties, achieving high accuracies of 84.98% and 92.71% on respective datasets. Lastly, multimodal fusion models integrate facial and physiological data, bolstering comprehensiveness and minimizing patient discomfort.

These models, while effective, face significant limitations. Their complex architectures render them uninterpretable, akin to black boxes. Clinical applicability is restricted by homogeneous training data and specialized hardware. Further challenges include data dependencies, annotation issues, and privacy concerns.

Future studies aim to bolster AI-driven facial pain prediction tools for non-verbal patients. Efforts will include enhancing interpretability via XAI, boosting generalizability through domain adaptation, refining multimodal fusion, and ensuring lightweight, privacy-preserving deployment for clinical use.

References

[1] Turk DC, Melzack R. Handbook of pain assessment. 4th ed. New York: Guilford Press; 2019.

[2] Duchenne GBA. The mechanism of human facial expression. Paris: Jules Renard; 1862.

[3] Empirical comparison of deep learning models for fNIRS pain decoding. Front Neuroinform. 2024;18:1320189.

[4] Computer aided pain detection and intensity estimation using compact CNN based fusion network. Comput Methods Programs Biomed. 2021;210:106381. doi:10.1016/j.cmpb.2021.106381.

[5] Kehlet H, Jensen TS, Woolf CJ. Persistent postsurgical pain: risk factors and prevention. Lancet. 2006;367(9522):1618–25.

[6] Bartlett C, Prkachin KM, Solomon S, et al. The UNBC-McMaster shoulder pain expression database. Proc IEEE Int Conf Multim Expo. 2005;417–20.

[7] van Doorn S, et al. Multimodal integrative pain database (MINT). Sci Data. 2020.

[8] Chen XJ, et al. Evaluating cancer pain via facial expressions using EfficientNet. Comput Digit Eng. 2023.

[9] Serraoui S, et al. Adaptive hierarchical method for pain assessment via spatio-temporal dynamic image analysis. Sensors. 2023.

[10] Ye XT, et al. Analyzing facial pain expressions using spatio-temporal information. Comput Appl Softw. 2023.

[11] Li ZP, Li YJ, Xu WQ, et al. Facial feature point cloud model for pain diagnosis based on spatio-temporal data analysis. Chin J Sci Instrum. 2025.

[12] Prkachin KM. The facial expression of pain: an evolutionary account. J Nonverbal Behav. 1992;16(2):119–35.