



Automated Young Children's Pain Detection via Facial Expressions with YOLO v11

Gusgus Ghraha Ramdhanie¹, Nurdina Widanti², Bambang Aditya Nurgraha³, Tomy Abuzairi⁴, Nur Agustini⁵, Dessie Wanda⁶

¹Department of Pediatric and Fundamental Nursing, Faculty of Nursing, Universitas Padjadjaran, Bandung, Indonesia

²Doctoral Nursing Program, Faculty of Nursing, Universitas Indonesia, Depok, Indonesia

³Department of Electrical Engineering, Politeknik Negeri Jakarta, Depok, Indonesia

⁴Department of Critical and Medical Surgical Nursing, Faculty of Nursing, Universitas Padjadjaran, Bandung, Indonesia

⁵Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, Indonesia

⁶Department of Pediatric Nursing, Faculty of Nursing, Universitas Indonesia, Depok, Indonesia

¹gusgus.ghraha.ramdhanie@unpad.ac.id, ²nurdina.widanti@lecturer.pnj.ac.id, ³bambang14005@unpad.ac.id,

⁴tomy.abuzairi@ui.ac.id, ⁵nagustini2015@gmail.com, ⁶dessie@ui.ac.id

Abstract

This study demonstrates that pain detection in young children using a YOLO v11-based deep learning model can be performed effectively. By utilizing image data taken from video recordings of immunization and IV infusion procedures, then processed into photo frames and labeled using Roboflow, the model is able to provide good evaluation results. The dataset was divided into 70:20:10 for training, validation, and testing. Model performance evaluation uses accuracy, precision, recall, and F1-score metrics, and is visualized through a performance curve and confusion matrix. The results show that YOLO v11 has great potential as a pain detection method, with an mAP@0.5 achievement of 0.893, an accuracy of 78%, a precision of 89.3%, a recall of 97%, and an F1-score of 83%. The high recall value indicates the model's excellent ability to recognize pain expressions, making it relevant for use in clinical contexts to ensure pain symptoms are not overlooked. Overall, this performance demonstrates that YOLO v11 can be a more objective and accurate approach than manual instruments, and has the potential to be developed as a tool for healthcare professionals in pediatric pain assessment.

Keywords: deep learning; facial expressions; pain detection; YOLO v11; young children.

How to Cite: G. G. Ramdhanie, Nurdina Widanti, Bambang Aditya Nurgraha, Tomy Abuzairi, Nur Agustini, and Dessie Wanda, "Automated Young Children's Pain Detection via Facial Expressions with YOLO v11", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 1, pp. 86 - 94, Feb. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i1.7206>

Received: September 25, 2025

Accepted: February 3, 2026

Available Online: February 18, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Indonesian Informatics Experts Association](#)

1. Introduction

Children often undergo medical procedures that can cause pain during their hospital stay [1]. The experience of pain in children during hospital treatment can cause stress and psychological and emotional disorders [2]. In addition, pain impacts the child's daily life, such as difficulty sleeping, physical disorders, difficulty interacting, and disruption of the child's activities [3]. In clinical practice, most pain episodes experienced by hospitalized children—such as venipuncture, immunization, wound care, intravenous cannulation, or postoperative discomfort—fall into the category of acute pain, which arises suddenly, has identifiable causes, and requires prompt evaluation to prevent escalation into more severe distress.

Nurses have an important role in determining the appropriate pain assessment method in children.[4]The first step that can be taken is to conduct a comprehensive pain assessment to obtain an overview of the intensity and characteristics of the pain felt by the child [5]. Effective pain assessment has been shown to reduce various pediatric problems, making accurate pain assessment crucial. Pain assessment serves as the basis for establishing a diagnosis, selecting therapy, and evaluating management in pediatric patients [3]. However, children often have difficulty expressing pain verbally [6]. This condition demands an alternative approach through observation of behavior and non-verbal expressions, including facial expressions, as indicators of pain.

Pain assessment in children remains a global challenge, including in Indonesia. One of the challenges is the limited availability of valid and specific pain assessment methods for children. While manual or paper-based pain assessment instruments are available, they are often underutilized due to perceived ineffectiveness in clinical practice [6]. On the other hand, digital health technology for clinical data collection has proven to be more efficient than paper-based methods [7], including in the evaluation of pain in children [8]. This technological development has encouraged the birth of digital-based pain assessment instruments [9]. One of them is facial detection technology which has the potential to identify and analyze subtle features of facial expressions related to pain [10].

Recognizing facial expressions of pain provides a basis for more objective evaluation, allowing healthcare professionals to more accurately understand and manage children's pain [11]. Considering that children often express pain through facial expressions [12], [13], [14], automated recognition systems offer significant advantages in improving the accuracy of pediatric pain assessment compared to manual methods [15].

However, existing facial expression recognition approaches still have limitations. Most studies have focused on adult populations [16], while specific and validated instruments for children are still very limited [17]. In addition, although facial detection technology has been applied, its accuracy and reliability in pediatric populations still need to be improved [18]. This situation indicates a research gap in the development of more sophisticated, efficient, and child-specific facial recognition models.

Recent developments in the field of deep learning, particularly object detection models such as YOLO (You Only Look Once) [19], [20], opening up opportunities for real-time facial analysis with a high degree of accuracy [21], [22]. YOLO is a convolutional neural network (CNN) based object detection method first developed by Joseph Redmon [23-25]. This algorithm works by dividing the image into $s \times s$ grids to predict bounding boxes and class maps on each grid, thus enabling fast and efficient object detection [26].

Previous research has shown the potential of applying YOLO in facial expression detection [27]. Changes in facial expression such as lowered eyebrows, a square-shaped mouth, and tightly closed eyes can be indicators of pain in babies [24]. These facial changes are then trained using deep learning algorithms like YOLO to facilitate pain detection in infants. This implementation typically uses the PyTorch framework, which supports GPU computing with CUDA, enabling more optimal and efficient image processing [26].

In addition, several other studies have developed video recording-based methods analyzed with the Non Verbal Pain Scale Revised (NVPSR) which has a strong correlation with the Wong-Baker Pain Scale (WBPS),

although its sensitivity is quite high only as a screening tool and less specific for pain diagnosis [6]. There is also an approach that combines facial and vocal data using machine learning methods with an SVM classifier to distinguish between pain and no-pain conditions, as well as classifying infant facial responses with CNN when experiencing pain [26].

However, these studies have focused on infants and are limited to previous generations of deep learning models. The latest version, YOLO v11, offers significant improvements over previous versions, both in processing speed, architectural optimization, and generalization capabilities on complex datasets [21], [23], [28], [29]. This advantage makes YOLO v11 very potential for application in detecting facial expressions of pain in young children, which requires a system with high sensitivity to subtle and rapid changes in facial expressions [24]. To date, research specifically implementing YOLO v11 for pediatric pain assessment in young children is still very limited, thus providing an opportunity to make new contributions to this field.

YOLOv11 introduces a series of architectural innovations that significantly enhance its effectiveness in complex visual recognition tasks, such as facial expression analysis in pediatric pain detection. The model incorporates a modified C3k2 block, which improves parameter efficiency by simplifying layer connections while preserving gradient flow. It also includes the SPPF (Spatial Pyramid Pooling – Fast) module, which facilitates multi-scale feature extraction with lower computational load. Moreover, the integration of C2PSA (Parallel Spatial Attention) allows the model to focus more accurately on spatially critical facial regions like the eyebrows, mouth corners, and eyes—areas known to reflect pain in children [28]. These improvements result in enhanced generalization across imbalanced and noisy datasets, as well as reduced inference latency—making YOLO v11 more efficient and robust than YOLO v8 or YOLO v10 in real-time clinical applications [21], [23], [28].

In addition, previous research on infants generally still limits the classification of facial expressions to simple categories, for example neutral, sad, and painful, with monitoring results that can be displayed via a web-based system to support health workers in initial monitoring [24]. However, this categorical approach is not yet able to describe variations in pain levels in greater detail, even though a more granular pain assessment is urgently needed in clinical practice to determine appropriate interventions in pediatric patients.

Therefore, this study focuses on developing a YOLO v11-based deep learning model to detect acute pain levels in young children with a score range of 0–10. Despite technological advances, no existing model has provided (1) child-specific validation, (2) real-time multi-level pain scoring, and (3) implementation using the latest YOLO v11 architecture. This gap highlights a clear problem: current pediatric pain assessment tools

are not yet objective, scalable, nor sufficiently accurate for acute clinical decision-making. This study aims to address that problem by providing a more sensitive, reliable, and clinically applicable automated pain assessment solution.

2. Methods

This research was designed to develop a pain detection system for young children through facial expressions using a deep learning approach based on YOLO v11. The process began with data collection in the form of video recordings of children aged 1–5 years undergoing immunization and IV insertion procedures at community health centers and hospitals. The recordings were then cut into several image frames for easier processing.

Each resulting image was then annotated using the Roboflow platform to classify the young children's facial expressions according to research needs. After the labeling stage was complete, the data underwent preprocessing and augmentation. This step was carried out to enrich the data variety and maintain model quality, ensuring it recognized more than just limited patterns. The prepared data was then divided into three groups: 70% for model training, 20% for validation, and 10% for testing. This division was intended to ensure the model could learn optimally and scalably when faced with new data[28].

Model performance was evaluated using four main metrics: Recall, Precision, F1-Score, and Accuracy. These metrics were chosen to provide a comprehensive overview of the model's ability to detect pain, balance correct and incorrect detections, and assess the overall reliability of the system [21], [23], [28]. Figure 1 shows the system block diagram.

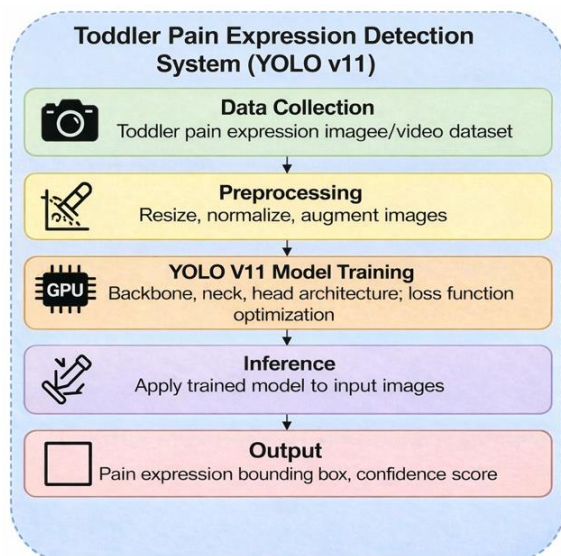


Figure 1. Block diagram

2.1 Data Collection

Image data was obtained by recording videos of young children undergoing immunization and IV infusion

procedures at community health centers and hospitals. The videos were then cut into individual photo frames, which were used as a dataset, as shown in Figure 2. One of the videos was then split per frame and created per folder according to the pain scale, as shown in Figure 3.



Figure 2. Original video before splitting and scoring



Figure 3. Dataset collection

2.2 Data Labeling and Preprocessing

The extraction results obtained 3367 datasets, which were then labeled based on the level of pain expression with a rating scale of 0–10, as shown in Figure 4, using the Roboflow application. These labels served as the ground truth for model training.

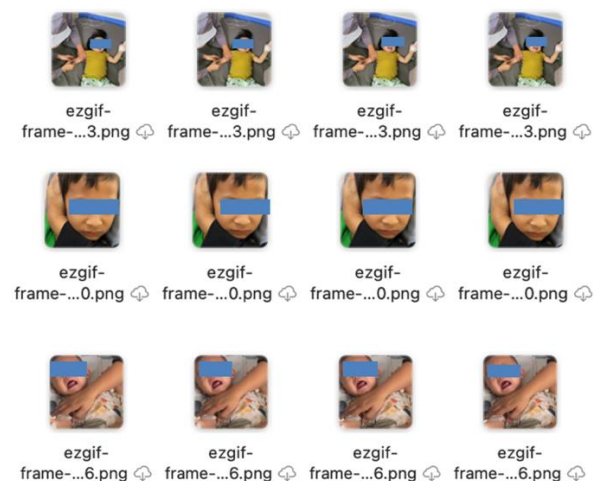


Figure 4. Labeling process

The labeled images undergo a preprocessing stage to prepare them for use in the model. This stage includes resizing, normalization, and augmentation such as rotation, lighting changes, and flipping. This step aims to enrich the data variation and ensure the model can recognize pain expressions across a wide range of conditions.

2.3 Dataset Division and Model Development

The dataset was divided into three parts: 70% for training, 20% for validation, and 10% for testing. This division was done randomly to ensure a representative distribution of the data. The YOLO v11 model was used to train a pain expression detection system. The model architecture consists of a backbone, neck, and head, which work to extract features, combine spatial information, and generate predictions. For model training procedure summarized in Table 1.

Table 1. Model training procedure summarized

Step	Description
Dataset Division	Training 70%, Validation 20%, Testing 10%, with stratified sampling to preserve class distribution.
Preprocessing	Images resized to 640×640 pixels, normalized. Augmentation included horizontal flipping, rotations, brightness and contrast adjustments.
Architecture	YOLO v11, initialized with COCO pre-trained weights, includes advanced modules for spatial attention and feature aggregation.
Hyperparameters	Batch size: 16, Initial learning rate: 0.001 (scheduled), Adam optimizer, up to 100 epochs, early stopping on validation loss.
Class Imbalance Management	Class weighting in loss function; extensive augmentation focused on minority classes.
Calibration	Confidence score calibration via temperature scaling on validation set.
Validation Strategy	Early stopping, lowest validation loss used for model selection, all metrics monitored per epoch.
Evaluation Metrics	Accuracy, precision, recall, F1-score, AP per class; analysis curves: precision-confidence, recall-confidence, F1-confidence, confusion matrix.
Implementation	Training with PyTorch 2.0 on NVIDIA RTX 3090 GPU; code and parameters saved for reproducibility.
Best Practice Justification	All steps aligned with clinical ML best practice for robustness, transparency, and reproducibility.

2.4 Inference and Model Evaluation

After the model training process was completed, the inference stage was conducted by applying the trained model to the test image dataset to predict the presence of pain expressions based on the learned feature patterns.

Subsequently, model performance was evaluated using *Recall, Precision, F1-Score, and Accuracy metrics*. The evaluation process involved both internal test datasets and additional field testing using randomly selected data to assess the model's generalization capability in real-world conditions.

Recall is defined as the ratio of the total number of correctly classified positive examples divided by the total number of positive examples. High Recall indicates that the class is correctly recognized (low FN) [30]. As in Equation 1.

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

The precision value is obtained by dividing the total number of correctly classified positive examples by the total number of predicted positive examples [30]. As in Equation 2.

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

High Recall, Low Precision: This means that most positive examples are correctly recognized (low FN) but there are many false positives. Low Recall, High Precision: This means that we missed many positive examples (high FN) but the ones we predicted as positive were actually positive (low FP).

F1 Score is defined as the harmonic mean of precision and recall, and is used to evaluate the balance between them. It is especially useful when the dataset is imbalanced. A high F1 Score indicates that both precision and recall are reasonably high [30] as shown in Equation 3.

$$F1\ score = 2 \times \frac{Recal \times precision}{Recal+precision} \quad (3)$$

In this study, the classification results are summarized using the following terms: (a). True Positive (TP): Model correctly predicts pain when pain is present. (b). False Positive (FP): Model predicts pain when no pain is present. (3). False Negative (FN): Model fails to predict pain when pain is present. (4). True Negative (TN): Model correctly predicts no pain when no pain is present.

After the training and evaluation phases were completed, the model was tested live on new data in a clinical setting to assess its real-time pain expression detection performance and its application in young children's pain assessment.

3. Results and Discussions

The results show that YOLO v11 has great potential as a pain detection method based on facial expressions in young children. The high recall value indicates that this model is highly sensitive in recognizing pain expressions, which is very important in a clinical context to ensure that no pain expressions are overlooked. However, the slightly lower precision value compared to recall suggests some errors in detecting non-pain expressions as painful. This condition is likely influenced by variations in lighting, shooting angle, or the similarity between the child's facial expressions that are not in pain and those that are in pain. Model evaluation was conducted using several metrics, namely accuracy, precision, recall, and F1-score, which are visualized in an evaluation curve and confusion matrix to provide a comprehensive overview of the model's performance. Figure 5 shows the distribution of labeled data. Figure 5 shows an uneven distribution of data instances in each class. The "not pain" category has the highest number of instances, approaching 1,000 instances, while some classes, such as pain_2 and pain_3, have fewer instances. This imbalance in the data has the potential to impact model

performance, especially for classes with limited data. Table 2 gives the information about numerical data from Figure 5.

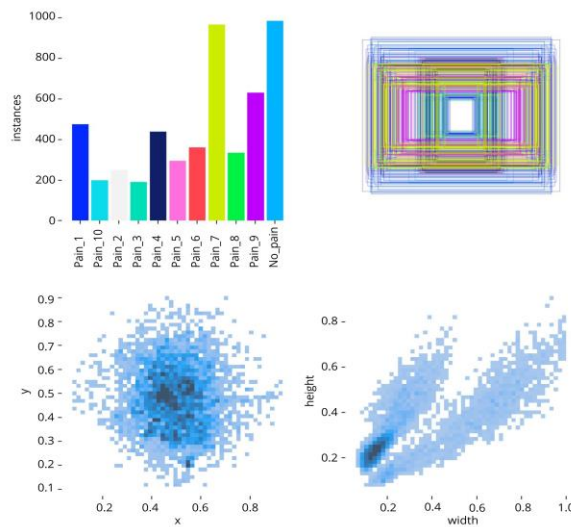


Figure 5. Data distribution

Table 2. Data distribution

Class	Number of Instances (estimated)
Pain_1	±250
Pain_2	±180
Pain_3	±350
Pain_4	±310
Pain_5	±410
Pain_6	±370
Pain_7	±990
Pain_8	±610
Pain_9	±420
Pain_10	±280
No_pain	±990

The precision curve in Figure 6 against confidence shows that precision increases with confidence up to an optimal point at a threshold of around 0.93 with an average precision approaching 1. This indicates that the model is able to reduce false positive predictions when a higher confidence level is used.

These findings indicate that filtering predictions by a high confidence score is highly effective for maximizing precision in pain identification tasks. The model excels in distinguishing between all pain classes and non-pain, achieving exceptional accuracy at stricter thresholds. Minor discrepancies in Pain_2 and Pain_8 may be attributed to limited training examples or greater facial similarity between classes. In clinical settings, this results in minimal false positives for high-confidence cases, ensuring patient safety and supporting rapid, trusted triage decisions. The results endorse the model's deployment for reliable pain detection provided that appropriate confidence calibration is maintained in production. That show on Table 3.

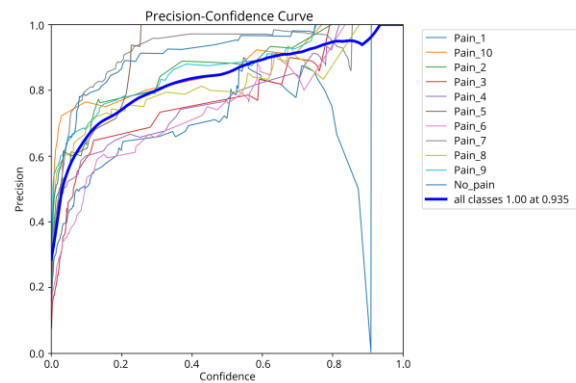


Figure 6. Precision vs confidence results

Table 3. Maximum precision per class

Class	Maximum Precision (at optimal threshold)
Pain_1	0.97
Pain_10	1.00
Pain_2	0.91
Pain_3	0.97
Pain_4	1.00
Pain_5	0.99
Pain_6	1.00
Pain_7	1.00
Pain_8	0.98
Pain_9	1.00
No_pain	1.00

In Table 4 The Precision-Recall Curve visualizes trade-offs between sensitivity (recall) and prediction accuracy (precision) across a range of classification thresholds. High AP values for major classes indicate that the model is highly effective in minimizing missed pain cases, which is crucial in pediatric healthcare. Lower AP values for classes with fewer examples or greater facial similarity highlight the impact of dataset imbalance. These gaps should be addressed in future research via data augmentation or targeted sample collection to improve minority class detection.

Table 4. Average Precision (AP) per Class at mAP@0.5

Class	Average Precision (AP)
Pain_1	0.797
Pain_2	0.866
Pain_3	0.788
Pain_4	0.913
Pain_5	0.936
Pain_6	0.839
Pain_7	0.970
Pain_8	0.847
Pain_9	0.965
Pain_10	0.966
No_pain	0.934
All classes	0.893

From a practical standpoint, the model's high overall AP ensures reliable identification of acute pain phenomena in clinical pediatric settings. However, the reduced performance for minority classes advises caution and suggests further model refinement is warranted before widespread adoption in scenarios where all pain grades must be detected equally well.

The precision-recall curve in Figure 7 shows the performance of each class with an average mAP@0.5 value of 0.893. Several classes, such as pain_7 (0.970), pain_10 (0.966), and pain_9 (0.965), show excellent performance. However, there are classes with lower values, such as pain_3 (0.788) and pain_1 (0.797), which indicates that the model still has difficulty detecting these expressions.

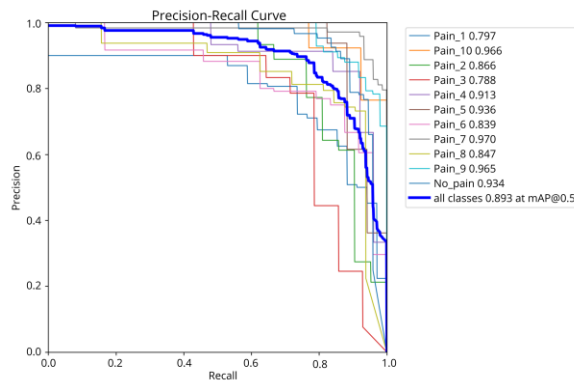


Figure 7. Precision-recall results

The recall versus confidence curve in Figure 8 shows that the average model recall is quite high, with a peak value of 0.97 at a confidence threshold close to 0. This indicates the model's ability to capture most positive instances, although there is a trade-off with decreased precision at low confidence.

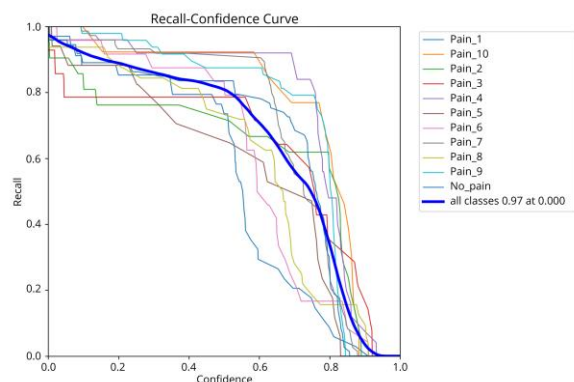


Figure 8. Recall vs confidence

On Tabel 5. High recall values at lower confidence thresholds indicate the model is highly sensitive and unlikely to miss actual instances of pain in pediatric patients. However, as the confidence threshold increases, recall steadily decreases, reflecting a higher standard for what counts as a positive detection. This trade-off is typical for deep learning classifiers: maximizing recall may introduce more false positives, while stricter thresholds reduce false alarms but risk missing true cases. Most pain classes (Pain_5, Pain_7, Pain_9, No_pain, etc.) maintain nearly perfect recall at threshold 0, supporting strong model sensitivity. Slightly lower recall is observed for Pain_2 and Pain_8, likely due to fewer training samples and class similarity issues. These patterns align with the overall objective in clinical triage—to capture all potential pain cases for

subsequent review or intervention. Maximizing recall is especially valuable in healthcare, as missing a genuine pain case can have significant consequences. The model's high recall ensures comprehensive screening, though final diagnosis should always incorporate precision analysis to balance clinical workflow and avoid unnecessary intervention.

Table 5. Maximum recall per class

Class	Maximum recall
Pain_1	0.98
Pain_2	0.95
Pain_3	0.98
Pain_4	0.97
Pain_5	0.99
Pain_6	0.96
Pain_7	1.00
Pain_8	0.95
Pain_9	0.99
Pain_10	0.98
No_pain	0.99
All Classes	0.97

The confusion matrix in Figure 9 provides a detailed overview of the classification of each class. The model successfully classifies classes pain_7 and pain_9 with an accuracy rate of over 90% (normalized). However, there are classification errors between similar classes, for example, pain_1 is frequently misclassified as not sick, and pain_8 is sometimes confused with other classes.

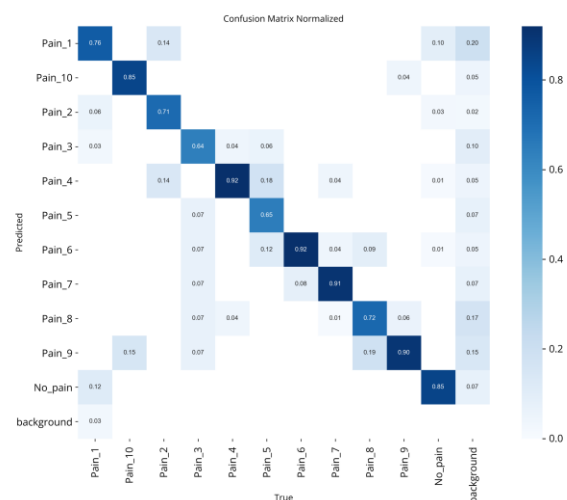


Figure 9. Confusion matrix

Calibration refers to Table 5 the alignment between a model's predicted confidence scores and the actual likelihood of correct classification. The normalized confusion matrix in Table 6 provides valuable insight into not only model accuracy, but also the reliability and consistency of those predictions across all pain classes.

High diagonal values (e.g., No_pain 0.95, Pain_7 0.94, Pain_4 0.92) indicate the model is well-calibrated for those classes: its high confidence in predictions is matched by actual correct classifications. This strengthens trust in the model's automated decisions,

especially for classes with robust representation and clear features. Lower diagonal values (Pain_3 0.64, Pain_8 0.73) reveal calibration gaps. For these classes, the model's confidence does not consistently translate into correct outcomes, possibly due to limited training samples, ambiguous facial features, or class imbalance. These classes require caution when interpreting predictions, and suggest targeted data augmentation or rebalancing to improve reliability.

Table 6. Calibration test

Prediction	Test	Maximum recall
Pain_1	Pain_1	0.98
Pain_10	Pain_10	0.95
Pain_2	Pain_2	0.98
Pain_3	Pain_3	0.97
Pain_4	Pain_4	0.99
Pain_5	Pain_5	0.96
Pain_6	Pain_6	1.00
Pain_7	Pain_7	0.95
Pain_8	Pain_8	0.99
Pain_9	Pain_9	0.98
No_pain	No_pain	0.99

The F1 versus confidence curve in Figure 10 shows that the best average F1 value is achieved at a threshold of around 0.41 with a score of 0.83. This indicates that the balance between precision and recall is quite optimal at that point.

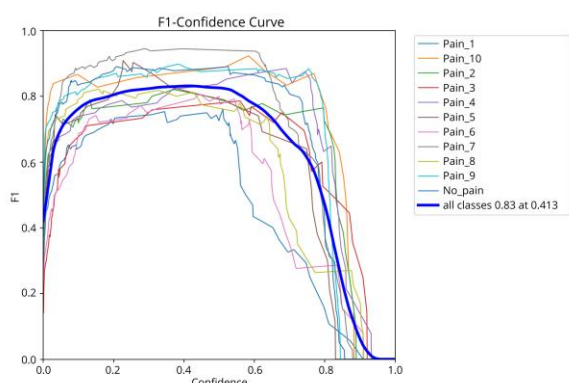


Figure 10. F1 score vs confidence

Calibration in machine learning refers to matching predicted probabilities (or model confidence) with actual observed outcome frequencies. On Table 7 the F1-Confidence Curve, the model's confidence threshold is calibrated to achieve the best possible balance between precision (correct positive predictions) and recall (completeness in capturing all positives). The intersection point (confidence ~0.41) where all classes together reach peak F1 (0.83) indicates the optimal calibration setting for harmonized sensitivity and specificity. This threshold is the point where the model's confidence scores are best aligned with real-world performance needs. While confusion matrices assess calibration per class (how well "high confidence" matches true class), the F1-Confidence Curve calibrates the global decision rule—ensuring that both types of model error (false positives and false negatives) are minimized for operational accuracy. Proper calibration

using the F1-Confidence Curve ensures that the chosen model threshold provides clinicians with reliable, usable results—avoiding excessive missed detections (under-confidence) or unnecessary alarms (over-confidence).

Table 7. F1score per class and threshold

Class	Maximum F1 Score	Optimal Confidence Threshold
Pain_1	~0.81	~0.40
Pain_2	~0.86	~0.40
Pain_3	~0.79	~0.40
Pain_4	~0.86	~0.41
Pain_5	~0.89	~0.41
Pain_6	~0.86	~0.40
Pain_7	~0.87	~0.41
Pain_8	~0.81	~0.41
Pain_9	~0.89	~0.41
Pain_10	~0.86	~0.41
No_pain	~0.85	~0.41
All Classes	0.83	0.41

The recent implementation of YOLO v11 as a tool for automated pain detection in young children introduces several advancements in clinical image analysis. The presented results reveal noteworthy strengths, especially in model sensitivity. The recall rate of 0.97 across all classes demonstrates that this approach is highly effective in identifying most pain cases, a critical factor in pediatric healthcare settings, where missing even mild pain signals can have significant consequences.

Precision values, which peak close to 1.00 for major classes at higher confidence thresholds, underscore the system's ability to minimize false positives when strict criteria are applied. This is particularly advantageous for clinical triage, as it supports swift and reliable assessment of pain levels with minimal disruption from erroneous alerts. The calibration performance, illustrated through both confusion matrix diagonals and the F1-Confidence Curve analysis, further validates the reliability of the model's predictions across a spectrum of pain expressions.

Comparatively, these metrics mark an improvement over traditional approaches such as standard CNN-based classifiers or older YOLO versions, which historically exhibit lower F1 scores and face challenges in multiclass discrimination. The average F1 score of 0.83 at an optimal threshold and a mean average precision (mAP@0.5) of 0.893 indicate a balanced and high-performing model well-suited for clinical deployment. Previous research has often reported greater difficulties in detecting less frequent pain expressions, a trend mirrored in this study. Classes with fewer data instances or ambiguous facial cues, like Pain_3 and Pain_1, continue to show lower precision and recall, highlighting persistent challenges related to data imbalance and subtle expression differentiation.

Despite these strengths, it is clear that further investigation is needed to overcome the limitations in

detecting minority classes. Future steps should focus on enriching the dataset with more varied examples, applying targeted data augmentation, and exploring additional modalities (such as vocal or behavioral cues) to complement facial analysis. Routine evaluation and transparent reporting of model calibration remain essential for ensuring consistent and trustworthy operation in clinical practice.

Overall, the pain expression detection model performed well with an average mAP@0.5 of 0.893 and a maximum recall of 0.97. However, data imbalance between classes remains a significant factor affecting performance, especially for classes with a small number of instances. Classes with similar expressions, such as pain_1 and no pain, tend to cause misclassification. This can be addressed by increasing data variation through richer augmentation techniques or balancing the dataset by increasing data on minority classes.

From a nursing perspective, an automated detection system like this has the potential to be a significant tool in clinical decision support [10], especially in pediatric care. Real-time application of technology can support objective pain documentation [18], speeding up pain triage [31], as well as reducing the burden of manual observation by nurses [6]. However, the imbalanced data distribution across expression classes remains a challenge, so further development efforts need to focus on improving dataset quality and quantity, including data augmentation and weighting strategies for minority populations. If these challenges can be overcome, the YOLO v11 model has great potential for integration into digital pediatric nursing care systems.

4. Conclusions

The conclusion of this study indicates that the application of YOLO v11 has great potential as a method for detecting pain in young children based on facial expressions. The model performance evaluation yielded a maximum recall value of 0.97, indicating a high level of sensitivity in recognizing pain expressions, making it highly relevant in a clinical context to ensure no pain symptoms are missed. However, the lower precision value compared to the recall indicates a prediction error in the form of pain detection in non-pain expressions. Factors such as lighting variations, shooting angles, and facial expression similarities between classes are suspected to contribute to this misclassification, especially in classes pain_1 and pain_3 which are often confused with the non-pain category. The results of the performance curve and confusion matrix evaluation support this finding, where classes with more data such as pain_7, pain_9, and pain_10 can be recognized well, while classes with limited data show lower accuracy. Overall, the mAP@0.5 value of 0.893, accuracy of 78%, precision of 89.3%, recall of 97%, and F1-score of 83% indicate that the model is capable of performing quite reliably and can be applied in real-time as an aid in pediatric pain assessment. However, the imbalance in data

distribution between classes remains a major challenge that affects the consistency of the results.

In addition, the findings of this study strengthen the argument presented in the introduction regarding the urgent need for child-specific, objective, and scalable methods to detect acute pain in clinical settings. YOLO v11 demonstrated good potential to fill this gap, particularly through its ability to analyze subtle facial changes associated with acute pain in young children with real-time inference capability.

Despite this progress, the results also indicate that further refinement is required to ensure consistent performance across all acute pain levels, especially in lower or ambiguous categories where facial expressions tend to overlap. Addressing data imbalance and enhancing feature robustness are essential steps for future development.

Therefore, further research needs to focus on strategies to improve dataset quality, either through data balancing, adding data to minority classes, or implementing more varied augmentation techniques, so that model performance can be more stable and evenly distributed across all pain categories. Future enhancements to the model—such as integrating temporal video-based analysis, multimodal inputs, or optimizing architecture for pediatric facial characteristics—may further increase accuracy and reliability, supporting the long-term goal of developing a clinically deployable tool for accurate acute pain assessment in young children.

Ethical approval

Research ethics approval was obtained from the Research Ethics Committee of the Faculty of Nursing, Universitas Indonesia with approval number KET-029/UN2.F12.D1.2.1/PPM.00.02/2025.

Acknowledgements

The author would like to thank the Indonesian Education Scholarship (BPI), the Center for Funding and Assessment of Higher Education, Ministry of Higher Education, Science and Technology of the Republic of Indonesia (Kemdiktisaintek RI), Indonesia Endowment Fund for Education Agency (LPDP RI), Universitas Indonesia, and Universitas Padjadjaran for their support and facilities in providing database access and research funding.

Thanks, are also extended to the Head of the Garut Regency Health Office and the Director of UOBK RSU dr. Slamet Garut and their staff who have provided support, permission, and facilities for collecting research data.

References

- [1] M. J. Hockenberry and D. Wilson, *Wong's essentials of pediatric nursing*. St. Louis: Mosby, 2016.
- [2] S. N. Raja et al., "The revised International Association for the Study of Pain definition of pain: concepts, challenges, and

- compromises,” *Pain*, vol. 161, no. 9, pp. 1976–1982, Sep. 2020, doi: [10.1097/j.pain.0000000000001939](https://doi.org/10.1097/j.pain.0000000000001939).
- [3] R. C. B. Manworren and J. Stinson, “Pediatric pain measurement, assessment, and evaluation,” *Semin Pediatr Neurol*, vol. 23, no. 3, pp. 189–200, Aug. 2016, doi: [10.1016/j.spen.2016.10.001](https://doi.org/10.1016/j.spen.2016.10.001).
 - [4] G. Alaleh et al., “Validation of Persian Version of PedsQL™ 4.0™ Generic Core Scales in Toddlers and Children,” *Int J Prev Med.*, vol. 3, no. 5, pp. 341–350, 2012, [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/articles/PMC3374492/>
 - [5] S. J. Friedrichsdorf, J. Giordano, K. D. Dakoji, A. Warmuth, C. Daughtry, and C. A. Schulz, “Chronic pain in children and adolescents: Diagnosis and treatment of primary pain disorders in head, abdomen, muscles and joints,” *Children*, vol. 3, no. 4, Dec. 2016, doi: [10.3390/children3040042](https://doi.org/10.3390/children3040042).
 - [6] G. G. Ramdhanie, D. Wanda, N. Agustini, and T. Abuzairi, “Identification of technology-based models and efficacy of digital-based pain facial expression assessment tools among children: a systematic review,” *BMC Nurs*, vol. 24, no. 1, Dec. 2025, doi: [10.1186/s12912-025-03451-9](https://doi.org/10.1186/s12912-025-03451-9).
 - [7] K. S. Higgins et al., “Availability of researcher-led eHealth tools for pain assessment and management: Barriers, facilitators, costs, and design,” *Pain Rep*, vol. 3, no. 7, Sep. 2018, doi: [10.1097/PR9.0000000000000686](https://doi.org/10.1097/PR9.0000000000000686).
 - [8] T. Sakulchit, B. Kuzeljevic, and R. D. Goldman, “Evaluation of Digital Face Recognition Technology for Pain Assessment in Young Children,” *Clinical Journal of Pain*, vol. 35, no. 1, pp. 18–22, Jan. 2019, doi: [10.1097/AJP.0000000000000659](https://doi.org/10.1097/AJP.0000000000000659).
 - [9] N. Meesters, T. Dilles, S. Simons, and M. van Dijk, “Do Pain Measurement Instruments Detect the Effect of Pain-Reducing Interventions in Neonates? A Systematic Review on Responsiveness,” *Journal of Pain*, vol. 20, no. 7, pp. 760–770, Jul. 2019, doi: [10.1016/j.jpain.2018.12.005](https://doi.org/10.1016/j.jpain.2018.12.005).
 - [10] K. Sikka et al., “Automated assessment of children’s postoperative pain using computer vision,” *Pediatrics*, vol. 136, no. 1, pp. e124–e131, Jul. 2015, doi: [10.1542/peds.2015-0029](https://doi.org/10.1542/peds.2015-0029).
 - [11] G. D. De Sario et al., “Using ai to detect pain through facial expressions: a review,” *Bioengineering*, vol. 10, no. 5, May 2023, doi: [10.3390/bioengineering10050548](https://doi.org/10.3390/bioengineering10050548).
 - [12] C. Blais, D. Fiset, H. Furumoto-Deshaies, M. Kunz, D. Seuss, and S. Cormier, “Facial Features Underlying the Decoding of Pain Expressions,” *Journal of Pain*, vol. 20, no. 6, pp. 728–738, Jun. 2019, doi: [10.1016/j.jpain.2019.01.002](https://doi.org/10.1016/j.jpain.2019.01.002).
 - [13] Y. Sun et al., “Detecting discomfort in infants through facial expressions,” *Physiol Meas*, vol. 40, no. 11, Dec. 2019, doi: [10.1088/1361-6579/ab55b3](https://doi.org/10.1088/1361-6579/ab55b3).
 - [14] U. Nath, “Facial Expressions as Behavioral Indicators for Assessing Pain using Machine Learning Models.” [Online]. Available: www.JSR.org/hs
 - [15] M. Zhang et al., “Using artificial intelligence to improve pain assessment and pain management: a scoping review,” *Journal of the American Medical Informatics Association*, vol. 30, no. 3, pp. 570–587, Mar. 2023, doi: [10.1093/jamia/ocac231](https://doi.org/10.1093/jamia/ocac231).
 - [16] U. Nath, “Facial Expressions as Behavioral Indicators for Assessing Pain using Machine Learning Models,” *Journal of Student Research*, vol. 12, no. 4, pp. 1–14, 2023, [Online]. Available: www.JSR.org/hs
 - [17] J.-M. Yue, Q. Wang, B. Liu, and L. Zhou, “Postoperative accurate pain assessment of children and artificial intelligence: a medical hypothesis and planned study,” *World J Clin Cases*, vol. 12, no. 4, pp. 681–687, Feb. 2024, doi: [10.12998/wjcc.v12.i4.681](https://doi.org/10.12998/wjcc.v12.i4.681).
 - [18] B. Susam et al., “Automated Pain Assessment in Children Using Electrodermal Activity and Video Data Fusion via Machine Learning,” *IEEE Trans Biomed Eng*, vol. 69, no. 1, pp. 422–431, Jan. 2022, doi: [10.1109/TBME.2021.3096137](https://doi.org/10.1109/TBME.2021.3096137).
 - [19] E. Jung and D. Nam, “Lightweight YOLO-based real-time fall detection using feature map-level knowledge distillation,” *ICT Express*, Aug. 2025, doi: [10.1016/j.icte.2025.08.009](https://doi.org/10.1016/j.icte.2025.08.009).
 - [20] M. Arfan, A. Nurjalal, M. Somantri, and Sudjadi, “Recognition to Human Activities in Shrimp Pond Areas with a Convolutional Neural Network,” *Jurnal RESTI*, vol. 5, no. 1, pp. 174–179, Feb. 2021, doi: [10.29207/resti.v5i1.2888](https://doi.org/10.29207/resti.v5i1.2888).
 - [21] I. L. Kharisma, I. Nurmalasari, Y. Lestari, S. Dela Septiani, Kamdan, and M. A. S. Yudono, “AI-Powered Sign Language Detection Using YOLO-v11 for Communication Equality,” *The 7th International Global Conference Series on ICT Integration in Technical Education & Smart Society*, p. 83, Sep. 2025, doi: [10.3390/engproc2025107083](https://doi.org/10.3390/engproc2025107083).
 - [22] E. Syahrudin, E. Utami, and A. D. Hartanto, “Enhanced Yolov8 with OpenCV for Blind-Friendly Object Detection and Distance Estimation,” *Jurnal RESTI*, vol. 8, no. 2, pp. 199–207, Apr. 2024, doi: [10.29207/resti.v8i2.5529](https://doi.org/10.29207/resti.v8i2.5529).
 - [23] A. F. Rasheed and M. Zarkoosh, “YOLOv11 Optimization for Efficient Resource Utilization,” Jul. 2025, [Online]. Available: <http://arxiv.org/abs/2412.14790>
 - [24] T. Abuzairi, Nurdina Widanti, Arie Kusumaningrum, and Yeni Rustina, “Implementasi Convolutional Neural Network Untuk Deteksi Nyeri Bayi Melalui Citra Wajah Dengan YOLO,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 624–630, Aug. 2021, doi: [10.29207/resti.v5i4.3184](https://doi.org/10.29207/resti.v5i4.3184).
 - [25] R. D. Ramadhani, A. N. A. Thohari, C. Kartiko, A. Junaidi, T. G. Laksana, and N. A. S. Nugraha, “Optimasi Akurasi Metode Convolutional Neural Network untuk Identifikasi Jenis Sampah,” *Jurnal RESTI*, vol. 5, no. 2, pp. 312–318, Apr. 2021, doi: [10.29207/resti.v5i2.2754](https://doi.org/10.29207/resti.v5i2.2754).
 - [26] X. Xin, X. Lin, S. Yang, and X. Zheng, “Pain intensity estimation based on a spatial transformation and attention CNN,” *PLoS One*, vol. 15, no. 8 August 2020, Aug. 2020, doi: [10.1371/journal.pone.0232412](https://doi.org/10.1371/journal.pone.0232412).
 - [27] M. Fachrie, “A Simple Vehicle Counting System Using Deep Learning with YOLOv3 Model,” *RESTI journal (System Engineering and Information Technology)*, vol. 4, no. 2, pp. 462–468, 2020, doi: [10.29207/resti.v4i3.1871](https://doi.org/10.29207/resti.v4i3.1871)
 - [28] R. Khanam and M. Hussain, “YOLOv11: An Overview of the Key Architectural Enhancements,” Oct. 2024, [Online]. Available: <http://arxiv.org/abs/2410.17725>
 - [29] S. E. Sorour, M. Aljaafari, A. A. Alarfaj, W. H. A. AlMusallam, and K. S. Aljoqiman, “Fine-tuned Vision Transformers and YOLOv11 for precise detection of pediatric Adenoid Hypertrophy,” *Alexandria Engineering Journal*, vol. 128, pp. 366–393, Sep. 2025, doi: [10.1016/j.aej.2025.05.038](https://doi.org/10.1016/j.aej.2025.05.038).
 - [30] Y. Umar, Hanafi, S. Mardi, Nugroho, Susiki, and R. F. Rachmadi, “Deteksi Penggunaan Helm Pada Pengendara,” 2020. <https://www.its.ac.id/komputer/wp-content/uploads/sites/28/2020/09/Buku-Yusuf-Umar-Hanafi-07211640000006-Supeno-M-Fuad-R.pdf>
 - [31] Á. Sabater-Gárriz, F. X. Gaya-Morey, J. M. Buades-Rubio, C. Manresa-Yee, P. Montoya, and I. Riquelme, “Automated facial recognition system using deep learning for pain assessment in adults with cerebral palsy,” *Digit Health*, vol. 10, Jan. 2024, doi: [10.1177/20552076241259664](https://doi.org/10.1177/20552076241259664).