Computers in Biology and Medicine 189 (2025) 109908

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Disclosing neonatal pain in real-time: AI-derived pain sign from continuous assessment of facial expressions

Leonardo Antunes Ferreira [a,*], Lucas Pereira Carlini [a], Gabriel de Almeida Sá Coutrin [a], Tatiany Marcondes Heiderich [a,b], Rita de Cássia Xavier Balda [b], Marina Carvalho de Moraes Barros [b], Ruth Guinsburg [b], Carlos Eduardo Thomaz [a]

[a] Department of Electrical Engineering, Centro Universitário FEI, Av. Humberto de Alencar Castelo Branco, 3972-B, São Bernardo do Campo, 09850-901, São Paulo, Brazil

[b] Department of Paediatrics, Federal University of São Paulo, R. Botucatu, 740, São Paulo, 04024-002, São Paulo, Brazil

ARTICLE INFO

Keywords:
Pain assessment
Deep-learning
Continuous monitoring
Pain sign

ABSTRACT

This study introduces an AI-derived pain sign for continuous neonatal pain assessment, addressing the limitations of existing pain scales and computational approaches. Traditional pain scales, though widely used, are hindered by inter-rater variability, discontinuity, and subjectivity. While AI, particularly Deep-Learning, has shown promise, prior research has largely prioritized model performance over clinical applicability, often delivering static, binary predictions that lack interpretability in clinical practice. To bridge this gap, we developed a real-time pain sign tracking tool using facial expression analysis, a primary and non-invasive pain indicator in neonates. Leveraging benchmark datasets (iCOPE, iCOPEvid, and UNIFESP) and Deep-Learning frameworks (VGG-Face, N-CNN, and ViT-B/16), the models analyze video frames to generate a continuous visual representation of pain probability. Our results reveal the limitations of single-label predictions for time intervals, emphasizing the utility of a continuous monitoring visualization tool. The proposed pain sign effectively tracks dynamic changes in neonatal facial expressions, providing actionable and interpretable insights for healthcare professionals. We categorized these insights into a novel classification scheme, such as stable, irregular, unstable, and indeterminate pain signs. By integrating this pain sign into clinical workflows as a potential vital sign, this approach enables personalized pain management and continuous monitoring of both current and historical pain states in neonates, enhancing neonatal care and improving outcomes for these vulnerable patients.

1. Introduction

In the past, neonatal pain was disregarded under the common assumption that the neural pathways responsible for pain in newborns were not fully developed, resulting in invasive procedures being performed without the use of analgesic or anesthetic agents [1,2]. Medical advancements have since debunked the notion that newborns do not experience pain [3]. However, even with this understanding, neonates admitted to NICUs (Neonatal Intensive Care Units) may still undergo more than 300 painful procedures during their stay [4] and two-thirds of neonates admitted into NICUs do not receive continuous pain assessment [5]. In fact, it is known that prolonged or repeated exposure to painful stimuli can lead to detrimental lifelong consequences [1,4,6].

Over 40 neonatal pain scales have since been proposed [4], like the Neonatal Facial Coding System (NFCS) [2], Neonatal Pain, Agitation and Sedation Scale (N-PASS) [7], Neonatal Infant Pain Scale (NIPS) [8], and Face, Legs, Activity, Cry, Consolability (FLACC) scale [9]. The existence of numerous scales underscores the importance of identifying and addressing pain, as well as the difficulty in establishing a universal method for neonatal pain assessment. Most of these scales share the analysis of facial expressions as a common feature, capturing highly sensitive and specific indicators of pain intensity and nature, such as a furrowed brow, narrowed eyes, nasolabial furrow, tense tongue, and a widened mouth angle [2]. Facial expression is, in fact, a primary and non-invasive form of manifestation of pain, being a common and universal expression present from birth regardless of gender and ethnicity [10], facilitating effective communication between non-verbal neonates and healthcare professionals.

Although pain scales are widely employed, their precision is compromised by the lack of inter-rater agreement among healthcare professionals [11–13]. Moreover, these scales are not continuous due to

* Corresponding author.
E-mail address: leferr@fei.edu.br (L.A. Ferreira).

https://doi.org/10.1016/j.compbiomed.2025.109908
Received 2 October 2024; Received in revised form 4 February 2025; Accepted 19 February 2025
Available online 1 March 2025

the labor-intensive nature of the process, requiring the participation of multiple skilled professionals, resulting in intermittent monitoring [5]. Additionally, their subjectivity is rooted in the innate human tendency to perceive pain subjectively in others [14].

To address these challenges, researchers have explored the use of Artificial Intelligence (AI), particularly Deep-Learning, for detecting neonatal pain through facial expressions [12,15,16]. These studies have demonstrated accurate and objective results and highlight the potential applicability of AI as a continuous tool for predictive early pain management [17]. However, a significant gap remains between computational and medical literature, specifically the need for a clinically comprehensible and user-friendly tool for neonatal pain assessment that delivers actionable insights with minimal training [13]. While previous computational studies have primarily focused on improving performance metrics by developing Deep-Learning architectures or integrating multidimensional data, they often fall short in producing practical clinical tools. In other words, if current research methods were implemented in clinical settings, healthcare professionals would still rely on static, discontinuous predictions, such as binary labels (“pain” or “no pain”) or single pain intensity estimates for a given time interval. These approaches lack the interpretability and visual clarity required in NICUs, where clinicians must frequently assess both the historical and current pain states of neonates to distinguish between acute, prolonged, persistent, or chronic pain [18].

To bridge the gap between computational methodologies and clinical practice in neonatal pain assessment, our study brings the following main contributions:

- *A real-time, continuous pain sign*: Proposed and implemented to track dynamic changes in neonatal facial expressions, addressing the limitations of static, single-label pain assessments;
- *A novel set of temporal classifications for the pain phenomenon*: Developed specifically for healthcare professionals in clinical practice, supporting actionable decision-making in NICU settings;
- *Correlation between model probabilities and pain facial expression intensity*: Our experimental results show a strong correlation between the probabilities generated by AI models and the intensity of pain-related facial expressions, even though the models were not explicitly trained for regression tasks.

2. Material and methods

2.1. Data sources and pre-processing

We utilized three benchmark datasets of neonatal pain focused on capturing neonatal facial expressions: iCOPE [19], UNIFESP [20], and iCOPEvid [21]. Both UNIFESP and iCOPE consist of images, whereas iCOPEvid is a video dataset. The characteristics of each dataset are detailed in Table 1.

Brahnam et al. [19] conducted one of the pioneering studies investigating neonatal pain through facial expressions, and subsequently created one of the earliest datasets for this purpose, named iCOPE dataset. Adhering to protocols and ethics directives for research involving human subjects at St. John’s Health System Inc., the authors obtained images of healthy neonates during predefined procedures, including transport from one crib to another, air stimulus, friction, and pain caused by a heel puncture. The captured images were categorized as follows: 63 images classified as resting, 18 depicting crying, 23 during air stimulus, 36 during friction, and 60 exhibiting pain. For our study, only images labeled as “rest” and “pain” were utilized.

The UNIFESP dataset was established with the objective of developing a software capable of automatically detecting facial expressions of pain in newborns by Heiderich et al. [20]. All newborns involved in this data collection were healthy, free from congenital malformations, and were hospitalized at Hospital São Paulo of Escola Paulista de Medicina at the Federal University of São Paulo. Image acquisition was granted approval by the Research Ethics Committee of the university under protocol number 1299/09. For each second, three images were extracted from 10-min videos recorded before, during, and after painful procedures such as injections, heel puncture, and venipuncture. Out of all the images acquired, 360 images were clinically evaluated by neonatal healthcare professionals using the NFCS [2] resulting in “no pain” and “pain” labels. Consequently, only these labeled images were included in this study.

Table 1
Datasets characteristics.

Dataset	iCOPE	UNIFESP	iCOPEvid
Data type	Image	Image	Video
Newborns	26	30	49
Males	13	15	26
Females	13	15	23
Postnatal age (hours)	24 h–144 h	18 h–72 h	34 h–70 h
Camera resolution	3008 × 2000	320 × 233	1920 × 1080
Total samples	123	360	120
Pain	60	164	71
No Pain	63	196	49

Later, Brahnam et al. [21] also introduced the iCOPEvid dataset adheres to the same procedures and ethical considerations as iCOPE for data acquisition. However, in this instance, videos of neonates experiencing both painful (heel puncture) and non-painful (movement, friction, and rest) stimuli were recorded for at least 1 min during each procedure. From the raw footage, a total of 234 video segments, each lasting 20-s, were extracted, capturing the precise moment when the noxious stimulus was administered. According to the authors of this dataset, when feasible, the non-painful video segments were hand-picked by two investigators to present challenging facial expressions for classification. This implies that non-painful facial expressions exhibit characteristics similar to those of painful expressions. To maintain consistency against datasets, we solely utilized the “rest” and “pain” categories, same and analogous as for the iCOPE and UNIFESP datasets.

All images and videos underwent a pre-processing step of face detection, wherein we employed the RetinaFace [22] face detector. This process involved cropping out only the faces of the neonates as the regions of interest.

2.2. Deep-learning framework

We utilized two benchmark models for automatic neonatal pain assessment [21,23–30], VGG-Face [31] and N-CNN (Neonatal Convolutional Neural Network) [32]. Additionally, we introduced a third model, a ViT (Vision Transformer), specifically the ViT-B/16 [33], to assess whether newer Deep-Learning architectures can perform comparably to established CNNs in this task.

UNIFESP and iCOPE datasets were merged for training and validating the Deep-Learning models, whereas iCOPEvid was reserved solely for testing, providing a verification task that is closer to clinical reality. To ensure robust evaluation and prevent data leakage during training, we employed the cross-validation method known as leave-some-subjects-out [27]. In this approach, the neonates were evenly divided into ten folds. During each iteration of cross-validation, one fold was set aside for validating the model, while the remaining folds were used for training. For each training fold, we applied standard data augmentation techniques, presented in Appendix A. All models were trained in a standard binary format, employing a single output neuron activated by a Sigmoid function. This neuron predicts the probability, ranging from zero to one, of the prediction belonging to the positive class, namely “pain”. For additional details on the training protocol, refer to Appendix B.

Table 2
Classification performance on validation and test datasets.

Datasets	Metrics	VGG-Face	N-CNN	ViT-B/16	*p*-value
Validation results on iCOPE + UNIFESP	Accuracy	88.6% ± 6.0%	83.6% ± 7.2%	82.9% ± 4.2%	0.2405
	Sensitivity	88.4% ± 9.0%	82.5% ± 10.8%	87.6% ± 6.9%	0.5916
	Specificity	89.3% ± 9.7%	85.3% ± 13.7%	79.5% ± 9.5%	0.0550
	AUC	93.0% ± 4.3%	89.0% ± 5.4%	89.1% ± 2.9%	0.2016
Test results on iCOPEvid	Accuracy	63.3%	66.7%	56.7%	–
	Sensitivity	85.7%	81.6%	83.7%	–
	Specificity	47.9%	56.3%	38.0%	–
	AUC	71.2%	74.0%	68.8%	–

Footnote: The *p*-values are regarding the statistical difference between models during validation. There is no standard deviation in the iCOPEvid results because only the best model across all available folds was used.

2.3. Explainability

To better understand the model's decision-making process towards its predictions, we employed two explainability methods: Grad-CAM (GC) [34] and Integrated Gradients (IG) [35]. Both are classified as attribution methods, generating heatmap masks highlighting the most crucial regions of the input image influencing the model's prediction. Nonetheless, GC and IG operate differently. GC identifies relevant regions of the image, and IG assesses the individual contribution of each pixel to the final classification.

To improve the visualization of these masks, we applied a post-processing technique proposed by Carlini et al. [29], where the color red indicates the most important image features, yellow represents moderately important features, and green highlights the least important regions of the image to the model's predictions.

2.4. Pain sign

With the goal of providing real-time and continuous insight into the occurrence probability of painful events in neonates, we introduced the pain sign real-time tracking concept. Through the analysis of facial expressions using Deep-Learning, the pain sign effectively communicates all changes in the neonate's facial expressions in a visual, time-sensitive, and continuous manner to the health professionals.

The pain sign is two-dimensional, with the x-axis representing time and the y-axis indicating the predicted probability outputted by the Deep-Learning model for the "pain" class. For each frame captured in a video recording, a prediction probability is associated with that frame. If no face is detected during the video, the model generates no prediction. In such cases, we employed linear interpolation to fill in the missing values. As a final step, we applied a moving average filter of 30 samples, equaling to 1-s of video, to reduce potential noise.

2.5. Evaluation metrics and statistical analysis

For each validation fold, we assessed the models' performance using the following metrics with standard deviation: accuracy, sensitivity, specificity, and AUC (Area Under the receiver operating characteristic Curve). To determine the optimal threshold between "pain" and "no pain" predictions, we utilized Youden's Index [36], calculated during the training folds to prevent information leakage. The statistical differences between models were evaluated using the Friedman test with a significance level set at $p < 0.05$.

For the temporal evaluation of the pain sign, we computed the average prediction probability, which served as the final classification of the videos. Additionally, we calculated entropy as a measure of sign unpredictability [37], and determined the total number of classification threshold crossings indicating transitions between facial expressions from "pain" to "no pain", and vice versa. For comparability between pain signs, we used the Pearson correlation coefficient (r). To further analyze the pain patterns represented in the pain signs, we applied the well known k-means unsupervised learning clustering algorithm using as input the proposed entropy and threshold crossing metrics for each video. The optimal number of clusters was determined experimentally by varying the cluster count from two to ten and selecting the configuration with the highest silhouette score [38].

3. Results

Table 2 presents classification results on iCOPE and UNIFESP validation folds. VGG-Face achieved higher classification metrics than N-CNN and ViT-B/16, although when looking at the overall performance, there was no relevant statistical significance in any metric. Nevertheless, the models exhibited higher levels of standard deviation, meaning that these models are highly susceptible to the neonates inside each fold.

After validation, only the top-performing VGG-Face, N-CNN and ViT-B/16 models were chosen to generate the pain sign, specifically the models with the highest AUC across all folds. When assessing the performance of these models on 20-s video recordings condensed into a single label, and consequently one prediction, the metrics in Table 2 indicate a decrease in performance compared to the validation data. For instance, the best AUC obtained by Brahnam et al. [21] was 79.8%, against ours VGG-Face 71.2%, N-CNN 74.0%, and ViT-B/16 68.8%.

Despite the high sensitivity at 80% level, the specificity suggests a high number of false positives. Initially, these results might appear indicative of overfitting during training. However, upon incorporating the temporal dimension into the predictions, we observed that the occurrence of false positives was associated with frames where newborns exhibited facial expressions of pain, despite the video being labeled as "no pain". Moreover, the facial expressions of newborns varied throughout the 20-s of video, contradicting the single label assigned to the data.

When addressing the pain signs, for all models, the optimal number of clusters was three. These clusters were then labeled as stable, irregular, and unstable based on their distinct characteristics. The results, including classification and temporal metrics for each sign type, are summarized in Table 3. Examples of the pain signs are shown in Figs. 1, 2, 3, and 4. In these figures, the colored lines depict the evolution of pain predicted probability over the 20-s duration of the video for each model. The markers indicate the moments when the pain sign crosses the classification threshold specific to each model. Below each sign, video frames are displayed at 1-s intervals, accompanied by their corresponding GC and IG heatmap masks. The legend on top of each figure includes, in order: the ground-truth label of the video, each model's average probability and final prediction, the calculated entropy, and the number of threshold crossings.

Stable signs, illustrated in Fig. 1, exhibited low entropy and few threshold crossings, remaining relatively constant throughout the video. These signs were characterized by high-confidence predictions close to zero, indicating a period of rest, or close to one, suggesting a prolonged painful event. The irregular pain sign, portrayed in Fig. 2, showed higher entropy but few threshold crossings, indicating irregularities in the models' predictions but without sufficient confidence to definitively switch between "pain" or "no pain". These irregularities

Table 3
Classification and temporal metrics for each type of pain sign and model.

Model	Sign type	Accuracy	Sensitivity	Specificity	AUC	Entropy[a]	Threshold crossings[a]	Samples Pain[b]	Samples No Pain[b]
VGG-Face	Stable	70.6%	93.8%	50.0%	70.9%	0.35 ± 0.46	0.16 ± 0.37	32	36
	Irregular	60.6%	69.2%	55.0%	50.8%	1.75 ± 0.64	2.52 ± 0.91	13	20
	Unstable	42.1%	75.0%	33.3%	68.3%	2.56 ± 0.46	6.79 ± 2.30	4	15
N-CNN	Stable	72.2%	89.5%	62.9%	83.6%	0.75 ± 0.66	0.11 ± 0.32	19	35
	Irregular	71.4%	80.0%	63.6%	75.2%	2.31 ± 0.47	2.43 ± 0.89	20	22
	Unstable	45.8%	70.0%	28.6%	50.7%	2.61 ± 0.52	6.75 ± 1.42	10	14
ViT-B/16	Stable	71.9%	89.7%	57.1%	74.9%	1.23 ± 0.79	0.19 ± 0.39	29	35
	Irregular	32.4%	69.2%	9.5%	33.0%	2.64 ± 0.46	3.00 ± 1.13	13	21
	Unstable	50.0%	85.7%	33.3%	61.0%	2.62 ± 0.39	8.00 ± 2.26	7	15

Footnote:
[a] Data are in mean ± standard deviation.
[b] The samples for "Pain" and "No Pain" are based on the original labels of the videos.

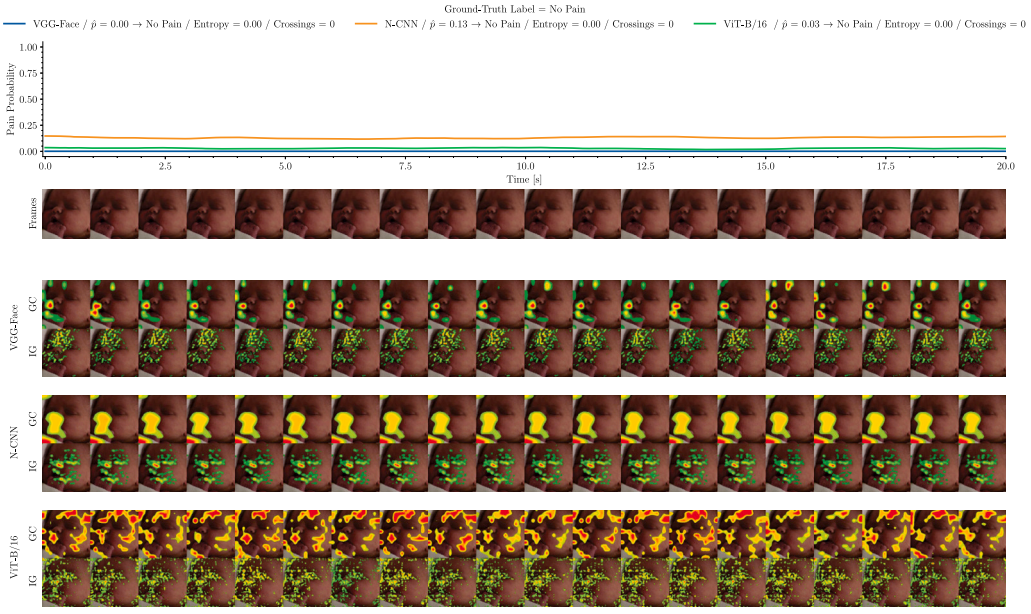


Fig. 1. Examples of stable pain signs.

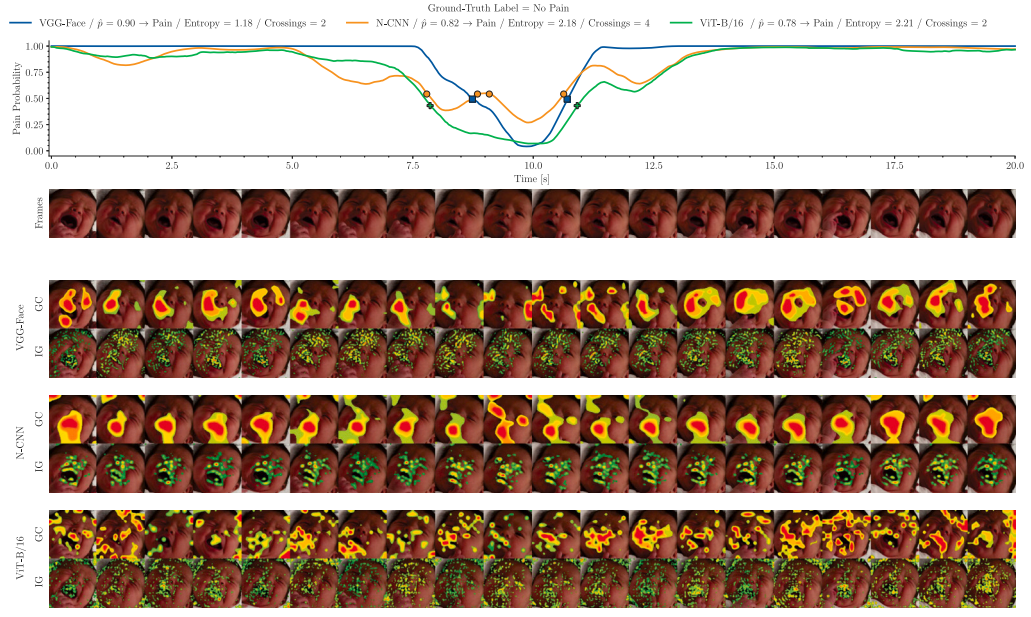


Fig. 2. Examples of irregular pain signs.

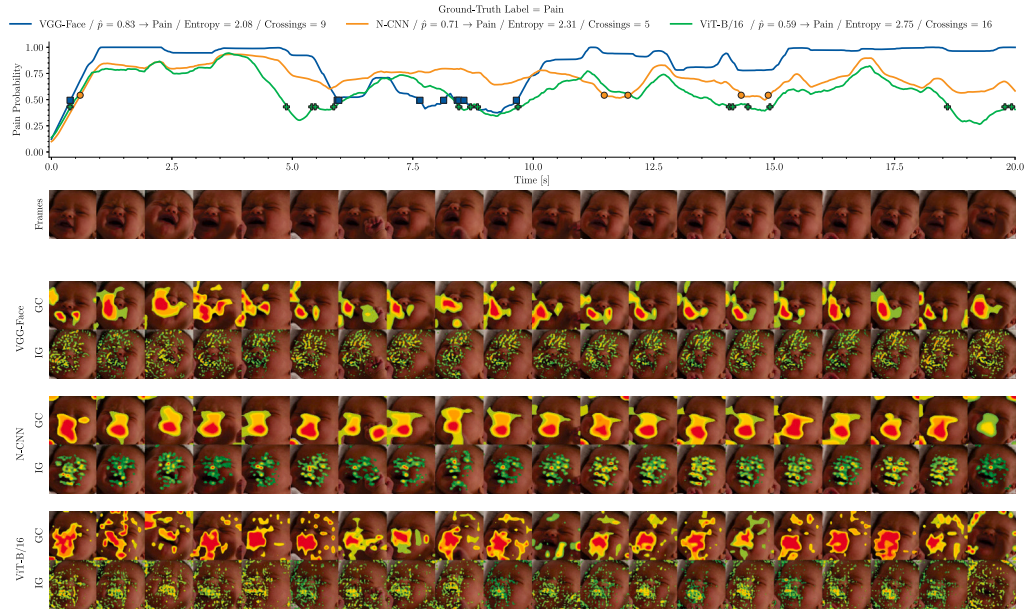


Fig. 3. Examples of unstable pain signs.

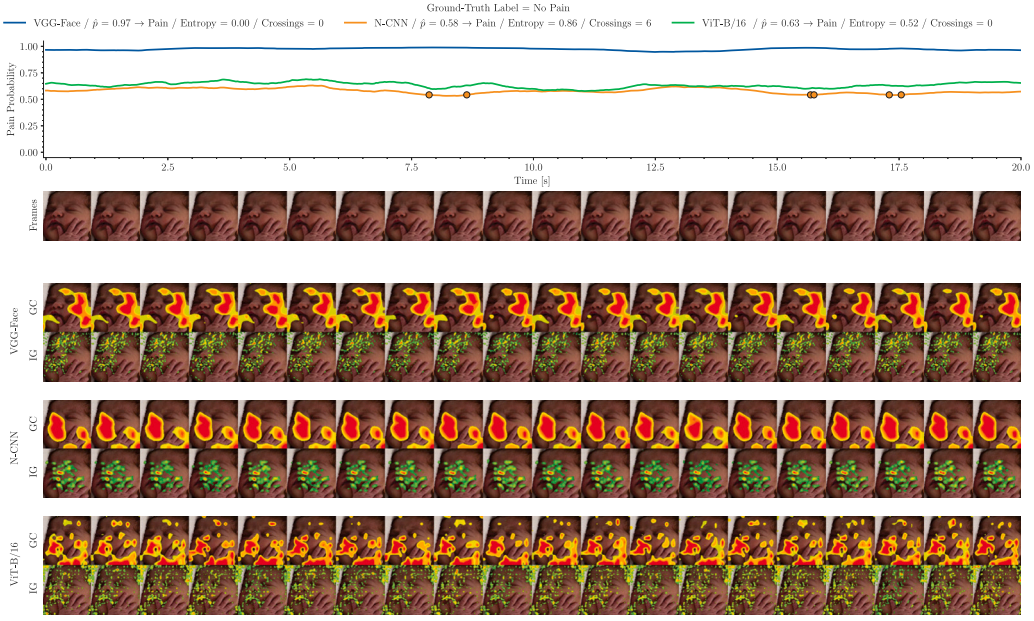
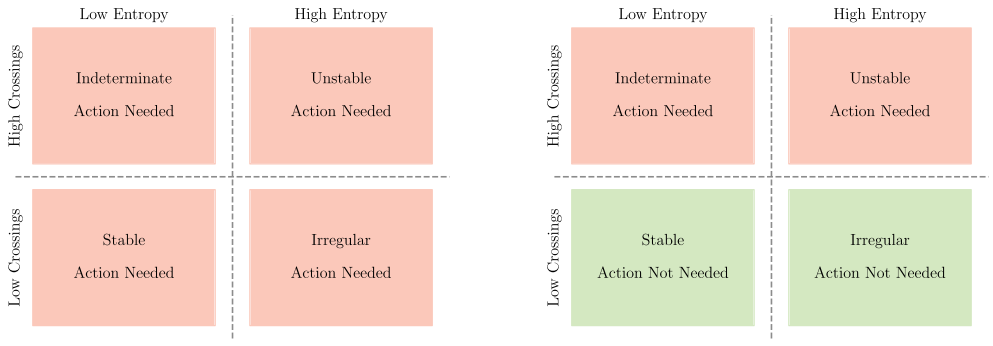


Fig. 4. Example of N-CNN indeterminate pain sign.

were often caused by hands occluding the face or sudden variations in facial expression. The unstable signs in Fig. 3 were defined by high entropy and numerous threshold crossings, where the models detected repeated painful and non-painful events, which could be interpreted as highly dynamic event or prediction instability.

As shown in Table 3, all models performed better when evaluating stable pain signs, reflecting the consistency of facial expressions throughout the video. If the prediction indicates "pain", immediate action by healthcare professionals is required. In contrast, irregular pain signs, characterized by fluctuations that impact the final prediction, reduced the models' predictive power. These signs also demand healthcare intervention, particularly if "pain" is predicted. For more complex pain signs, such as unstable ones, classification metrics declined significantly, as a single prediction or label could not adequately represent the constantly varying facial expressions, necessitating human supervision for both "pain" and "no pain" predictions. We translated these findings into decision matrices showed in Figs. 5(a) and 5(b), which guide healthcare professionals on whether to take action for specific pain sign types. For consistency across models, we used abstract terms such as high/low entropy and high/low threshold crossings in Figs. 5(a) and 5(b), with each sign type's cluster characteristics defined in Table 3. For instance, high entropy for VGG-Face corresponds to values within the intervals for irregular (1.75 ± 0.64) and unstable (2.56 ± 0.46) signs.

In fact, a fourth type of sign is also possible, as observed in the upper left quadrant of the decision matrices in Figs. 5(a) and 5(b). The fourth pain sign, referred to here as "indeterminate", is characterized by a high number of threshold crossings but low entropy, indicating a sign that fluctuates around the classification threshold without a clear designation as "pain" or "no pain". Only one sign from N-CNN met this criterion, as shown in Fig. 4. Due to the limited number of



(a) Pain predictions. (b) No pain predictions.

Fig. 5. Decision matrices for actionable responses for each type of pain sign.

samples for this pain sign type, it was not included in Table 3. Even though the ViT-B/16 pain sign closely resembles that of N-CNN, its lower classification threshold of 43.1%, compared to N-CNN's 54.2%, resulted in no detected threshold crossings, leading to its categorization as stable. Additionally, the mean predicted probability of pain for N-CNN is 58%, which is closer to the classification threshold, making it more indicative of an indeterminate prediction.

In Figs. 1, 2, 3, and 4, we also present the results of the GC and IG explainability methods. For VGG-Face, both GC and IG highlighted the entire face of the neonate as important. However, when the "pain" probability nears zero, GC emphasized peripheral regions of the image, as no face regions were deemed significant for classification. Conversely, for N-CNN, the mouth of the neonate consistently emerged as a region of interest, although GC also identified some regions outside the face as relevant. For ViT-B/16, the GC resulting masks appear more dispersed, due to its transformer-based architecture, which relies on a global attention mechanism rather than the localized operations typical of CNNs. Additionally, the ViT-B/16 used in this study was pre-trained on ImageNet, which lacks domain-specific features like neonatal facial expressions reflecting general patterns from pre-training rather than task-specific focus. However, we observed a tendency for the ViT-B/16 model to highlight the mouth region, especially when predicting pain.

Interestingly, IG remained mostly static for all models, highlighting the same face regions regardless of the predicted probability. Additionally, it is worth noting that GC is highly sensitive to the model's predicted pain probability, unlike IG. This granularity can offer real-time insights for healthcare professionals into why the pain probability fluctuates across different time intervals.

Even with comparable classification performance, the average Pearson correlation between signs was moderate: $r = 0.52 \pm 0.36$ for VGG-Face and N-CNN, $r = 0.49 \pm 0.33$ for VGG-Face and ViT-B/16, and $r = 0.57 \pm 0.37$ for N-CNN and ViT-B/16. Despite this, 60% of the videos were classified under the same sign type across models, as reflected in the number of "pain" and "no pain" samples in Table 3. Furthermore, applying the Friedman test to the entropy and threshold metrics for each model and its corresponding sign type revealed no significant differences, suggesting that the characterization of pain signs is not exclusive of the underlying model architecture.

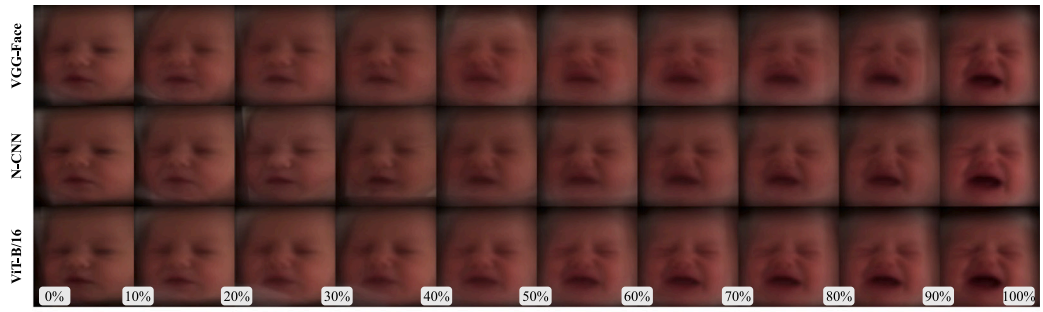
The models effectively tracked subtle changes in facial expressions, predicting "pain" probabilities in a smooth and consistent manner. These probabilities demonstrated a clear correlation with the intensity of the pain-related facial expressions. Thus, expressions with more pronounced pain characteristics are more likely to be predicted as "pain". Fig. 6 provides a frame-by-frame analysis of all videos in the iCOPEvid database, comparing the ground-truth video labels with the models' frame-level predictions. Frames where the predicted label matched the video label were classified as correct (Fig. 6(a)), whereas mismatched predictions were classified as incorrect (Fig. 6(b)). To enhance interpretability, the facial images were centered to a common reference point and averaged within probability intervals of 10%. From 0% to 50%, there is no evident facial expression of pain, but for probabilities over 50%, an increase in facial expression intensity is clear as the predicted probability rises. This finding aligns with the classification thresholds found by the Youden Index, 54.2% for N-CNN, 49.4% for VGG-Face, and 43.1% for ViT-B/16. Additionally, these figures demonstrate that a single label cannot accurately capture the nuanced facial expressions of neonates within a given time interval. This is evident in the incorrectly classified frames shown in Fig. 6(b), where consistent pain-related facial characteristics appear in "no pain" videos, and vice versa.

4. Discussion

Overall, our study highlights that while previous research on computational neonatal pain assessment has prioritized enhancing model performance using multidimensional data or developing new AI methodologies [21,23,24,26,28,30,39], practical usability in clinical settings by clinicians remains largely overlooked. By aiming to bridge the gap between computational and clinical approaches, our experimental results show the potential of an AI-derived pain sign that is both clinically comprehensible and user-friendly, providing actionable and interpretable information for healthcare professionals.

This work employed two well-established Deep-Learning models in this field, VGG-Face and N-CNN [21,23–30], along with the ViT-B/16 architecture [33], to showcase the versatility and potential of our proposed methodology across different model frameworks. We utilized three distinct benchmark datasets containing images and videos of newborns exposed to both painful and non-painful stimuli for training, validation, and testing. The validation results revealed no statistically significant differences in the classification metrics of the three models. In short, while VGG-Face, N-CNN, and ViT-B/16 achieved high AUC values during validation (93.0% ± 4.3%, 89.0% ± 5.4%, and 89.1% ± 2.9%, respectively), their performance dropped significantly when tested on 20-s video segments (71.2%, 74.0%, and 68.8%, respectively). This decline points to the limitations of assigning a single label to an entire time interval and subsequently generating only one prediction.

Similarly, Manworren et al. [40] observed that using frame-level labels obtained from nurses, significantly improved AUC values to 97%. However, 12% of frames labeled as "pain" were classified as "no pain" by nurses, while 40% of "no pain" frames were labeled as "pain". These findings highlight the inherent variability and subjectivity in pain assessment and support our results, as our models also mirror this disagreement, with lower classification performance. Thus, according to the authors, frame-level ground-truth labels are recommended for training AI models, as not all frames correspond to the single assigned



(a) Correct predictions.

(b) Incorrect predictions.

Fig. 6. Average facial expressions by predicted pain probability and model.

label to the video. Even so, obtaining clinically accurate labels for over 72,000 frames in the iCOPEvid dataset is both labor-intensive and time-consuming for clinical staff.

The results in Fig. 6 highlight the potential of the employed models to automatically generate frame-level labels, thereby facilitating future model training. This is evident from the positive correlation between the intensity of pain-related facial expressions and the predicted pain probability, similarly to the NFCS score [2], where the presence of more facial characteristics indicative of pain increases the likelihood of the neonate being in pain. The models' ability to capture this relationship, despite not being trained for regression tasks, explains visually the pain sign contribution. Consequently, the pain sign continuously monitors frame-by-frame subtle changes in neonatal facial expressions in real-time, offering a visual, time-sensitive explanation of pain probability that assists healthcare professionals in managing NICU workloads.

While similar signs have been identified in adult pain estimation research [41–43], to the best of our knowledge, this study is the first to apply such an analysis to the neonatal domain. Furthermore, we improved the interpretability of these signs by incorporating two metrics – entropy and threshold crossings – that quantitatively defined three distinct and novel sign types: stable, irregular, and unstable. We also introduced a potential fourth type, the indeterminate pain sign. This temporal classification scheme, though separately defined for each model, exhibited consistent characteristics across all models, suggesting that sign types are not exclusive of the underlying AI model architecture. However, each model demonstrated varying interpretive capabilities when analyzing videos from the same time intervals, often leading to different predictions and sign types, resembling the inter-rater variability observed among human evaluators in the neonatal pain assessment task [11–13].

For clinical practice, we translated these computational and mathematical findings into decision matrices, illustrated in Fig. 5, to guide healthcare professionals. The matrices suggest prioritizing actions based on the following observations: (I) Stable signs predicting painful events, where the neonate's facial expression remains consistent, indicating a prolonged painful state; (II) Irregular signs predicting painful events, similar to stable signs but with facial occlusions or sudden variations in expression; (III) Unstable signs, regardless of the final prediction, representing highly dynamic events that the model cannot reliably predict; and (IV) Indeterminate signs, also regardless of the final prediction, due to significant uncertainty in the model's outputs. During clinical practice, the pain sign would automatically be assigned to one of these categories using the clustering algorithm proposed here. Additionally, triggers can be implemented to detect sudden surges or drops in the pain sign, promptly alerting healthcare professionals to changes in the neonate's condition.

However, it is important to highlight that the 20-s duration of iCOPEvid videos, which were recorded in a controlled environment, are not fully representative of real-world NICU scenarios, where neonates often have medical apparatus obscuring their faces and need to be monitored full-time. Our approach can be adapted to longer videos by segmenting them into smaller intervals and analyzing sign behavior over extended periods, such as the past minute, hour, or even longer. Therefore, it might be possible to align our temporal classification scheme with clinical terms of acute, prolonged, persistent, or chronic pain [18].

5. Conclusion

To the best of your knowledge, this retrospective multi-dataset study is the first to implement an AI-derived pain sign for continuous tracking of neonatal facial expressions with a novel set of temporal classifications designed specifically for healthcare professionals usage during clinical practice.

Unlike previous methods that offer single predictions and lack NICU-specific interpretability, our approach provides continuous, real-time insights into neonates' current and historical pain states. Bridging computational and clinical needs, it delivers actionable, user-friendly information with minimal training, supporting timely decisions and improving care in NICU settings. Furthermore, our approach can be extended to other AI models in the pain assessment domain that generate a prediction probability.

We believe that the proposed pain sign has the potential to serve as a vital sign, seamlessly integrating into vital sign monitors for ongoing visualization and monitoring of newborns. By categorizing pain signs into distinct types, healthcare professionals can effectively screen neonates and gain deeper insights into individual pain patterns experienced by newborns, enabling personalized treatment. With the implementation of the pain sign, we can advance towards ensuring optimal pain management and promoting the well-being of these vulnerable patients.

CRediT authorship contribution statement

Leonardo Antunes Ferreira: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lucas Pereira Carlini:** Writing – review & editing, Validation, Methodology, Formal analysis. **Gabriel de Almeida Sá Coutrin:** Writing – review & editing, Validation, Methodology, Formal analysis. **Tatiany Marcondes Heiderich:** Data curation. **Rita de Cássia Xavier Balda:** Project administration, Data curation, Conceptualization. **Marina Carvalho de Moraes Barros:** Project administration, Data curation, Conceptualization. **Ruth Guinsburg:** Writing – review & editing, Project administration, Funding acquisition, Data curation, Conceptualization. **Carlos Eduardo Thomaz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Ethics statement

All data used are benchmark databases already approved by the appropriate institutional committees.

Code availability

All codes are open source and available at https://github.com/leferr-code/article. Trained models can be requested to the corresponding author.

Funding

This work was supported by FAPESP (2018/13076-9) and CAPES (Finance Code 001).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Leonardo Antunes Ferreira reports financial support was provided by Coordination of Higher Education Personnel Improvement. Carlos Eduardo Thomaz reports financial support was provided by State of Sao Paulo Research Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Centro Universitário FEI, and the Brazilian funding agencies FAPESP and CAPES.

Appendix A. Data augmentation

Data augmentation is the process of generating artificially new data from existing data and is commonly used in Deep-Learning to prevent overfitting and regularizing the training procedure [44]. For our purposes, the available face images went through a series of transformations altering its characteristics, including rotation (30°), shear (0.15), width and height shift (0.20), brightness adjustment (0.50–1.1), zoom (0.70–1.5), and horizontal flip. These augmentations resulted in 20 new images for each original image. We ensured the neonate's face was fully detectable by applying the RetinaFace detector [22].

Appendix B. Training protocol

During training, the value of the Cross-Entropy error function was used to determine the convergence or divergence between training and validation. In each epoch of the 100 epochs scheduled for training, if the error value decreased in the validation data, a model checkpoint was saved. Otherwise, if there was no improvement in this error for five consecutive epochs, training was stopped. Only the best model, that is, the one that obtained the lowest error value in the validation data during training was used in the following steps. In the case of the VGG-Face, if the error did not improve for five consecutive epochs, the fine-tuning process started, allowing the training of the last two convolution layers groups. For ViT-B/16, only the classifier head was trained, and N-CNN was trained from scratch.

We also used the DropOut technique during training, for the VGG-Face the value was set at 50%, for N-CNN we followed what Zamzmi et al. [32] proposed, and for ViT-B/16 we used a rate of 10%. For all models, the batch size was fixed on 16, the optimizer used was RMSProp with a learning rate of $\eta = 1 \times 10^{-4}$.

References

[1] K.J. Anand, P.R. Hickey, Pain and its effects in the human neonate and fetus, N. Engl. J. Med. 317 (21) (1987) 1321–1329.
[2] R.E. Grunau, K.D. Craig, Pain expression in neonates: facial action and cry, Pain 28 (3) (1987) 395–410.
[3] K.J. Anand, D.B. Carr, The neuroanatomy, neurophysiology, and neurochemistry of pain, stress, and analgesia in newborns and children, Pediatr. Clin. North Am. 36 (4) (1989) 795–822.
[4] M. Perry, Z. Tan, J. Chen, T. Weidig, W. Xu, X.S. Cong, Neonatal pain: perceptions and current practice, Crit. Care Nurs. Clin. 30 (4) (2018) 549–561.
[5] K.J. Anand, M. Eriksson, E.M. Boyle, A. Avila-Alvarez, R.D. Andersen, K. Sarafidis, T. Polkki, C. Matos, P. Lago, T. Papadouri, et al., Assessment of continuous pain in newborns admitted to NICU's in 18 European countries, Acta Paediatr. 106 (8) (2017) 1248–1259.
[6] R.E. Grunau, Neonatal pain in very preterm infants: long-term effects on brain, neurodevelopment and pain reactivity, Rambam Maimonides Med. J. 4 (4) (2013) e0025.
[7] P. Hummel, M. Puchalski, S. Creech, M. Weiss, Clinical reliability and validity of the N-PASS: neonatal pain, agitation and sedation scale with prolonged pain, J. Perinatol. 28 (1) (2008) 55–60.
[8] J. Lawrence, D. Alcock, P. McGrath, J. Kay, S.B. MacMurray, C. Dulberg, The development of a tool to assess neonatal pain., Neonatal Netw.: NN 12 (6) (1993) 59–66.
[9] D.J. Crellin, D. Harrison, N. Santamaria, F.E. Babl, Systematic review of the face, legs, activity, cry and consolability scale for assessing pain in infants and children: is it reliable, valid, and feasible for use? Pain 156 (11) (2015) 2132–2151.
[10] M. Schiavenato, J.F. Byers, P. Scovanner, J.M. McMahon, Y. Xia, N. Lu, H. He, Neonatal pain facial expression: Evaluating the primal face of pain, Pain 138 (2) (2008) 460–471.
[11] R.X. Balda, R. Guinsburg, M.F.B. de Almeida, C. de Araújo Peres, M.H. Miyoshi, B.I. Kopelman, The recognition of facial expression of pain in full-term newborns by parents and health professionals, Arch. Pediatr. Adolesc. Med. 154 (10) (2000) 1009–1016.
[12] G.D. De Sario, C.R. Haider, K.C. Maita, R.A. Torres-Guzman, O.S. Emam, F.R. Avila, J.P. Garcia, S. Borna, C.J. McLeod, C.J. Bruce, et al., Using AI to detect pain through facial expressions: A review, Bioeng. 10 (5) (2023) 548.
[13] A. Llerena, K. Tran, D. Choudhary, J. Hausmann, D. Goldgof, Y. Sun, S.M. Prescott, Neonatal pain assessment: Do we have the right tools? Front. Pediatr. 10 (2023) 1022751.

[14] M.D. Cruz, A. Fernandes, C. Oliveira, Epidemiology of painful procedures performed in neonates: a systematic review of observational studies, Eur. J. Pain 20 (4) (2016) 489–498.
[15] G. Zamzmi, R. Kasturi, D. Goldgof, R. Zhi, T. Ashmeade, Y. Sun, A review of automated pain assessment in infants: features, classification tasks, and databases, IEEE Rev. Biomed. Eng. 11 (2017) 77–96.
[16] T.M. Heiderich, L.P. Carlini, L.F. Buzuti, R.C. Balda, M.C. Barros, R. Guinsburg, C.E. Thomaz, Face-based automatic pain assessment: challenges and perspectives in neonatal intensive care units, J. Pediatr. (RioJ) 99 (6) (2023) 546–560.
[17] M.S. Salekin, P.R. Mouton, G. Zamzmi, R. Patel, D. Goldgof, M. Kneusel, S.L. Elkins, E. Murray, M.E. Coughlin, D. Maguire, et al., Future roles of artificial intelligence in early pain management of newborns, Paediatr. Neonatal Pain 3 (3) (2021) 134–145.
[18] K.J. Anand, Defining pain in newborns: need for a uniform taxonomy? Acta Paediatr. 106 (9) (2017) 1438–1444.
[19] S. Brahnam, C.-F. Chuang, F.Y. Shih, M.R. Slack, Machine recognition and representation of neonatal facial displays of acute pain, Artif. Intell. Med. 36 (3) (2006) 211–222.
[20] T.M. Heiderich, A.T.F.S. Leslie, R. Guinsburg, Neonatal procedural pain can be assessed by computer software that has good sensitivity and specificity to detect facial movements, Acta Paediatr. 104 (2) (2015) e63–e69.
[21] S. Brahnam, L. Nanni, S. McMurtrey, A. Lumini, R. Brattin, M. Slack, T. Barrier, Neonatal pain detection in videos using the iCOPEvid dataset and an ensemble of descriptors extracted from Gaussian of local descriptors, Appl. Comput. Inform. 19 (1/2) (2020) 122–143.
[22] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 5202–5211.
[23] C. Li, A. Pourtaherian, L. van Onzenoort, W.E.T.a. Ten, P.H.N. de With, Infant facial expression analysis: Towards a real-time video monitoring system using R-CNN and HMM, IEEE J. Biomed. Heal. Inform. 25 (5) (2021) 1429–1440.
[24] M.S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, Y. Sun, First investigation into the use of deep learning for continuous assessment of neonatal postoperative pain, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, 2020, pp. 415–419.
[25] L.P. Carlini, L.A. Ferreira, G.A. Coutrin, V.V. Varoto, T.M. Heiderich, R.C. Balda, M.C. Barros, R. Guinsburg, C.E. Thomaz, A convolutional neural network-based mobile application to bedside neonatal pain assessment, in: 2021 34th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI, IEEE, 2021, pp. 394–401.
[26] M.S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, Y. Sun, Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment, Comput. Biol. Med. 129 (2021) 104150.
[27] G.A. Coutrin, L.P. Carlini, L.A. Ferreira, T.M. Heiderich, R.C. Balda, M.C. Barros, R. Guinsburg, C.E. Thomaz, Convolutional neural networks for newborn pain assessment using face images: A quantitative and qualitative comparison, in: R. Su, Y. Zhang, H. Liu, A. F. Frangi (Eds.), International Conference on Medical Imaging and Computer-Aided Diagnosis, Springer Nature Singapore, Singapore, 2023, pp. 503–513.
[28] Y. Zhao, H. Zhu, X. Chen, F. Luo, M. Li, J. Zhou, S. Chen, Y. Pan, Pose-invariant and occlusion-robust neonatal facial pain assessment, Comput. Biol. Med. 165 (2023) 107462.
[29] L.P. Carlini, G.A. Coutrin, L.A. Ferreira, J. do Carmo Azevedo Soares, G.V.T. Silva, T.M. Heiderich, R.C. Balda, M.C. Barros, R. Guinsburg, C.E. Thomaz, Human vs machine towards neonatal pain assessment: A comprehensive analysis of the facial features extracted by health professionals, parents, and convolutional neural networks, Artif. Intell. Med. 147 (2024) 102724.
[30] J. Hausmann, M.S. Salekin, G. Zamzmi, P.R. Mouton, S. Prescott, T. Ho, Y. Sun, D. Goldgof, Accurate neonatal face detection for improved pain classification in the challenging NICU setting, IEEE Access 12 (2024) 49122–49133.
[31] O. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: BMVC 2015-Proceedings of the British Machine Vision Conference 2015, British Machine Vision Association, 2015.
[32] G. Zamzmi, R. Paul, D. Goldgof, R. Kasturi, Y. Sun, Pain assessment from facial expression: Neonatal convolutional neural network (N-CNN), in: 2019 International Joint Conference on Neural Networks, IJCNN, IEEE, 2019, pp. 1–7.
[33] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[34] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 618–626.
[35] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.
[36] W.J. Youden, Index for rating diagnostic tests, Cancer 3 (1) (1950) 32–35.
[37] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[38] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[39] G. Zamzmi, C.-Y. Pai, D. Goldgof, R. Kasturi, T. Ashmeade, Y. Sun, A comprehensive and context-sensitive neonatal pain assessment using computer vision, IEEE Trans. Affect. Comput. 13 (1) (2022) 28–45.
[40] R.C. Manworren, S. Horner, R. Joseph, P. Dadar, N. Kaduwela, Performance evaluation of a supervised machine learning pain classification model developed by neonatal nurses, Adv. Neonatal Care 24 (3) (2024) 301–310.
[41] J. Zhou, X. Hong, F. Su, G. Zhao, Recurrent convolutional neural network regression for continuous pain intensity estimation in video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 84–92.
[42] S. Thuseethan, S. Rajasegarar, J. Yearwood, Deep hybrid spatiotemporal networks for continuous pain intensity estimation, in: Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26, Springer, 2019, pp. 449–461.
[43] D. Erekat, Z. Hammal, M. Siddiqui, H. Dibeklioğlu, Enforcing multilabel consistency for automatic spatio-temporal assessment of shoulder pain intensity, in: Companion Publication of the 2020 International Conference on Multimodal Interaction, in: ICMI '20 Companion, Association for Computing Machinery, New York, NY, USA, 2021, pp. 156–164.
[44] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.