

Automated facial recognition system using deep learning for pain assessment in adults with cerebral palsy

DIGITAL HEALTH
Volume 10: 1–22
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241259664
journals.sagepub.com/home/dhj



Álvaro Sabater-Gárriz^{1,2,3,4} , F Xavier Gaya-Morey⁵ ,
José María Buades-Rubio^{3,5}, Cristina Manresa-Yee^{3,5}, Pedro Montoya^{3,4,6}
and Inmaculada Riquelme^{2,3,4}

Abstract

Objective: Assessing pain in individuals with neurological conditions like cerebral palsy is challenging due to limited self-reporting and expression abilities. Current methods lack sensitivity and specificity, underlining the need for a reliable evaluation protocol. An automated facial recognition system could revolutionize pain assessment for such patients.

The research focuses on two primary goals: developing a dataset of facial pain expressions for individuals with cerebral palsy and creating a deep learning-based automated system for pain assessment tailored to this group.

Methods: The study trained ten neural networks using three pain image databases and a newly curated CP-PAIN Dataset of 109 images from cerebral palsy patients, classified by experts using the Facial Action Coding System.

Results: The InceptionV3 model demonstrated promising results, achieving 62.67% accuracy and a 61.12% F1 score on the CP-PAIN dataset. Explainable AI techniques confirmed the consistency of crucial features for pain identification across models.

Conclusion: The study underscores the potential of deep learning in developing reliable pain detection systems using facial recognition for individuals with communication impairments due to neurological conditions. A more extensive and diverse dataset could further enhance the models' sensitivity to subtle pain expressions in cerebral palsy patients and possibly extend to other complex neurological disorders. This research marks a significant step toward more empathetic and accurate pain management for vulnerable populations.

Keywords

Pain assessment, automated facial recognition, deep learning, pain expression image dataset, cerebral palsy

Submission date: 9 January 2024; Acceptance date: 7 May 2024

¹Department of Research and Training, Balearic ASPACE Foundation, Marratxí, Spain

²Department of Nursing and Physiotherapy, University of the Balearic Islands, Palma de Mallorca, Spain

³Research Institute on Health Sciences (IUNICS), University of the Balearic Islands, Palma de Mallorca, Spain

⁴Health Research Institute of the Balearic Islands (IdISBa), Palma de Mallorca, Spain

⁵Department of Mathematics and Computer Science, University of the Balearic Islands, Palma de Mallorca, Spain

⁶Center for Mathematics, Computation and Cognition, Federal University of ABC, São Bernardo do Campo, Brazil

Corresponding author:

Inmaculada Riquelme, Department of Nursing and Physiotherapy and Research Institute on Health Sciences (IUNICS-IdISBa), University of Balearic Islands, Carretera de Valldemossa km 7.5, 07122 Palma, Spain.

Email: inma.riquelme@uib.es



Highlights

- We construct a pioneering dataset of facial images illustrating pain in cerebral palsy.
- Our automated facial recognition system can enhance pain assessment in cerebral palsy.
- This novel system may be extrapolated to other complex neurological conditions.

Introduction

Pain assessment is exceedingly challenging in individuals who, in addition to lacking self-report capabilities, present complex neurological conditions that impact both facial and bodily expressions. Cerebral palsy (CP) stands as a group of enduring neurological disorders impacting motor function, ranking among the most prevalent lifetime medical conditions.^{1,2} Its prevalence, estimated at 2–3 per 1000 live births in developed countries,³ highlights its significant impact. Due to disruptions in brain development, individuals with CP often manifest a range of non-motor comorbidities, notably pain (74%–82%), intellectual disability (50%), speech impairment (25%), epilepsy (25%), urinary/fecal incontinence (25%), and behavioral or sleep disorders (20% to 25%).^{4,5}

This chronic condition, spanning from birth throughout one's lifespan, requires ongoing therapy services.⁶ This demand places substantial financial burdens on families and healthcare systems and leads to extended inpatient stays.^{7,8} Given its chronic nature, cognitive and speech-related comorbidities, and the unique motor impairments that can affect non-verbal communication, CP serves as a valuable model for investigating complex chronic pathologies.

Pain, with its exceedingly high prevalence, emerges as one of the main limitations to carrying out activities of daily living for individuals with CP.⁹ The complexity of accurately diagnosing pain in those with cognitive or communication impairments¹⁰ has led to a historical pattern of underestimation and inadequate treatment of pain in these individuals.^{11–14} It is posited that the self-reports from individuals with CP regarding their pain should be considered paramount in identifying the presence, intensity, and the overall effect this pain has on their daily lives.¹⁵ Consequently, initiating the pain assessment process with efforts to obtain either verbal or non-verbal self-reports of pain—utilizing alternative and augmentative communication systems, deliberate hand gestures, or nodding, for example—is recommended, even for those with cognitive impairments.¹⁶ Given that, as mentioned previously, an estimated 50% of individuals with CP are affected by intellectual disabilities,⁴ which may hinder self-reporting abilities, coupled with evidence indicating that the majority of adults with CP experience various communication disorders,¹⁰ the challenge of pain assessment becomes markedly

pronounced. This difficulty in pain identification poses a substantial challenge not only for clinical and socio-health institutions tasked with caring for these individuals but also for their families, who frequently describe pain identification as a particularly daunting task.^{17–20} Both stakeholders underscore the critical need for the development and implementation of reliable and accurate methods for evaluating pain in those unable to self-communicate their pain experiences.²¹ However, the scientific community has yet to establish a standard for pain assessment in this demographic,²² pointing out an area that urgently needs attention and improvement.

Recent years have seen various approaches to address pain assessment in non-communicative populations. These range from methods based on physiological signs to proxy observation of painful behaviors. Research has explored the utility of specific pain biomarkers such as salivary metabolites,^{23,24} brain activity,^{25,26} cardiorespiratory vital signs, skin conductance, muscle tension, or heart rate variability^{27–29} for identifying the presence of acute or chronic pain.

Observational behavioral scales are the most used tools to assess pain in this population.³⁰ However, their use is not without controversy, as they can yield subjective, observer-dependent data,^{31–33} and some may lack specificity or sensitivity.³⁴ Observers might also confuse other emotions, such as fear or stress, for pain.³¹ Further, studies comparing pain assessments in children with CP by their parents have uncovered both overestimations and underestimations of pain by parents.^{35,36} This disparity in pain assessment extends to healthcare professionals as well.³⁷

In this context, the Facial Action Coding System (FACS),³⁸ initially designed to reduce subjectivity and provide objective descriptions of facial expressions for basic emotions, can be employed to categorize pain more objectively.³⁹ However, mastering this system requires significant effort, and its microanalytic methods can be challenging to apply in routine clinical pain evaluations.⁴⁰

In order to automate the facial expression recognition, systems using deep learning (DL) approaches have led to significant advancements, particularly in the context of emotions. DL has already shown successful results in other tasks related to the identification of patterns in image-related tasks such as segmentation in medical imaging. Traditionally, human experts in the field rely on their knowledge and prior experience to perform these tasks. However, the advent of deep learning has revolutionized this process by enabling automatic pattern learning.⁴¹ Consequently, numerous studies have leveraged deep learning techniques across various tasks in medical imaging. For instance, deep learning has been applied to segmentation tasks in brain images,^{42,43} image registration,⁴⁴ and image fusion for Alzheimer's Disease/Mild Cognitive Impairment (AD/MCI) diagnosis.⁴⁵ Additionally, it has been utilized for image annotation in chest X-rays,⁴⁶ diagnosis of

brain disorders,⁴⁷ segmentation of brain tumors,⁴⁸ and analysis of microscopic images.⁴⁹ These applications highlight the versatility and effectiveness of deep learning in addressing various challenges in medical image analysis, paving the way for improved diagnosis and treatment in healthcare.

In order to automate the facial expression recognition, systems using deep learning (DL) approaches have led to significant advancements, particularly in the context of emotions, and they rely on audiovisual databases containing emotional expressions.⁵⁰ Nowadays, we find multiple datasets housing collections of facial expressions depicting pain, both as standalone expressions and within broader emotional expression datasets.^{51,52} These datasets encompass images or videos capturing spontaneous pain, such as shoulder, neck, or lumbar pain, as well as pain induced by thermal or electrical stimuli. Some even include neonates or preschool-age children receiving injections. Within these repositories, one can find images portraying at least two levels of pain intensity, and they predominantly involve healthy individuals.⁵² These datasets have paved the way for the application of artificial intelligence methods, resulting in impressive achievements in pain detection, even distinguishing between spontaneous and simulated facial pain expressions.⁵³

Automated systems grounded in the FACS, including FaceReader,⁵⁴ OpenFace,⁵⁵ AFAR toolbox,⁵⁶ or PainCheck®,⁵⁷ have been developed for pain assessment in diverse non-communicative populations, including infants,⁵⁸ and individuals with advanced dementia.⁵⁹ However, these solutions may lack the specificity required to assess pain in complex neurological pathologies, such as individuals with CP, who may exhibit motor dysfunctions affecting facial expressions of pain.

Further, to understand the decision-making processes of the AI system and explore the idiosyncrasies on how it operates with individuals with CP, eXplainable Artificial Intelligence (XAI) methods can be applied.⁶⁰ XAI techniques provide comprehensive explanations elucidating the decision-making processes and output generation of DL models. This facilitates the comprehension and interpretation of results by human users, enhances model trustworthiness, discerns causality among data variables, or informs decision taking.^{60,61} In the case of facial expression recognition, we find works applying XAI techniques to emotion recognition such as sadness or happiness^{62,63} or the work by Weitz et al. (2019),⁶⁴ whose research focused on the differences among facial expressions like anger or happiness, and pain.

The aim of this work is to improve pain assessment accuracy in individuals with complex neurological conditions, potentially revolutionizing the way we address their pain-related needs. In this research, our emphasis centers on deep neural networks trained using pain expression images. This focus stems from the inherent capability of these networks to adeptly handle the intricate challenges

associated with recognition in real-world, or “in the wild”, scenarios, as evidenced by Li and Deng (2020).⁶⁵ Our study is confined to the analysis of static images, as opposed to dynamic sequences or video data. The current study sets out to pioneer an automated facial recognition system, grounded in DL, to evaluate pain in individuals with complex neurological pathologies, specifically adults with CP. The proposed DL system will undergo training using a variety of existing pain datasets that capture diverse pain conditions. Additionally, a built purpose-designed dataset tailored for individuals with CP, denoted as CP-PAIN, will be used to evaluate the system’s effectiveness.

Subsequently, the automated system’s pain scores will be compared with evaluations by clinicians employing three commonly used observational scales within the CP population: The Wong Baker FACES® Pain Rating Scale,⁶⁶ the Facial Action Coding System (FACS), and The Non-Communicating Adults Pain Checklist (NCAPC).⁶⁷

Finally, to delve deeper into the mechanisms underlying DL techniques, we include eXplainable Artificial Intelligence (XAI) techniques to understand the rationale behind the outcomes of the pain perception mechanisms employed by DL models and explore potential commonalities among diverse trained DL models, especially when used with people with CP.

The work is organized as follows. Section 2 describes the methods used to build the dataset of pain expressions in individuals with CP and the automated pain recognition system. Section 3 describes the results achieved and the last Section discusses the main findings and concludes the work.

Methods

Within this section, we describe the methodology encompassing the construction, labelling, and evaluation of a dataset of pain expressions in individuals with CP (CP-PAIN), as well as the training phases of the pain recognition artificial intelligence (AI) system.

Constructing and assessing the CP-PAIN database

Prior to the construction of the CP-PAIN database, we adhered to a robust ethical framework. Written informed consent, encompassing the use of images for informative purposes, was obtained. This process involved participants with CP who had the legal capacity to provide consent ($n = 10$), as well as legal representatives of remaining participants ($n = 43$). The documentation was thoughtfully written in standard form and an accessible easy-to-read format to ensure its comprehension by as many participants as possible.

Ethical compliance and approvals. This research meticulously adhered to the principles outlined in the Declaration of Helsinki (1991). The research protocol received approval from both the ASPACE Foundation's ethics committee and the Research Ethics Commission of the Balearic Islands (protocol number IB4046/19), affirming its ethical rigor.

Participants. The study extended invitations to all users diagnosed with CP, or their legal representatives, affiliated with the Adult Life Promotion Services of the Cerebral Palsy Association (SPACE) in the Balearic Islands (Majorca, Spain) and Toledo (Castilla-La Mancha, Spain). Participants were exclusively selected based on two criteria: a diagnosis of CP and being over 18 years old. The only exclusion criterion was the lack of consent for the use of personal images by the institutions involved. A total of 53 individuals (mean age = 37.57 (9.88) years, age range = 21–69 years, including 19 females) agreed to participate. Subsequently, they or their respective legal representatives formally filled in the informed consent.

Table 1 displays the clinical characteristics of the 53 participants with CP.

Measures. In addition to gathering sociodemographic and clinical data (e.g., age, sex, type of CP, level of motor, and communication impairment) from medical records, the study implemented the following measures:

Observational scales. 1. The Facial Action Coding System (FACS): Designed to minimize subjective judgments in assessing facial activity, FACS is a widely employed system for coding emotional facial expressions in scientific studies.⁷⁰ It has been successfully applied to individuals with communication disorders and CP,⁷¹ and a recent systematic review and meta-analysis concluded that it is the preferred scale for individuals with CP who have communication impairments.³⁰

FACS dissects facial expressions into 44 individual components of muscle movement, known as Action Units (AUs), rating them on a 6-point scale (0 = no expression, 5 = extreme expression). Pain is identified through six specific AUs: lowering of the eyebrows (AU4), elevation of the cheeks and compression of the eyelids and/or contraction of the cheekbones (AU6/AU7), wrinkling of the nose and/or raising the upper lip (AU9/AU10), and closing the eyes (AU43).^{40,72}

The total pain score is calculated as follows: Pain score = AU4 + (AU6/AU7) + (AU9/AU10) + AU43, yielding a 20-point scale.

2. The Non-Communicating Adults Pain Checklist (NCAPC): This 18-item scale assesses pain behaviors through six components: vocal response, emotional response, facial expression, body language, protective responses, and physiological responses. Derived from the Non-

Communicating Children Pain Checklist (NCCPC), the NCAPC offers optimal utility irrespective of the evaluator's familiarity with the individual.⁷³ The NCAPC has demonstrated strong psychometric properties and the capacity to identify pain and its intensity in adults with intellectual and developmental disabilities.⁶⁷ Evidence points to this scale, along with FACS, as the optimal ones for assessing pain in adults with CP.⁷⁴

3. The Wong Baker FACES® Pain Rating Scale: This scale rates pain on a scale from 0 (no pain) to 10 (worst possible pain) by comparing the patient's facial expression to the provided scale images. While commonly used in pediatric populations, it has also found utility applied to individuals with disabilities and communication disorders.⁷⁵

As depicted, these scales, especially FACS and NCAPC, were selected over others due to their specific relevance, applicability, and proven ability to address the unique challenges of assessing pain in adults with CP, particularly those facing communication barriers. The selection of the Wong Baker FACES® Pain Rating Scale was primarily driven by its ease of use and the consideration that it is a tool for individuals with CP and cognitive impairment could utilize to rate their level of pain. This combination of scales allows for a comprehensive and nuanced approach to pain evaluation in this specific population, leveraging the strengths of each tool and addressing the particular needs of our study cohort.

Procedure for data collection. Video recordings capturing facial and body expressions were conducted in situations where participants with CP either underwent potential painful procedures or when caregivers identified signs of pain in other care procedures (e. g. feeding, personal hygiene, assistive...). For scheduled painful procedures such as therapies or intramuscular injections, the video recording initiated 2 min prior to the potentially painful stimuli and continued for 2 min thereafter. In cases where participants had the cognitive capacity (n = 24, resulting in a total of 43 recorded expressions), they were kindly asked to self-rate their pain on a scale of 1 to 10 using the Wong-Baker Faces pain rating scale.

Efforts to reduce potential biases in the video recordings were meticulously implemented. Standardized protocols, with instructions for camera placement, lighting, and the duration of the recordings were established to minimize variation in recording conditions. Moreover, observers involved in the video recording underwent training including pain recognition, and the recording equipment and protocol, in order to homogenize the observers' understanding and approach. In addition, observers were blinded (when possible) to the participants' medical histories to avoid preconceived notions influencing the recording process. Finally, video recordings were independently review by evaluators who were not involved in the recording sessions.

Table 1. Clinical characteristics of participants.

GMFCS	n	%	CFCS	n	%	CP Subtype	n	%
Level I	0	0	Level I	6	11.3	Spastic	42	79.2
Level II	5	9.4	Level II	7	13.2	Dyskinetic	3	5.7
Level III	1	1.9	Level III	11	20.8	Ataxic	1	1.9
Level IV	17	32.1	Level IV	14	26.4	Mixed	7	13.1
Level V	19	35.8	Level V	15	28.3			

CP: cerebral palsy; GMFCS: Gross Motor Function Classification System,⁶⁸ CFCS: Communication Function Classification System.⁶⁹ These scales classify the person into five levels of function, lower scores indicating lower impairment of function.

Throughout the study duration, a comprehensive total of 127 recordings were successfully acquired. These recordings depicted various sources of pain in the painful images, which were classified as follows:

- Intramuscular injection: 77 images (60.6%)
- Muscular stretching: 32 images (25.2%)
- Other sources of pain: 18 images (14.2%)

Expert evaluation of video recordings. All video recordings underwent meticulous offline evaluation by two highly experienced physiotherapists, with more than 10 years of expertise in treating individuals with CP. These evaluators independently applied the three observation scales: the Wong-Baker Faces pain rating scale, NCAPC, and FACS.

In this study, efforts to reduce potential biases in the physiotherapists' evaluations were meticulously implemented. Recognizing the inherent subjective biases in traditional observational scales, thorough training was provided to the two clinicians involved to standardize their application of these scales, aiming for a consensus in their evaluations to minimize individual subjective interpretations. Furthermore, by using three different observational scales we aim to provide a comprehensive view of pain assessment and to mitigate the limitations inherent in any single scale. Additionally, to further prevent bias, the design of the study ensured that evaluators using traditional scales were blinded to each other's results, promoting independent evaluations free from preconceived expectations about the pain scores. Rigorous statistical analysis, specifically using the Intraclass Correlation Coefficient (ICC) model 2,k, was conducted to assess inter-rater reliability among the clinicians (MSP, MSR, MSE k, and n used in the equation to refer, respectively, Mean Square Persons, Mean Square Raters, Mean Square Error [from a two-way ANOVA], and number of persons).⁷⁶

$$ICC(2, k) = \frac{MSP - MSE}{MSP + (MSR - MSE) / n}$$

Physiotherapists might be more accurate recognizing pain expressions resulting from their own interventions, such as muscle stretching, which they have observed on a regular basis, rather than pain expressions caused by procedures less related to their profession, such as intramuscular injections. Thus, the ICC was separately calculated for each type of painful stimulus, in order to ascertain whether the evaluator's familiarity with the particular painful procedure or situation influenced the agreement.

The interpretation of the ICC scores followed the categories proposed by Fleiss (1986):⁷⁷ low agreement (ICC < 0.40), fair/good agreement (ICC 0.41 to 0.75), and excellent agreement (ICC > 0.75).

For a precise comparison with the AI system's assessments, we rigorously applied a mathematical process to reclassify the mean scores from each observational scale into binary outcomes. Specifically, a "0" score was designated to signify "no pain", whereas any score above 0 denoted the detection of pain. Consequently, an image was categorized as "no pain" only if there was a unanimous '0' score consensus among all evaluators. For any deviation from this consensus, the image was systematically labelled as "pain". This systematic categorization facilitated a clear and unbiased distinction between pain and no pain expressions based on quantifiable measures.

Challenges with self-reports. Despite our initial efforts to collect retrospective self-reports from participants following the video recordings, we encountered notable challenges, primarily due to cognitive and attentional issues experienced by the participants. Consequently, only a limited number of self-reports, totaling 10, could be successfully gathered. The challenge of obtaining only 10 self-reports, however, significantly impacted our methodology, compelling us to abandon the initially proposed comparative approach. This adjustment underscores the broader difficulties inherent in pain assessment for individuals with CP, and reflects the critical need to adapt research methodologies to effectively address these complexities.

Deep learning pain recognition

Datasets. Aiming at building a pain recognition system applicable to images of individuals with CP, we merge three extensively used databases: the UNBC-McMaster Shoulder Pain Expression Archive Database,⁷⁸ the Multimodal Intensity Pain dataset (MInt PAIN),⁷⁹ and the Delaware Pain Database.⁵¹ For the sake of brevity, we will subsequently denote this merged dataset as PAIN-DB.

The UNBC-McMaster Shoulder Pain Expression Archive Database comprises 25 adult patients afflicted with shoulder pain. It features a collection of 200 distinct range of motion tests, encompassing both affected and unaffected limbs. Data acquisition involves videos capturing facial expressions, albeit in low resolution, which also include social interaction and verbal communication. Annotations include self-report measurements via Visual Analog Scales (VAS) encompassing sensory and affective aspects, along with pain intensity assessed by both self-report and observer (Observer-Assessed Pain Intensity, OPI). Moreover, annotations encompass limb information (affected/unaffected) and FACS coding.

The MInt PAIN Database presents a collection of images obtained through electrical muscle pain stimulation, involving 20 subjects. Each subject participated in two trials during data acquisition, with each trial encompassing 40 pain stimulation sweeps. Within these sweeps, two types of data were captured: one representing the absence of pain, and the other portraying pain at four varying intensities.

The Delaware Pain Database is an extensive compilation of fully characterized photographs featuring 127 female and 113 male subjects, with an emphasis on painful and neutral expressions. The dataset's primary hallmark lies in its remarkable diversity across dimensions of race, gender, and expression intensity.

Upon the integration of the three datasets, a relabeling into two classes was performed: images featuring pain

and images devoid of pain. Consequently, images encompassing varying degrees of pain were grouped within the first class, while the remaining images were categorized under the second class. This reduction served two primary objectives: foremost, the normalization of data across the datasets, aligned with the two-class structure in the Delaware Pain Database; and secondly, the transformation of the task into a binary classification problem.

Furthermore, we did not use all available frames from UNBC-McMaster and MInt PAIN to avoid overfitting on the users they contain. Since there is little variation in consecutive frames, instead of using them all, a sample of twenty frames per user and class was taken. The final PAIN-DB dataset, composed by images from the three described databases, is summarized in Table 2. Figure 1 depicts some example images from each database.

To assess the efficacy of the PAIN-DB-trained pain recognition system on individuals with CP, a testing dataset was curated from video recordings of study participants with CP, designated as CP-PAIN. By using pin-pointed moments of pain expression as determined by one physiotherapist specialized in neurological rehabilitation, two frames were extracted from each video: one capturing the moment of pain manifestation and another from a non-pain moment, thus ensuring a balanced dataset composition. Regrettably, a subset of videos proved unsuitable for inclusion due to two primary reasons: substantial occlusions obscuring facial features during pain instances, and participants wearing surgical masks that obscured half of the face. Although human observers can bypass these occlusions to discern pain expressions, our automated system lacked training for such scenarios, which could introduce unpredictable outcomes. Consequently, these videos were omitted. The resulting CP-PAIN dataset comprised 109 images, representing a dataset suitable for evaluation. Two example images from this dataset can be found on Figure 1, in the rightmost column.

Table 2. Breakdown of PAIN-DB dataset into the datasets forming it.

Dataset	Users	Images	Levels	Resolution (width × height)
MInt	20	800	5	1920 × 1080
Delaware	240	803	2	5152 × 3864
UNBC-McMaster	25	980	5	320 × 240
Total	285	2.583	2	Mixed

Levels column refers to the number of pain levels used to label the images of the dataset. Notice that not all images from the MInt PAIN and the UNBC-McMaster datasets were used, and that the final dataset contains only two classes: pain and no pain.

Image preprocessing. We conducted a pre-processing phase to enhance the performance of the neural networks, which comprised cropping and background subtraction.

Images were cropped into a squared region containing the face (similar to Haque et al. (2018) approach⁷⁹). This was achieved through the application of the well-established multitask cascaded convolutional networks for face detection.⁸⁰

Furthermore, the background subtraction was accomplished through the integration of U2-Net,⁸¹ which facilitates precise object segmentation. In our case, we used their pretrained weights that had been fine-tuned for human segmentation. Following the extraction of a binary mask outlining the background pixels, these regions were replaced with white color, thereby retaining only the human subject within the image frame.



Figure 1. Example of images with and without pain, from the four datasets employed in this study.

The combined cropping and background subtraction procedures were devised to streamline the pain recognition task for the neural networks. This dual process serves both to eliminate any interference introduced by factors other than the person's facial features and to standardize the images. This standardization ensures that the facial attributes consistently occupy an approximate spatial alignment. Finally, the resultant images were resized to the input dimensions required by the distinct network architectures utilized in our study. Examples of the preprocessed images for each class and dataset are shown in Figure 2.

Models. We built ten neural networks aiming at recognizing pain from images (see Table 3). For the sake of simplicity, we denoted the models introduced in Song et al. (2014),⁸² Li et al. (2015),⁸³ and Ramis et al. (2022),⁶³ as SongNet, WeiNet, and SilNet, respectively.

All the employed networks share a foundation in convolutional architecture; however, they diverge in terms of architectural constituents, including the arrangement of convolutional layers, pooling layers, fully connected layers, and other elements. This architectural disparity is reflected in the count of parameters employed by each network. The input size of the networks is 224×224 , with the exception of WeiNet and SilNet, which is 64×64 and 150×150 respectively.

Six out of the ten models—VGG16, VGG19, ResNet50, ResNet101V2, Xception, and InceptionV3—were initialized with pre-trained weights sourced from the ImageNet dataset.⁸⁹ For these models, final fully connected layers were replaced with an average 2D pooling layer and a new fully connected layer with the appropriate number of outputs: two, pain and no pain. The remaining models—AlexNet, SongNet, WeiNet, and SilNet—were trained from scratch.

By integrating these varied architectures, we can compare their performance and analyze the impact of architectural selections concerning tasks related to pain recognition.

The models were implemented using the Python programming language, harnessing the capabilities of the Keras library. In the case of the AlexNet, WeiNet, SongNet, and SilNet models, we constructed each layer sequentially within the Keras framework. Conversely, the remaining models were conveniently available through the Keras API, inclusive of their pre-trained weights sourced from the ImageNet dataset.

Metrics. To assess the performance of the different models, common metrics used when evaluating classification tasks, namely accuracy, precision, recall, and F1-score, were used (TP, TN, FP, and FN used in the equations to refer, respectively, to the number of true positives, true negatives, false positives, and false negatives):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F}_1 \text{ score} = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

- Accuracy (equation (1)): it measures the proportion of correctly classified instances among the total instances in the dataset. It provides an overall view of the model's correctness. However, it might not be suitable when dealing with imbalanced datasets where one class dominates, as it can be misleading.

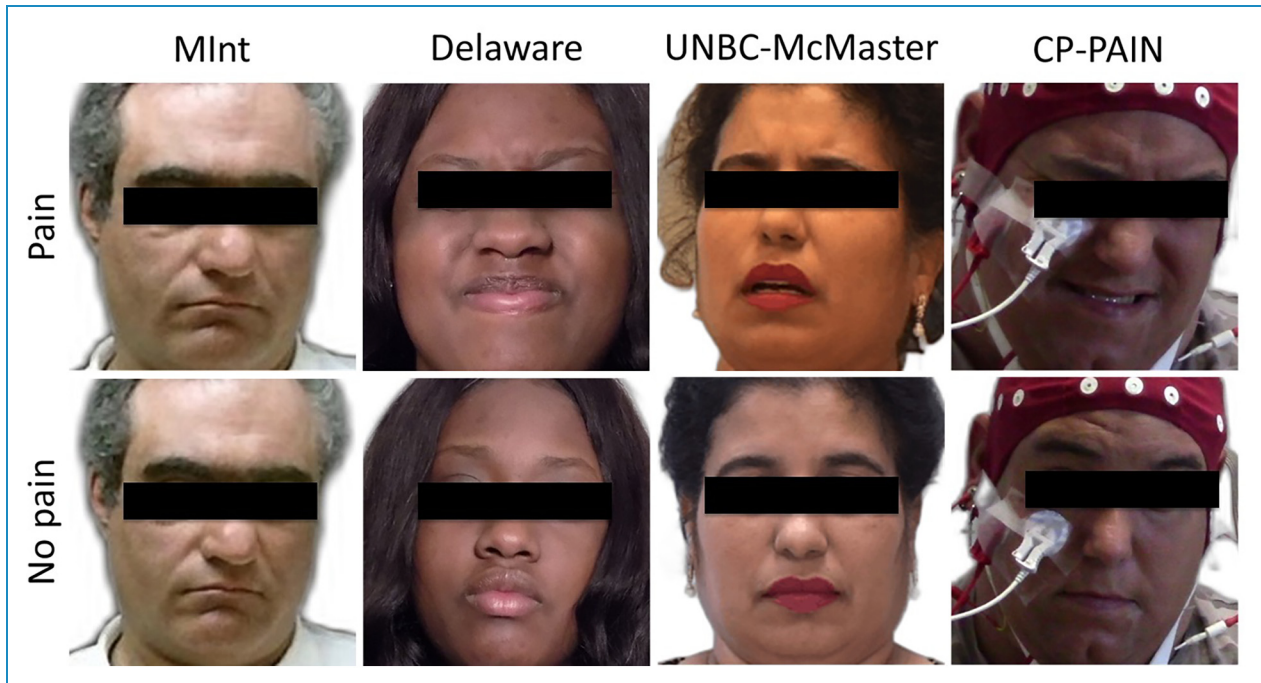


Figure 2. Resulting images after the pre-processing performed before any training or testing on the four databases, featuring a face crop and background subtraction.

Table 3. List of network architectures used in this study for pain recognition.

Work	Model/s	Parameters
Krizhevsky et al. (2012) ⁸⁴	AlexNet	88.7 M
Song et al. (2014) ⁸²	SongNet	172.7 K
Li et al. (2015) ⁸³	WeiNet	1.7 M
Simonyan and Zisserman (2015) ⁸⁵	VGG16	14.7 M
Simonyan and Zisserman (2015) ⁸⁵	VGG19	20 M
He et al. (2015) ⁸⁶	ResNet50	23.6 M
He et al. (2015) ⁸⁶	ResNet101V2	42.6 M
Szegedy et al. (2015) ⁸⁷	InceptionV3	21.8 M
Chollet (2017) ⁸⁸	Xception	20.9 M
Ramis et al. (2022) ⁶³	SilNet	184.9 M

- **Precision** (eq. (2)): it is a measure of how many of the instances predicted as positive by the model are actually true positives. It focuses on the correctness of positive predictions. High precision indicates that the model is careful in labeling instances as positive.

- **Recall** (eq. (3)): it quantifies how many of the actual positive instances were correctly predicted by the model. It emphasizes the model's capability to capture all positive instances.
- **F1-Score** (eq. (4)): it is the harmonic mean of precision and recall. It combines both precision and recall into a single value, providing a balanced assessment of a model's performance. It is particularly useful when dealing with imbalanced datasets or situations where both false positives and false negatives are of concern.

When evaluating a model's performance in image classification, these metrics collectively provide a comprehensive understanding of its strengths and weaknesses. Accuracy gives a global view of performance, precision focuses on correct positive predictions, recall emphasizes capturing all true positives, and F1-score offers a balanced perspective considering both precision and recall. Using this set of metrics ensures a nuanced assessment of a model's effectiveness in classifying images accurately and reliably, even in scenarios where data distribution might be uneven or where the cost of false positives and false negatives differs significantly.

Models' explanation. We applied the model-agnostic XAI technique known as Local Interpretable Model-agnostic Explanations (LIME).⁹⁰ LIME operates by introducing perturbations to the image under examination, thereby generating multiple modified versions of the image. These

perturbed samples are subsequently processed through the model, and the resulting prediction outcomes are employed to establish the relevance of each region within the image with respect to a specific class. To further understand the classification processes employed by the models, we generalized the local explanations derived from a subset of samples to formulate global explanations for each class. This approach followed the methodology introduced in Manresa-Yee et al. (2023),⁹¹ where local explanation masks are transformed into a normalized space and then averaged to produce a comprehensive heatmap that delineates the significance of individual facial features in the classification process. For a better comprehension of the explanation process, Figure 3 provides a visual representation of the distinct steps involved.

Procedure. All trainings for the ten networks were performed on a computer featuring an NVIDIA 4090 GPU and an i9 9900KF CPU, generously supplied by the Universitat de les Illes Balears. This same hardware configuration was consistently employed for the evaluation phase. Overall, we did not encounter any notable challenges or limitations with the selected hardware configuration. It adequately supported the execution of our experiments without impeding our ability to achieve meaningful results.

Following a series of preliminary experimental iterations, we conducted several training sessions, each comprising 50 epochs, on a training–validation partition of the dataset. Through this process, we systematically varied the learning rate across three orders of magnitude (0.01, 0.001, and 0.0001) and assessed the resulting performance on the validation set. Based on these evaluations, we determined that a learning rate of 0.001 yielded the highest accuracy. Similarly, we monitored the training and validation loss curves across epochs to gauge the model’s convergence and generalization capabilities. After observing diminishing returns beyond 30 epochs, we determined that this value struck a favorable balance between performance and training time. Additionally, continuing training beyond this point did not yield significant improvements in validation loss, indicating the risk of overfitting. Additionally, data augmentation layers were incorporated into the training pipeline, encompassing randomized processes such as rotation, mirroring, and contrast adjustment. This integration aimed at introducing invariance to these intrinsic properties while concurrently augmenting the spectrum of image variations during training.

Regarding the utilization of pretrained weights, we adopted a cautious approach to mitigate the risk of overfitting, given the relatively small size of our training dataset. We experimented with various fine-tuning strategies, including freezing lower blocks of the models and training only the last layers, training the entire model, and a combination of both. Surprisingly, freezing the lower blocks did not yield performance improvements in our case; instead,

it hindered the convergence speed during training. As a result, we opted to fine-tune the entire model on our dataset and task, with no frozen layers left.

The experimental design comprised a total of three training scenarios for each of the ten models detailed in the Models Section. Further, we employed K-cross validation with $K = 5$ to ensure more stable and reliable results, for each of the three training scenarios, 150 trainings were performed in total. Regarding the choice of $K = 5$ in the K-cross validation, this decision was made to strike a balance between the number of training iterations and the reliability of the results, providing a reasonable approximation of model performance while avoiding excessive computational overhead. Overall, the K-cross validation approach enhanced the stability and reliability of our results by mitigating the variability that can arise from different data partitions or training iterations.

Initially, the merged PAIN-DB dataset was partitioned into training and testing subsets, maintaining an 80%–20% and with varying test subsets, according to the K-cross validation procedure. This training phase involved models being trained on nearly all users’ data, thereby assessing their performance in predicting pain based on images from users they had previously encountered. This training configuration aimed to yield the most optimal outcomes among the three.

The second training iteration closely resembled the initial one, using the PAIN-DB dataset. However, in this iteration, the partitioning was user-centric rather than image-centric, while preserving the same train-test ratio. Here, models were trained using images from a subset of users, with the remaining users reserved for the testing phase. This approach sought to evaluate the models’ ability to accurately perceive pain within images of users unseen during training.

Finally, a training with the entire PAIN-DB dataset was conducted, testing on the pain images collected from individuals of the CP-PAIN dataset. The primary distinction here was that models were evaluated on users external to the PAIN-DB dataset. For this particular training scenario, there was no K-cross validation, but rather five different trainings on the same data, since the testing set is external to the PAIN-DB dataset. This third scenario aimed to validate the extrapolation of knowledge gained from the PAIN-DB dataset to the CP-PAIN dataset, which presented additional challenges due to its external nature and inclusion of users with cerebral palsy. The main challenge with the CP-PAIN dataset was its limited size, which precluded direct training on it. Instead, we sought to generalize the models trained on the PAIN-DB dataset to the CP-PAIN dataset.

The three-tiered approach to experimentation was devised to sequentially validate network performance on well-established pain recognition datasets. This stratification allowed for enhanced assessment of model effectiveness on a novel dataset, the size of which precluded fine-tuning strategies.

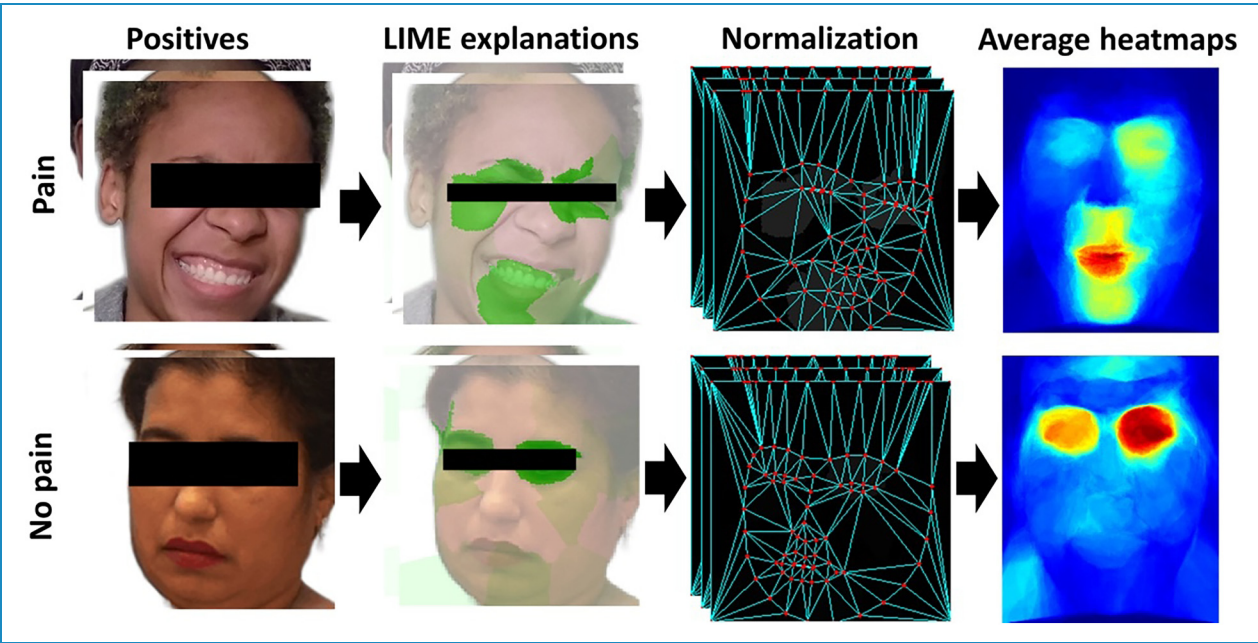


Figure 3. Visualization of the explanation process followed for the identification of the important regions for a specific model to classify into a class.

Results

Inter-rater agreement among physiotherapists

In the course of the study, both physiotherapists conducted a grand total of 762 assessments, all stemming from the scrutiny of 127 video recordings. The inter-rater agreement, measured through the ICC, exhibited the following hierarchy of agreement levels for the observational measures:

- FACS = 0.751 (excellent)
- Wong-Baker Faces pain rating scale = 0.639 (fair/good)
- NCAPC = 0.551 (fair/good)

For a more comprehensive understanding of the inter-rater agreement contingent on the type of painful stimulus, refer to Table 4.

These findings show a higher ICC across all scales when pain was induced by muscular stretching (mean of 3 scales = .825, excellent ICC), followed by an unknown source of pain (mean of 3 scales = .632, fair ICC), and lastly, intramuscular injection (mean of 3 scales = .522, fair ICC). Given that FACS exhibited the highest level of agreement, it was employed to label the image into the “pain”/“no pain” classes.

Deep learning pain recognition

The outcomes derived from the three distinct training scenarios are presented in Figure 4. As expected, the

Table 4. ICC of observational measures and pain/no pain measurement proportion.

Source of pain	Measure		
	Wong-Baker	NCAPC	FACS
Intramuscular injection	0.412	0.590	0.565
Muscular stretching	0.845	0.739	0.891
Other	0.661	0.469	0.764
Videos rated as “no pain”	20	19	16
Videos rated as “pain”	107	108	111

NCAPC: Non-Communicating Adults Pain Checklist; FACS: Facial Action Coding System; Wong-Baker: the Wong-Baker Faces pain rating scale.

initial scenario, where networks were evaluated on users they had encountered during training, yielded the most favorable results overall. Notably, upon assessment with previously unencountered users, a marginal reduction was observed in both accuracy and F1 score across most networks. The third scenario, involving testing on users from the distinct CP-PAIN dataset, led to slightly diminished metrics. Among the models assessed, merely three attained performance levels surpassing 70%: InceptionV3, ResNet101V2, and Xception. Remarkably, InceptionV3 exhibited the most promising outcomes on the CP-PAIN dataset, attaining an accuracy of

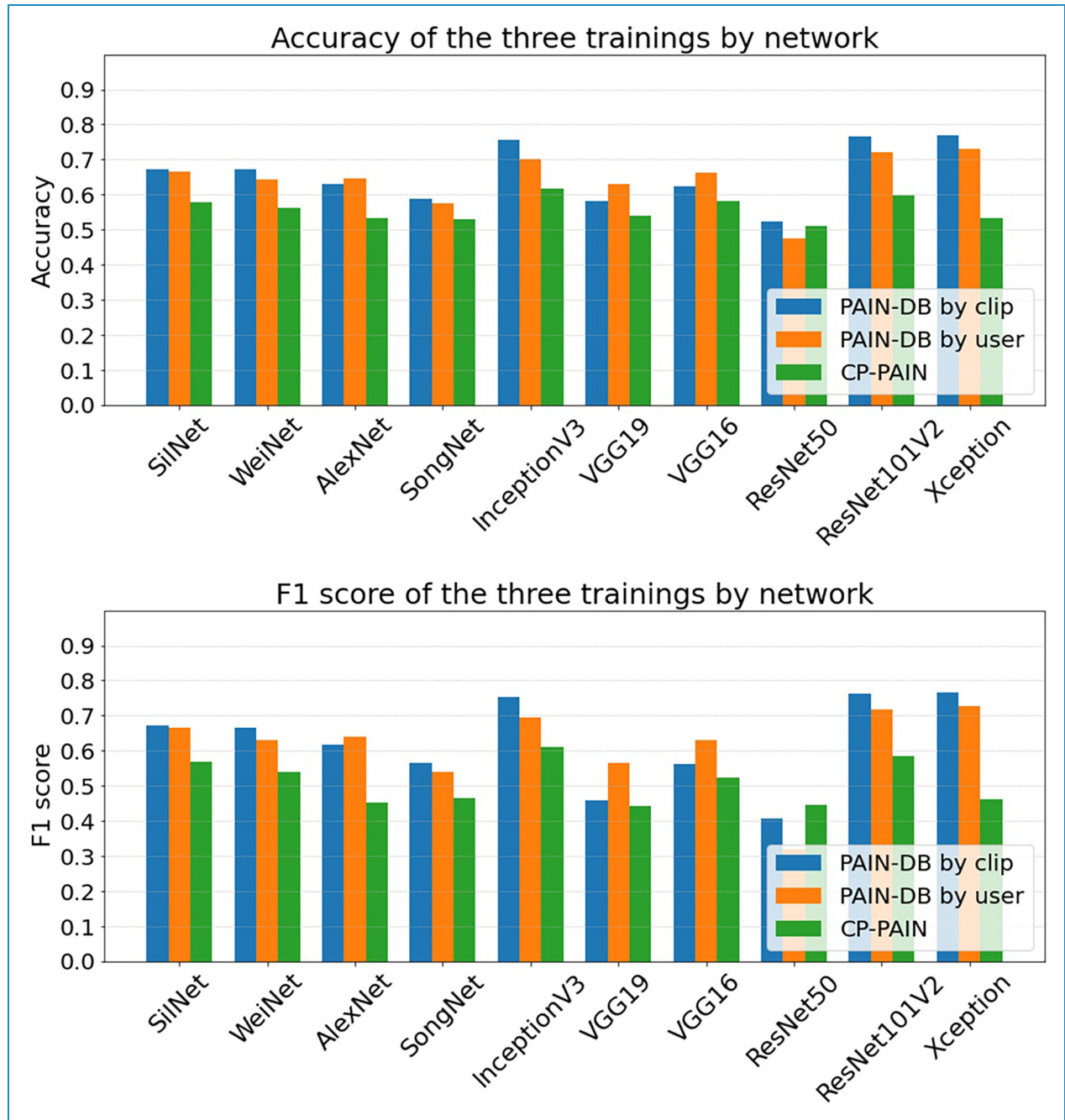


Figure 4. Accuracy (top) and F1 score (bottom) of the pain prediction results of the three different trainings performed, displayed by network. The values correspond to the average between the five validation splits, for each training scenario.

62.67% alongside an F1 score of 61.12%. Examples of the pain recognition task of this model on the different datasets, including true positives and false negatives, can be found in the appendix section. We expect these examples to provide a more tangible understanding of the model's capabilities.

Figure 5 illustrates precision and recall values for each class (pain and no pain) on the CP-PAIN dataset. The

precision values across networks display a relatively balanced distribution between classes. Interestingly, the recall values, which highlight the model's aptitude for identifying specific classes, underscore a consistent trend toward pain prediction over no pain prediction for all networks. This trend is particularly pronounced in certain models such as AlexNet, VGG19, ResNet50, and Xception, which achieve a recall of 80% or higher for pain

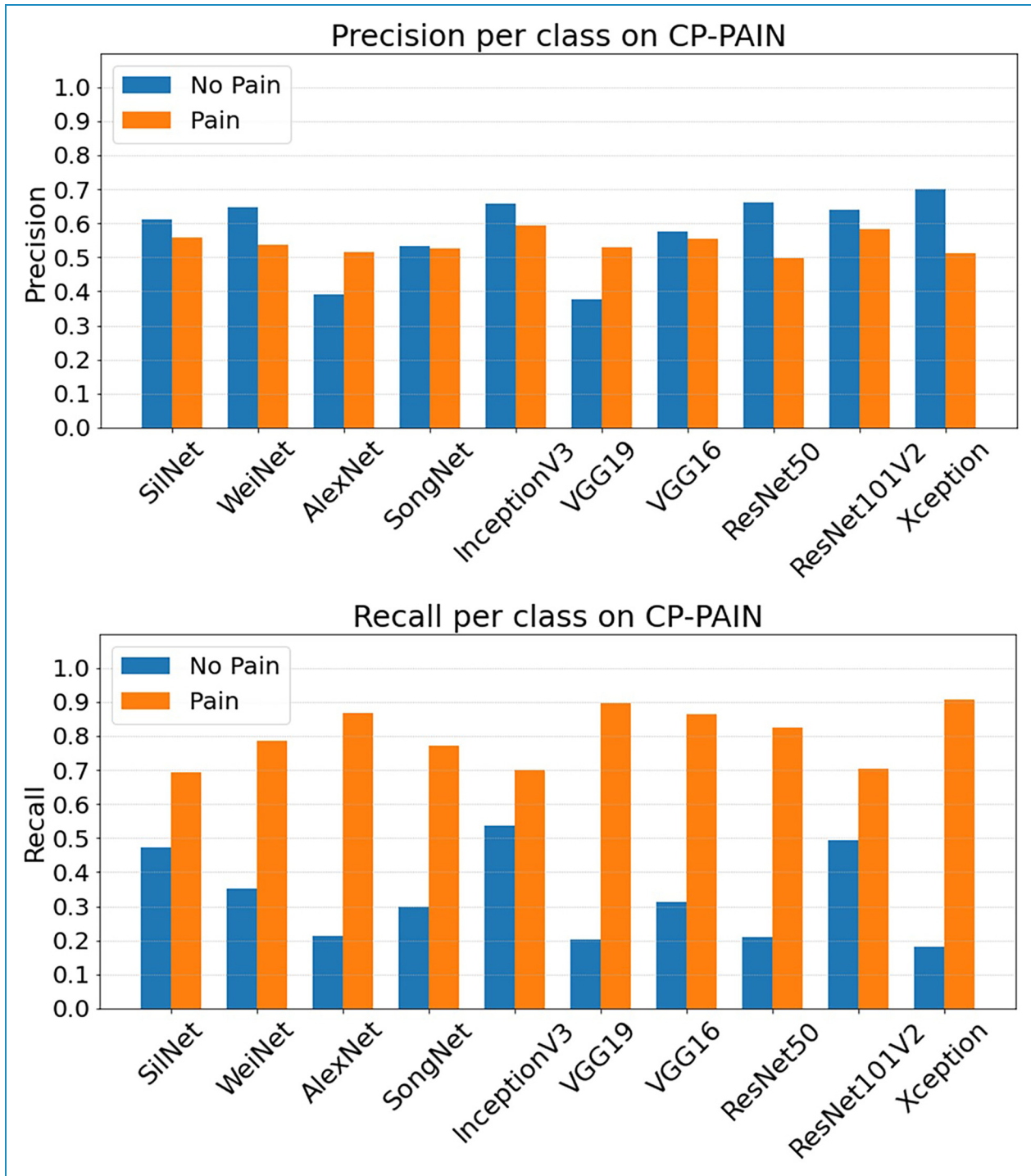


Figure 5. Precision (top) and recall (bottom) of the pain prediction results by class on the CP-PAIN dataset, displayed by network. The values correspond to the average between the five validation splits, for each training scenario.

but exhibit figures of 20% or lower for no pain instances. In contrast, InceptionV3, ResNet101V2, and SilNet exhibit more balanced recall values. This balance is further translated into higher F1 scores, as depicted in the lower chart of Figure 4.

Figure 6 presents heatmaps that illustrate the significance of distinct facial regions in predicting a specific class for each of the trained models. Evidently, there exists subtle divergence among the models regarding the specific facial regions they emphasize when discerning

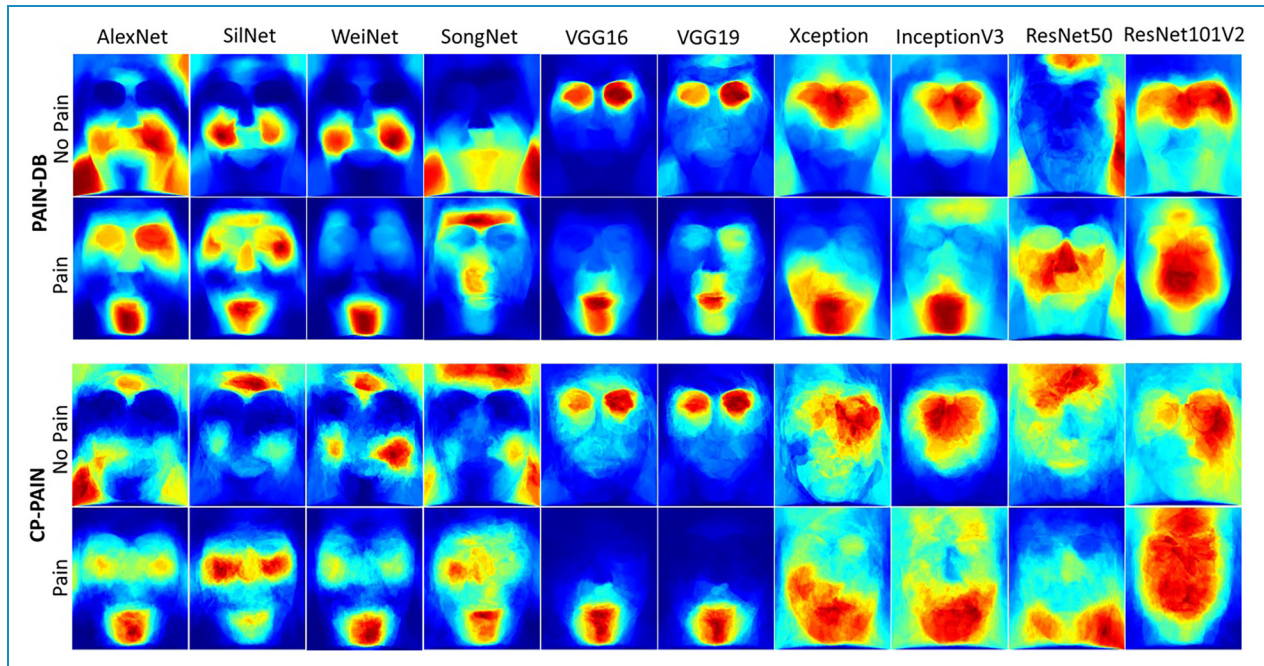


Figure 6. Heatmaps representing face regions importance for the different models, classes, and databases.

the presence or absence of pain. However, noticeably, the heatmaps from models with pre-training and those without pre-training appear to segregate into two distinct clusters. This observation suggests a proclivity for models within each category to prioritize particular facial regions to a greater or lesser extent during the classification process.

In an ideal scenario, the models should exhibit similar heatmaps when applied to both the PAIN-DB and CP-PAIN datasets. However, as depicted in Figure 6, certain models deviate from this ideal alignment. SilNet, WeiNet, and SongNet prominently emphasize the forehead region when classifying instances as “Not pain” in the CP-PAIN dataset, a pattern not evident in the PAIN-DB dataset. This disparity underscores the substantial impact of database differences on model predictions. Similarly, ResNet50 also demonstrates distinct results by focusing on the lower facial region for pain recognition in CP-PAIN, while primarily centering on the central facial area in the case of the PAIN-DB dataset. In contrast, the remaining models appear to exhibit relatively consistent reliance on the same facial regions, irrespective of the database under consideration. Another noteworthy observation pertains to the importance of the lower portion of the face, particularly encompassing the mouth and its vicinity, which emerges as a pivotal factor in pain recognition. Conversely, the absence of pain is predominantly associated with the upper facial region, particularly focusing on the eyes and their surroundings.

Discussion

The primary objective of this study was to develop an automated facial recognition system based on DL for the

assessment of pain in adults with CP. To achieve this goal, we developed and trained this system using a specific dataset of images of individuals with CP (CP-PAIN) curated for this very purpose and three existing well-known pain databases. Subsequently, we compared pain scores obtained from the automated facial recognition system with the pain scores provided for each image of the CP-PAIN dataset, obtained by consensus by two independent physiotherapists experienced in caring for individuals with CP. Our findings revealed a 60% accuracy rate for the facial recognition system, thus confirming the feasibility of adapting pain detection from images for patients with CP.

To the best of our knowledge, this study marks a pioneering initiative introducing pain detection through image analysis for individuals with CP. This innovation has the potential to offer valuable solutions for evaluating pain within this population, a challenge often noted by clinicians and family members alike.²¹ Beyond addressing this pressing need, the application of image-based pain detection has the added advantage of mitigating the intrinsic subjectivity inherent in human pain assessments,^{31–33} thereby reducing associated inaccuracies. Our findings reflect a noteworthy aspect of this subjectivity. Even in cases where the evaluator has a deep familiarity with the individual with CP under assessment, our results revealed the presence of a subjective bias that depends on the evaluator’s level of familiarity with the specific painful procedure being evaluated. To this regard, physiotherapists exhibited a higher level of agreement when assessing pain induced

by muscle stretching compared to pain resulting from intramuscular injection or an unknown source of pain. This subjectivity in pain assessment has been demonstrated in other studies that measured the degree of agreement between individuals with cerebral palsy and their parents,^{35,36} as well as with healthcare professionals.³⁷ This highlights the importance of objective, technology-assisted approaches in enhancing the precision and objectivity of pain assessment in individuals with CP.

Given the scarcity of relevant patient data for training, we devised a strategy involving the training of pain detection models on established datasets tailored for this task. Subsequently, the models' performance was evaluated on a minimal test set, CP-PAIN, constructed specifically for this study, which comprises images of users with CP, both with and without pain. Despite the limited number of images contained in CP-PAIN, this dataset is the first of its kind. Cognitive factors and motor impairment may affect the gestures, body movement, and mimics in individuals with neuromotor disorders, such as individuals with CP, what could lead to idiosyncratic pain expressions or mask pain of low intensity.^{71,92–94} Our approach incorporates a specific dataset of pain in this population that may help enhance DL pain recognition. Further refinement and expansion of our dataset can harness the power of DL to create a more robust and reliable pain detection system. The need for a larger and more diverse image dataset becomes evident as we strive to train our DL model to better recognize nuanced pain expressions in this unique population. Although limited and prone to expansion and quality improvement, the CP-PAIN database could be a first contribution to the DL analysis of pain expressions in complex populations with pain facial expressions diverse from those of the general population,^{71,92–94} collected in other existing databases.

Initially, the performance achieved on the training set reached a maximum of 70% for users unknown to the models. Consequently, this outcome can be considered an upper-bound estimation for the performance attainable on CP-PAIN. Acknowledging this modest performance, likely attributed to the limited training data, we plan to enhance these results in future work by diversifying the training datasets. Additionally, we emphasize the necessity for meticulous analysis of images from each database, particularly those extracted from video recordings, since the imprecision in annotations during video frame selection could introduce a significant number of erroneously labeled images, warranting careful consideration.

Our findings on the CP-PAIN dataset unveiled 60% accuracy, thereby establishing the viability of adapting pain detection from images to patients with CP. A noteworthy trend emerged where the models exhibited a pronounced trend towards detecting pain in images, surpassing instances of identifying no pain. This proclivity could be attributed to the frequent appearance of subjects in

the test images exhibiting open mouths and exposed teeth, a phenomenon potentially associated by the networks with pain expressions seen during training. This accuracy in pain detection, while seemingly modest, marks a significant milestone in clinical research, as it underscores the potential for developing a specialized assessment model tailored to individuals with complex neuromotor conditions, a development that holds promise in reducing the inherent subjectivity and biases often associated with human evaluations of pain.^{20–22}

Comparing the results obtained in this study with existing literature in terms of measurable performance is challenging due to several factors.⁹⁵ Firstly, different datasets were used, making direct comparison difficult. Secondly, this study does not exploit temporal information, which was considered in other works. Thirdly, many works aim for pain level classification, rather than solely classification between pain and no pain in images. No other work employing the same four datasets (UNBC-McMaster Shoulder Pain Expression Archive Database, Mint PAIN Database, Delaware Pain Database, and CP-PAIN) as this study has been found. Typically, existing studies focus on testing new models or strategies and optimizing performance on a single dataset rather than generalizing results to unseen datasets. For instance, many studies prefer using the UNBC dataset, with varying results over the years: 86.10% accuracy,⁹⁶ 87.20%,⁹⁷ 76.00%,⁹⁸ and 93.16%,⁹⁹ using the Leave-One-Subject-Out (LOSO) cross-validation method. It is worth noting that these results, while higher than those obtained in this study, were achieved on a single dataset using a different cross-validation method, making direct comparisons unfair. No other works utilizing the MInt pain database with solely 2D information were found, and no deep learning-based approaches were identified for the Delaware dataset, likely due to its reduced size. Therefore, while comparisons with existing literature may be challenging, this study contributes valuable insights by leveraging multiple datasets, which may enhance the generalizability and robustness of the findings. Furthermore, we were unable to find prior work that included users with CP, making direct comparisons with our results on the CP-PAIN dataset unfeasible. As pioneers in studying automatic pain recognition among individuals with cerebral CP, our study lays the foundation for future research in this area.

Given the focus on real-world applicability, the scalability of our automated facial recognition system is a critical consideration. While the system was not specifically trained on individuals with cerebral CP, the results obtained from the CP-PAIN dataset can reasonably generalize to other individuals within the CP population. This assumption is grounded in the understanding that the users in the CP-PAIN dataset represent a sample from the broader CP population and are likely to share common characteristics in their pain expressions. However, challenges may arise when deploying the system in diverse environments or

when faced with varying degrees of pain expressions within the CP population. As the system is deployed in different settings, the availability of more data from individuals with CP will likely increase. This presents an opportunity for fine-tuning the models using CP-specific data, thereby enhancing the system's performance in pain recognition tailored to this population. Moreover, further exploration of the similarities in pain expression among individuals with CP can provide valuable insights for improving the system's accuracy and reliability in real-world scenarios.

For the models exhibiting the highest performance, such as InceptionV3 and ResNet101V2, the application of XAI techniques have unveiled an interesting finding: when employing models trained on the PAIN-DB dataset on users with CP, the essential regions crucial for accurate pain identification remain largely unaffected. The consistent focus of these two models on identical facial regions for pain recognition both for users with CP and without suggests a noteworthy similarity in the expression of pain between these two distinct groups. However, it is essential to acknowledge that this finding does not consider potential idiosyncrasies in the facial expression of pain in subjects with complex neurological disorders, which may have led to the presence of false positives. It is important to note that some idiosyncratic behaviors, such as crying, moaning, flinching, having red cheeks, grunting, or sticking out the tongue can be misleadingly labeled as pain within the CP population, while they can express other physiological or emotional events.⁹³ On the contrary, other facial gestures, such as smiling or laughing, may be used to express pain in individuals with poorer communication or motor ability.⁹⁴ This underscores the need for further training of the system using a more diverse dataset of images of individuals with CP. We are committed to addressing potential variations in the expression of pain and enriching the robustness of our models for pain recognition across diverse user groups. This approach would ensure a more precise and reliable application of our technology in clinical and medical settings.

This study was not without limitations. In addition to the previously mentioned constraint of a limited number of images in our database, we also faced the unforeseen challenge of not being able to collect self-reports from most individuals with CP, as originally planned, what would have been valuable to evaluate the facial recognition system's measurements in a more accurate way. Another potential limitation was the unequal representation of different types of painful stimuli.

DL pain recognition allows envisioning the possibility of mitigating the limitations of human-based assessments of pain in complex conditions, such as CP, and achieving a level of objectivity and consistency that can significantly benefit the care and well-being of individuals with communication problems and cognitive and neuromotor disorders affecting pain expression. Therefore, our work not only showcases the potential of adapting technology for healthcare applications but also emphasizes the importance of

ongoing research and data collection to advance and extrapolate the capabilities of our model in the future.

Conclusions

Our study addresses the profound challenges in pain assessment for individuals with CP, highlighting the limitations of existing methodologies and the need for approaches tailored to this unique population. Unlike conventional studies that often utilize datasets and models designed for neurotypical individuals, our research introduces an innovative approach by developing and employing the CP-PAIN dataset. This dataset is specifically designed to capture the diverse and often unique expressions of pain characteristic of individuals with CP, highlighting a significant departure from the broader application DL models prevalent in current research. Our efforts to customize DL models for the CP population underscore the complex nature of pain expressions in these individuals, which are not adequately addressed by standard facial recognition systems.

Despite showing promise, with an initial accuracy rate of 60%, our findings also signal the need for further enhancements and the expansion of the CP-PAIN database. A more comprehensive collection of data is essential for refining DL models to accurately identify the nuanced pain expressions of individuals with CP, a step towards developing a universally applicable pain assessment tool. Our study not only contributes to the specialized field of pain assessment in CP but also emphasizes the critical importance of ongoing research. By focusing on the unique needs of individuals with CP, we underscore the broader challenge of creating more inclusive and accessible pain assessment methodologies. This endeavor not only advances our understanding of pain assessment in the context of CP but also sets a precedent for the importance of tailoring research and technology to meet the specific needs of underserved populations.

Finally, in addressing the ethical and privacy implications of using automated facial recognition for pain assessment in clinical settings, it is crucial to obtain transparent informed consent, protect sensitive biometric data through strict adherence to data protection laws and advanced encryption, and prevent bias to ensure equitable patient assessments. Continuous stakeholder engagement and system evaluation are imperative for upholding ethical standards and patient privacy.

Abbreviations

AI	Artificial Intelligence.
CP-PAIN	dataset of facial pain expression images for individuals with cerebral palsy based on video recorded during potentially painful procedures.
CP	Cerebral Palsy.
FACS	Facial Action Coding System.
LIME	Local Interpretable Model-agnostic Explanations.

MInt PAIN Multimodal Intensity Pain dataset.
 NCAPC Non-Communicating Adults Pain Checklist.
 PAIN-DB dataset including the UNBC-McMaster Shoulder Pain Expression Archive Database, the Multimodal Intensity Pain Dataset, and the Delaware Pain Database.

Acknowledgements: The authors would like to thank the healthcare staff of the ASPACE Balearic Foundation and ASPACE Toledo for their various contributions during the development of this study.

Contributorship: AS-G and IR researched the literature and conceived the study. AS-G, IR, and PM were involved in protocol development, gaining ethical approval, and recruiting patients. JMBR and CMY worked on the conceptualization of the experiment. FXGM and JMBR conducted the experiments. AS-G carried out the data analysis. AS-G, FXGM, and CMY wrote the initial draft of the manuscript. All authors reviewed, edited, and approved the final version of the manuscript.

Ethical approval: The research protocol received approval from both the ASPACE Foundation's ethics committee and the Research Ethics Commission of the Balearic Islands (protocol number IB4046/19), affirming its ethical rigor.

Guarantor: IR

Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by MICIU/AEI/ 10.13039/501100011033 Spain, grants PID2020-114967GA-I00 (SENTS?) and PID2019-104829RA-I00, EXPLAInable Artificial Intelligence systems for health and well-beING (EXPLAINING) and for the Ministry of Science and Technology of Spain /Spanish Foundation for Science and Technology (FECYT, grant FCT-20-16485). In addition, the authors also acknowledge the funding of the FPU scholarship from the Ministry of European Funds, University and Culture of the Government of the Balearic Islands.

ORCID iDs: Álvaro Sabater-Gárriz  <https://orcid.org/0000-0001-5770-7240>

F Xavier Gaya-Morey  <https://orcid.org/0000-0003-1231-7235>
 Inmaculada Riquelme  <https://orcid.org/0000-0003-4705-8325>

Supplemental material: Supplemental material for this article is available online.

References

1. Bax M, Goldstein M, Rosenbaum P, et al. Proposed definition and classification of cerebral palsy, April 2005. *Dev Med Child Neurol* 2005; 47: 571–576.
2. Peterson MD and Hurvitz EA. Cerebral palsy grows up. *Mayo Clin Proc* 2021; 96: 1404–1406.
3. Jonsson U, Eek MN, Sunnerhagen KS, et al. Cerebral palsy prevalence, subtypes, and associated impairments: a population-based comparison study of adults and children. *Dev Med Child Neurol* 2019; 61: 1162–1167.
4. Novak I, Hines M, Goldsmith S, et al. Clinical prognostic messages from a systematic review on cerebral palsy. *Pediatrics* 2012; 130: e1285–e1312.
5. Tarsuslu Şimşek T and Livanelioğlu A. Serebral paralizili bireylerde ağrının aktivite bağımsızlığı ve sağlıkla ilgili yaşam kalitesi üzerine etkisi [the effect of pain on activity independence and health-related quality of life in cerebral palsied individuals]. *Agri* 2011; 23: 107–113.
6. Kerr C, McDowell BC, Parkes J, et al. Age-related changes in energy efficiency of gait, activity, and participation in children with cerebral palsy. *Dev Med Child Neurol* 2011; 53: 61–67.
7. Yang KT, Yin CH, Hung YM, et al. Continuity of care is associated with medical costs and inpatient days in children with cerebral palsy. *Int J Environ Res Public Health* 2020; 17: 2913. Published 2020 Apr 23.
8. Schaible B, Colquitt G, Caciula MC, et al. Comparing impact on the family and insurance coverage in children with cerebral palsy and children with another special healthcare need. *Child Care Health Dev* 2018; 44: 370–377.
9. du Toit J, Eken MM, Lamberts RP, et al. Adults with spastic diplegic cerebral palsy living in a low-to-middle income country: a six-year follow-up study on pain, functional mobility, activity and participation. *Disabil Health J* 2021; 14: 101130.
10. Schölderle T, Staiger A, Lampe R, et al. Dysarthria in adults with cerebral palsy: clinical presentation and impacts on communication. *J Speech Lang Hear Res* 2016; 59: 216–229.
11. Axmon A, Sandberg M, Ahlström G, et al. Prescription of potentially inappropriate medications among older people with intellectual disability: a register study. *BMC Pharmacol Toxicol* 2017; 18: 68. PMID: 29070067; PMCID: PMC5657112.
12. Duerden EG, Taylor MJ, Lee M, et al. Decreased sensitivity to thermal stimuli in adolescents with autism spectrum disorder: relation to symptomatology and cognitive ability. *J Pain* 2015; 16: 463–471. Epub 2015 Feb 19. PMID: 25704841.
13. McGuire BE, Daly P and Smyth F. Chronic pain in people with an intellectual disability: under-recognised and under-treated? *J Intellect Disabil Res* 2010; 54: 240–245. PMID: 20387264.
14. Rodby-Bousquet E, Alriksson-Schmidt A and Jarl J. Prevalence of pain and interference with daily activities and sleep in adults with cerebral palsy. *Dev Med Child Neurol* 2021; 63: 60–67. Epub 2020 Sep 19. PMID: 32951227; PMCID: PMC7756851.
15. Pasero C and McCaffery M. *Pain assessment and pharmacologic management-E-Book*. Saint Louis, MO, USA: Elsevier Health Sciences, 2010.
16. Herr K, Coyne PJ, Ely E, et al. Pain assessment in the patient unable to self-report: clinical practice recommendations in support of the ASPMN 2019 position statement. *Pain Manag Nurs* 2019; 20: 404–417. Epub 2019 Oct 12. PMID: 31610992.
17. McKinnon C, White J, Morgan P, et al. Clinician perspectives of chronic pain management in children and adolescents with cerebral palsy and dyskinesia. *Phys Occup Ther Pediatr* 2021; 41: 244–258. Epub 2020 Nov 29. PMID: 33251932.

18. Bernal-Celestino RJ, León-Martín A, Martín-López MM, et al. Evaluating and handling the pain of people with intellectual disability. *Pain Manag Nurs* 2022; 23: 311–317. Epub 2021 Sep 4. PMID: 34493439.
19. Ostojic K, Paget S, Kyriagis M, et al. Acute and chronic pain in children and adolescents with cerebral palsy: prevalence, interference, and management. *Arch Phys Med Rehabil* 2020; 101: 213–219. Epub 2019 Sep 12. PMID: 31521713.
20. Penner M, Xie WY, Binopal N, et al. Characteristics of pain in children and youth with cerebral palsy. *Pediatrics* 2013; 132: e407–e413. Epub 2013 Jul 15. PMID: 23858420.
21. Gutysz-Wojnicka A, Ozga D, Mayzner-Zawadzka E, et al. Psychometric assessment of physiologic and behavioral pain indicators in Polish versions of the pain assessment scales. *Pain Manag Nurs* 2019; 20: 292–301.
22. Ostojic K, Paget SP and Morrow AM. Management of pain in children and adolescents with cerebral palsy: a systematic review. *Dev Med Child Neurol* 2019; 61: 315–321. Epub 2018 Oct 31. PMID: 30378122.
23. Hu S, Loo JA and Wong DT. Human saliva proteome analysis and disease biomarker discovery. *Expert Rev Proteomics* 2007; 4: 531–538.
24. Cantón-Habas V, Rich-Ruiz M, Moreno-Casbas MT, et al. Correlation between biomarkers of pain in saliva and PAINAD scale in elderly people with cognitive impairment and inability to communicate. *J Clin Med* 2021; 10: 1424. Published 2021 Apr 1.
25. Benromano T, Pick CG, Granovsky Y, et al. Increased evoked potentials and behavioral indices in response to pain among individuals with intellectual disability. *Pain Med* 2017; 18: 1715–1730.
26. Arbour C, Gélinas C, Loisel CG, et al. An exploratory study of the bilateral bispectral index for pain detection in traumatic-brain-injured patients with altered level of consciousness. *J Neurosci Nurs* 2015; 47: 166–177.
27. Jeitziner MM, Schwendimann R, Hamers JP, et al. Assessment of pain in sedated and mechanically ventilated patients: an observational study. *Acta Anaesthesiol Scand* 2012; 56: 645–654.
28. Koenig J, Jarczok MN, Ellis RJ, et al. Heart rate variability and experimentally induced pain in healthy adults: a systematic review. *Eur J Pain* 2014; 18: 301–314.
29. Bradley MM, Silakowski T and Lang PJ. Fear of pain and defensive activation. *Pain* 2008; 137: 156–163.
30. Sabater-Gárriz Á, Molina-Mula J, Montoya P, et al. Pain assessment tools in adults with communication disorders: systematic review and meta-analysis. *BMC Neurol* 2024; 24: 66. PMID: 38368314; PMCID: PMC10873938.
31. Klarer N, Rickenbacher H, Kasser S, et al. Electrophysiological measurement of noxious-evoked brain activity in neonates using a flat-tip probe coupled to electroencephalography. *J Vis Exp* 2017; 129: 56531.
32. Nerella S, Cupka J, Ruppert M, et al. Pain action unit detection in critically ill patients. *Proc COMPSAC* 2021: 645–651. doi:10.1109/compsac51774.2021.00094
33. Roué JM, Morag I, Haddad WM, et al. Using sensor-fusion and machine-learning algorithms to assess acute pain in non-verbal infants: a study protocol. *BMJ Open* 2021; 11: e039292. Published 2021 Jan 6.
34. Kappesser J, Voit S, Lautenbacher S, et al. Pain assessment for cognitively impaired older adults: do items of available observer tools reflect pain-specific responses? *Eur J Pain* 2020; 24: 851–862.
35. Böling S, Tarja V, Helena M, et al. Measuring quality of life of Finnish children with cerebral palsy. *J Pediatr Rehabil Med* 2013; 6: 121–127.
36. Ramstad K, Jahnsen R, Skjeldal OH, et al. Characteristics of recurrent musculoskeletal pain in children with cerebral palsy aged 8 to 18 years. *Dev Med Child Neurol* 2011; 53: 1013–1018.
37. Riquelme I, Cifre I and Montoya P. Are physiotherapists reliable proxies for the recognition of pain in individuals with cerebral palsy? A cross sectional study. *Disabil Health J* 2015; 8: 264–270.
38. Ekman P and Friesen WV. *Facial action coding system : investigator's guide*. Palo Alto, Calif.: Consulting Psychologists Press, 1978.
39. Prkachin KM and Solomon PE. The structure, reliability and validity of pain expression: evidence from patients with shoulder pain. *Pain* 2008; 139: 267–274.
40. Rojo R, Prados-Frutos JC and López-Valverde A. Evaluación del dolor mediante el sistema de codificación de la acción facial. Revisión sistemática [pain assessment using the facial action coding system. A systematic review]. *Med Clin (Barc)* 2015; 145: 350–355.
41. Shen D, Wu G and Suk HI. Deep learning in medical image analysis. *Annu Rev Biomed Eng* 2017; 19: 221–248. Epub 2017 Mar 9. PMID: 28301734; PMCID: PMC5479722.
42. Kleesiek J, Urban G, Hubert A, et al. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. *Neuroimage* 2016; 129: 460–469. Epub 2016 Jan 22. PMID: 26808333.
43. Zhang W, Li R, Deng H, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuroimage* 2015; 108: 214–224. Epub 2015 Jan 3. PMID: 25562829; PMCID: PMC4323729.
44. Wu G, Kim M, Wang Q, et al. Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE Trans Biomed Eng* 2016; 63: 1505–1516. Epub 2015 Nov 2. Erratum in: *IEEE Trans Biomed Eng.* 2017 Jan;64(1):250. PMID: 26552069; PMCID: PMC4853306.
45. Suk H-I and Shen D. Deep learning in diagnosis of brain disorders. In: Lee S-W, Bülthoff H and Müller KR (ed.) *Recent progress in brain and cognitive engineering*. Netherlands: Springer, 2015, pp.203–213. https://doi.org/10.1007/978-94-017-7239-6_14
46. Shin H-C, Roberts K, Lu L, et al. Learning to read chest X-rays: recurrent neural cascade model for automated image annotation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp.2497–2506. <https://doi.org/10.1109/CVPR.2016.274>
47. Suk HI, Lee SW and Shen D; Alzheimer's Disease Neuroimaging Initiative. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage* 2014; 101: 569–582. Epub 2014 Jul 18. PMID: 25042445; PMCID: PMC4165842.
48. Pereira S, Pinto A, Alves V, et al. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE*

- Trans Med Imaging* 2016; 35: 1240–1251. Epub 2016 Mar 4. PMID: 26960222.
49. Cireşan DC, Giusti A, Gambardella LM, et al. Mitosis detection in breast cancer histology images with deep neural networks. *Med Image Comput Assist Interv* 2013; 16: 411–418. PMID: 24579167.
 50. Jia S, Wang S, Hu C, et al. Detection of genuine and posed facial expressions of emotion: databases and methods. *Front Psychol* 2021; 11: 580287. Published 2021 Jan 15.
 51. Mende-Siedlecki P, Qu-Lee J, Lin J, et al. The delaware pain database: a set of painful expressions and corresponding norming data. *Pain Rep* 2020; 5: e853. Published 2020 Oct 21.
 52. Al-Eidan M, Al-Khalifa RH and Al-Salman A. Deep-learning-based models for pain recognition: a systematic review. *Appl Sci* 2020; 10: 5984.
 53. Tavakolian M, Cruces CGB and Hadid A. Learning to detect genuine versus posed pain from facial expressions using residual generative adversarial networks. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019, May, pp.1–8. IEEE.
 54. Baltrušaitis T, Mahmoud M and Robinson P. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, May, Vol. 6, pp.1–6. IEEE.
 55. Baltrušaitis T, Zadeh A, Lim YC, et al. Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp.59–66. IEEE.
 56. Ertugrul IO, Jeni LA, Ding W, et al. AFAR: a deep learning based tool for automated facial affect recognition. *Proc Int Conf Autom Face Gesture Recognit* 2019; 2019: 10.1109/FG.2019.8756623.
 57. Allsop MJ, Johnson O, Taylor S, et al. Multidisciplinary software design for the routine monitoring and assessment of pain in palliative care services: the development of PainCheck. *JCO Clin Cancer Inform* 2019; 3: 1–17.
 58. Hoti K, Chivers PT and Hughes JD. Assessing procedural pain in infants: a feasibility study evaluating a point-of-care mobile solution based on automated facial analysis. *Lancet Digit Health* 2021; 3: e623–e634.
 59. Babicova I, Cross A, Forman D, et al. Evaluation of the psychometric properties of PainChek® in UK aged care residents with advanced dementia. *BMC Geriatr* 2021; 21: 337. Published 2021 May 28.
 60. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf Fusion* 2020; 58: 82–115. ISSN 15662535.
 61. Gunning D and Aha D. DARPA's explainable artificial intelligence (XAI) program. *AI Mag* 2019; 40: 44–58.
 62. Heimerl A, Weitz K, Baur T, et al. Unraveling ml models of emotion with nova: multi-level explainable ai for non-experts. *IEEE Trans Affective Comput* 2020; 13: 1155–1167.
 63. Ramis S, Buades JM, Perales FJ, et al. A novel approach to cross dataset studies in facial expression recognition. *Multimedia Tools Appl* 2022; 81: 39507–39544. ISSN 1380-7501.
 64. Weitz K, Hassan T, Schmid U, et al. Deep-learned faces of pain and emotions: elucidating the differences of facial expressions with the help of explainable AI methods. *tm-Technisches Messen* 2019; 86: 404–412.
 65. Li S and Deng W. Deep facial expression recognition: a survey. *IEEE Trans Affective Comput* 2020; 13: 1195–1215.
 66. Wong DL and Baker CM. Pain in children: comparison of assessment scales. *Pediatr Nurs* 1988; 14: 9–17.
 67. Lotan M, Moe-Nilssen R, Ljunggren AE, et al. Measurement properties of the non-communicating adult pain checklist (NCAPC): a pain scale for adults with intellectual and developmental disabilities, scored in a clinical setting. *Res Dev Disabil* 2010; 31: 367–375.
 68. Palisano RJ, Rosenbaum P, Bartlett D, et al. Content validity of the expanded and revised gross motor function classification system. *Dev Med Child Neurol* 2008; 50: 744–750. PMID: 18834387.
 69. Hidecker MJ, Paneth N, Rosenbaum PL, et al. Developing and validating the communication function classification system for individuals with cerebral palsy. *Dev Med Child Neurol* 2011; 53: 704–710. Epub 2011 Jun 27. PMID: 21707596; PMCID: PMC3130799.
 70. Bartlett M, Littlewort G, Vural E, et al. Data mining spontaneous facial behavior with automatic expression coding. In: verbal and nonverbal features of human–human and human–machine interaction: COST Action 2102 International Conference, October 29–31, 2007, Patras, Greece, 2008. Revised Papers, pp.1–20. Springer Berlin Heidelberg.
 71. Benromano T, Pick CG, Merick J, et al. Physiological and behavioral responses to calibrated noxious stimuli among individuals with cerebral palsy and intellectual disability. *Pain Med (Malden, Mass.)* 2017; 18: 441–453.
 72. Lucey P, Cohn JF, Prkachin KM, et al. Painful data: the unbc-mcmaster shoulder pain expression archive database. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2011, pp.57–64. doi: 10.1109/FG.2011.5771462
 73. Breau LM and Camfield CS. The relation between children's pain behaviour and developmental characteristics: a cross-sectional study. *Dev Med Child Neurol* 2011; 53: e1–e7.
 74. Lotan M, Moe-Nilssen R, Ljunggren AE, et al. Reliability of the non-communicating adult pain checklist (NCAPC), assessed by different groups of health workers. *Res Dev Disabil* 2009; 30: 735–745. Epub 2008 Nov 25. PMID: 19036559.
 75. Ozkan D, Gonen E, Akkaya T, et al. Popliteal block for lower limb surgery in children with cerebral palsy: effect on sevoflurane consumption and postoperative pain (a randomized, double-blinded, controlled trial). *J Anesth* 2017; 31: 358–364.
 76. Laschinger HK. Intraclass correlations as estimates of interrater reliability in nursing research. *West J Nurs Res* 1992; 14: 246–251. PMID: 1561790.
 77. Fleiss JL. *The design and analysis of clinical experiments*. New York: Wiley-Interscience, 1986.
 78. Lucey P, Cohn JF, Prkachin KM, et al. Painful data: the UNBC-McMaster shoulder pain expression archive database. In: 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2011, March, pp.57–64. IEEE. doi: 10.1109/ACII.2009.5349321
 79. Haque MA, Bautista RB, Noroozi F, et al. Deep multimodal pain recognition: a database and comparison of spatio-

- temporal visual modalities. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, May, pp.250–257. IEEE. doi: 10.1109/FG.2018.00044
80. Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process Lett* 2016; 23: 1499–1503.
 81. Qin X, Zhang Z, Huang C, et al. U2-net: going deeper with nested u-structure for salient object detection. *Pattern Recognit* 2020; 106: 107404.
 82. Song I, Kim HJ and Jeon PB. Deep learning for real-time robust facial expression recognition on a smartphone. In: 2014 IEEE International Conference on Consumer Electronics (ICCE), 2014, pp.564–567. doi:10.1109/ICCE.2014.6776135
 83. Li W, Li M, Su Z, et al. A deep-learning approach to facial expression recognition with candid images. In: 2015 14th IAPR International Conference on Machine Vision Applications (MVA), 2015, pp.279–282. IEEE.
 84. Krizhevsky A, Sutskever I and Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. New York, NY, USA: Communications of the ACM, 2012, pp.84–90.
 85. Simonyan K and Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint; 2014.:1409-1556.
 86. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp.770–778.
 87. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, 2016, pp.2818–2826.
 88. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp.1251–1258.
 89. Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009 June, pp.248–255. IEEE. doi: 10.1109/CVPR.2009.5206848
 90. Ribeiro MT, Singh S and Guestrin C. “Why should i trust you?” explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp.1135–1144. doi: 10.1145/2939672.2939778
 91. Manresa-Yee C, Ramis S and Buades JM. Analysis of gender differences in facial expression recognition based on deep learning using explainable artificial intelligence. *Int J Inter Multimed Artif Intell*, In Press 2023; 4: 1–10. ISSN 1989-1660.
 92. Lacerda DC, Ferraz-Pereira KN, Bezerra de Moraes AT, et al. Oro-facial functions in experimental models of cerebral palsy: a systematic review. *J Oral Rehabil* 2017; 44: 251–260.
 93. Hadden KL and von Baeyer CL. Pain in children with cerebral palsy: common triggers and expressive behaviors. *Pain* 2002; 99: 281–288.
 94. Kunz M, Prkachin K and Lautenbacher S. The smile of pain. *Pain* 2009; 145: 273–275.
 95. Gkikas S and Tsiknakis M. Automatic assessment of pain based on deep learning methods: a systematic review. *Comput Methods Programs Biomed* 2023; 231: 107365. Epub 2023 Feb 8. PMID: 36764062.
 96. Pedersen H. Learning appearance features for pain detection using the UNBC-McMaster shoulder pain expression archive database. In: Nalpantidis L, Krüger V, Eklundh JO, et al (eds) *Computer vision systems*. Copenhagen, Denmark: Springer International Publishing, 2015, pp.128–136.
 97. Kharghanian R, Peiravi A and Moradi F. Pain detection from facial images using unsupervised feature learning approach. *Annu Int Conf IEEE Eng Med Biol Soc* 2016; 2016: 419–422. PMID: 28268362.
 98. Rudovic O, Tobis N, Kaltwang S, et al. Personalized federated deep learning for pain estimation from face images. 2021. arXiv preprint arXiv:2101.04800.
 99. Kharghanian R, Peiravi A, Moradi F, et al. Pain detection using batch normalized discriminant restricted Boltzmann machine layers. *J Vis Commun Image Represent* 2021; 76: 103062
-

Appendix

See Figures 7–10.



Figure 7. Examples of images correctly classified as “pain” by the InceptionV3 model on the CP-PAIN dataset.

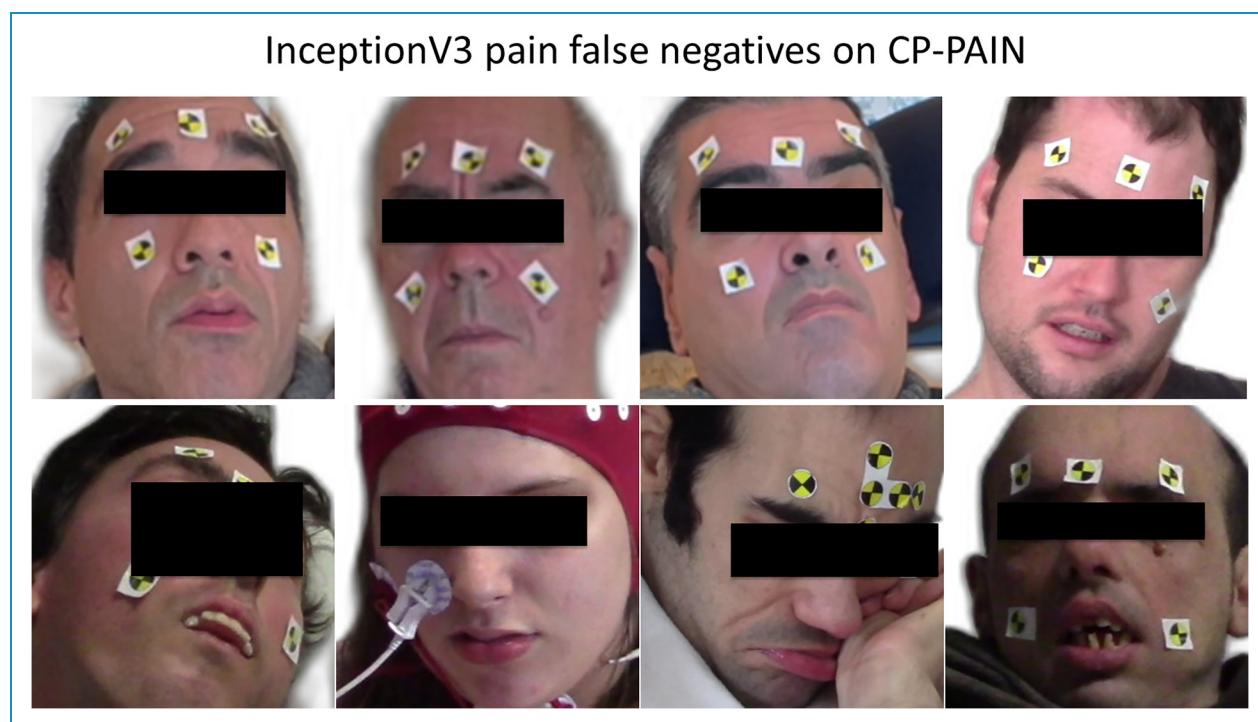


Figure 8. Examples of images incorrectly classified as “no pain” by the InceptionV3 model on the CP-PAIN dataset.

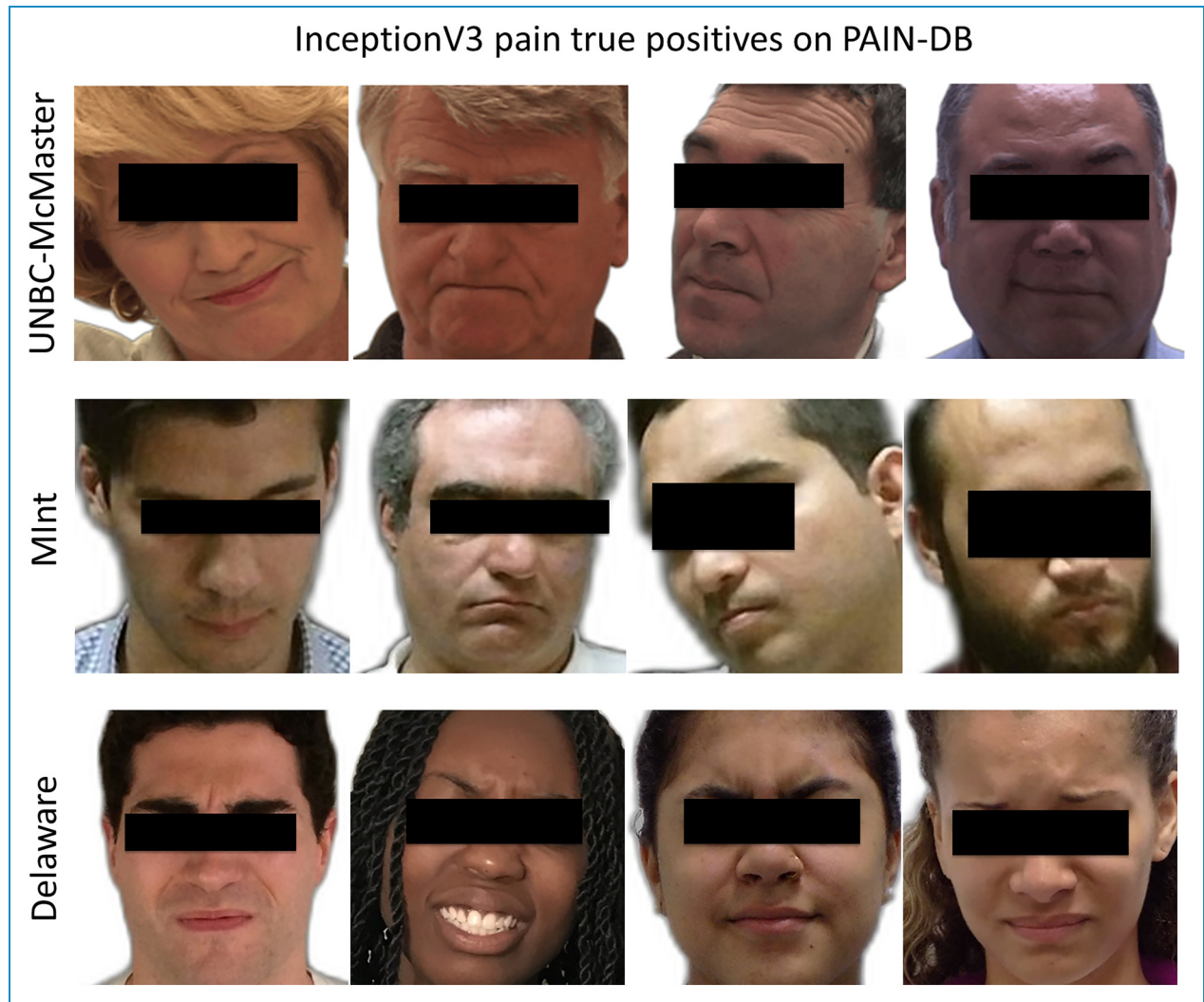


Figure 9. Examples of images correctly classified as “pain” by the InceptionV3 model on the PAIN-DB dataset, grouped by original dataset (by rows).

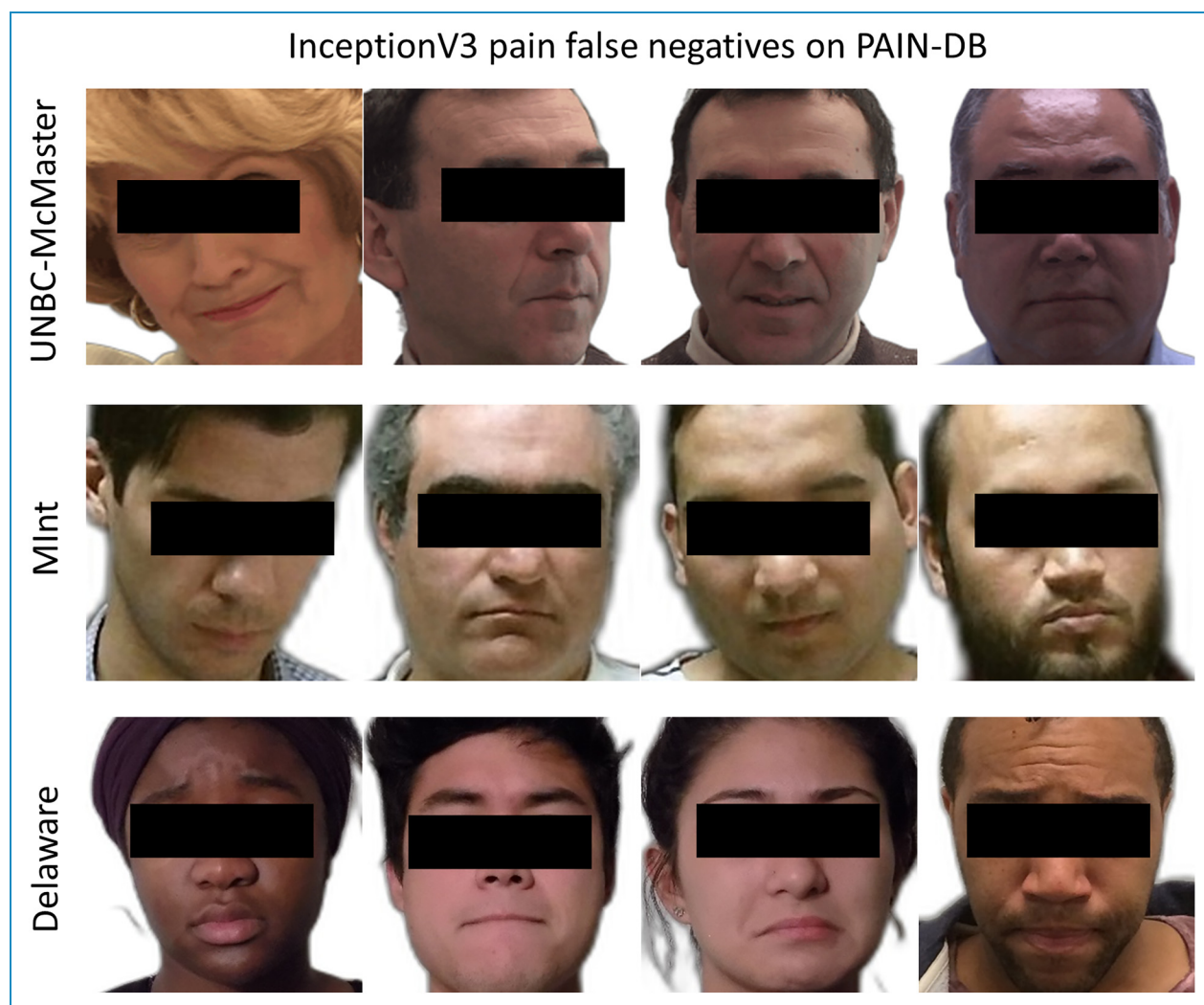


Figure 10. Examples of images incorrectly classified as “no pain” by the InceptionV3 model on the PAIN-DB dataset, grouped by original dataset (by rows).