Jurnal Teknologi Informasi dan Pendidikan
Volume 19, No. 2, September 2026
https://doi.org/10.24036/jtip.v19i2.1120

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

Lightweight YOLO Models for Robust Facial Expression Detection

Achmad Indra Aulia[1*]✉, Albert Jofrandi Hutapea[1], Amril Mutoi Siregar[1], Surjandy[2]
[1]informatics study program, Institut Teknologi Sains Bandung, Cikarang, Indonesia
[2]Digital Business study program, Institut Teknologi Sains Bandung, Cikarang, Indonesia
**Corresponding Author: achmadindra@itsb.ac.id*

Article Information

Article history:

No. 1120
Rec. March 08, 2026
Rev. April 02, 2026
Acc. April 15, 2026
Pub. April 16, 2026
Page. 1449 – 1463

Keywords:

- *Facial Expression Detection*
- *YOLO*
- *Data Augmentation*
- *Confidence Threshold*
- *Object Detection*

ABSTRACT

Facial expression detection is a fundamental component of artificial intelligence systems, particularly in human–machine interaction. However, achieving robust detection accuracy remains challenging due to variations in lighting, facial orientation, and limited training data diversity. While recent lightweight YOLO architectures—YOLOv8n, YOLOv10n, and YOLO11n—have demonstrated strong performance in general object detection, comparative studies evaluating these models specifically for facial expression detection remain limited. This study addresses this gap by systematically comparing these three nano-variant models on a dataset of 2,000 labeled facial images across four expression categories: flat face, angry, sad, and smile. The dataset was divided into training (70%), validation (20%), and test (10%) subsets. Experiments were conducted under two scenarios—with and without data augmentation—using identical training configurations on an NVIDIA GeForce RTX 4070 GPU. Augmentation techniques included mosaic composition, HSV variation, geometric transformations, and flipping. Results show that augmentation improved the F1 score of YOLOv10n from 0.68 to 0.72 and YOLO11n from 0.65 to 0.72, with the latter achieving the highest overall precision of 0.82. YOLOv8n exhibited stable performance with an F1 score of 0.75 under both conditions. Confidence threshold optimization revealed distinct optimal operating points for each model, ranging from 0.1 to 0.6, confirming that per-model threshold tuning is necessary to maximize detection performance. These findings provide practical guidance for selecting and configuring lightweight YOLO models for facial expression detection in resource-constrained environments.

How to Cite:

Aulia, A. I., & et al. (2026). Lightweight YOLO Models for Robust Facial Expression Detection. Jurnal Teknologi Informasi Dan Pendidikan, 19(2), 1449-1463. https://doi.org/10.24036/jtip.v19i2.1120

Jurnal Teknologi Informasi dan Pendidikan
Volume 19, No. 2, September 2026
https://doi.org/10.24036/jtip.v19i2.1120

1. INTRODUCTION

Facial expression detection has become a key capability in a growing number of intelligent systems, including emotion-aware interfaces, assistive communication tools, mental health monitoring applications, and human-robot interaction platforms [1], [2]. Despite substantial progress driven by deep learning, several persistent challenges continue to limit detection accuracy in practice, including variable illumination conditions, partial facial occlusions, inter-individual morphological differences, and the scarcity of diverse, well-annotated training data [3].

In the domain of real-time object detection, the YOLO (You Only Look Once) family of architecture has established itself as a leading paradigm, offering single-pass detection that balances inference speed with accuracy [4], [5], [6]. The continued evolution of YOLO models has produced increasingly efficient variants, with each generation introducing architectural innovations aimed at improving feature extraction, reducing computational overhead, or enhancing detection precision. The nano (n) variants of each architecture are specifically designed for deployment on resource-constrained hardware, making them particularly relevant for edge computing and real-time applications [7].

Among the most recent lightweight YOLO models, YOLOv8n, YOLOv10n, and YOLO11n each introduce distinct architectural contributions. YOLOv8n employs a C2f (Cross-Stage Partial with 2F connections) backbone with a decoupled anchor-free detection head, offering a balanced speed–accuracy trade-off [5]. YOLOv10n introduces a dual-assignment training strategy that eliminates the need for Non-Maximum Suppression (NMS) during inference, thereby reducing post-processing latency [8]. YOLO11n, the most recent of the three, incorporates C3k2 blocks, a Spatial Pyramid Pooling-Fast (SPPF) module, and a Convolutional block with Parallel Spatial Attention (C2PSA) for enhanced spatial feature representation [9].

Despite their demonstrated effectiveness in general object detection benchmarks [10], comprehensive comparisons of these three architectures specifically within the domain of facial expression detection are lacking. Most prior comparative studies either focus on general detection tasks or evaluate older YOLO versions [5], [8]. This leaves an open question regarding which lightweight YOLO variant is most suitable for the specific characteristics of facial expression detection, where subtle intra-class differences and variable image quality present unique challenges [11].

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

This study aims to address this gap through a systematic comparative evaluation of YOLOv8n, YOLOv10n, and YOLO11n for detecting four basic facial expressions: flat face, angry, sad, and smile. Beyond the model comparison itself, this research investigates two additional factors that influence detection performance but have received limited attention in this specific context:

1) **Data Augmentation Impact**: The study evaluates whether applying augmentation techniques to a relatively small training set can meaningfully improve detection accuracy for each model, providing insights into data preparation strategies for scenarios where large-scale datasets are unavailable [12], [13], [14].
2) **Per-Model Confidence Threshold Optimization**: The study conducts a fine-grained sweep of confidence thresholds to identify the operating point that best balances precision and recall for each model and augmentation scenario. While threshold selection is standard practice, its model-specific behavior across these three architectures has not been previously documented for facial expression detection.

The practical contribution of this work lies in providing empirical guidance for developers and researchers seeking to deploy lightweight YOLO models for facial expression detection, particularly in settings where computational resources are limited and labeled data is scarce.

2. RESEARCH METHOD

This section describes the experimental methodology employed to compare YOLOv8n, YOLOv10n, and YOLO11n for facial expression detection. The overall workflow encompasses dataset preparation, preprocessing, augmentation, model training, evaluation, and confidence threshold optimization. The described process was performed on computer with an NVIDIA GeForce RTX 4070 GPU.

2.1. Dataset Preparation

The facial expression dataset used in this study was sourced from Kaggle with the dataset name of "Ekspresi Wajah" [15]. This dataset was selected for its relevance to the research objective, its manageable size suitable for lightweight model experimentation, and its straightforward annotation structure.

The dataset comprises 2,000 images with varying resolutions (96×96, 800×400, and 600×600 pixels) spanning four facial expression categories: flat face, angry, sad, and smile, with approximately 500 images per class ensuring a balanced distribution.. All images were annotated with bounding boxes delineating the facial region and corresponding expression class labels. Annotations were converted to the standard YOLO format (.txt) following the convention: `class_id center_x center_y width height`.

The dataset was divided into three subsets: a training set (70%, 1,400 images), a validation set (20%, 400 images), and a test set (10%, 200 images). The training set was used for model optimization, the validation set for monitoring training progress and guiding hyperparameter decisions, and the test set for final objective evaluation.

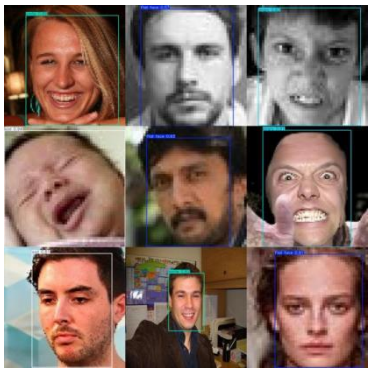


Figure 1. Example of a Sample Dataset with Facial Expression Annotations

Figure 1 presents representative samples from the dataset. The images capture individuals of varying ages under different conditions of lighting, image quality, and facial orientation. This variability, while limited in scale compared to large-scale benchmarks such as AffectNet [16], provides a controlled setting for evaluating the comparative performance of lightweight models under data-scarce conditions—a scenario commonly encountered in domain-specific or custom deployment settings.

2.2. Data Preprocessing

All images were rescaled to a standardized resolution of 640×640 pixels, which serves as the default input size for the YOLO model family. Following the default Ultralytics pipeline, the original aspect ratio was preserved via letterboxing (scaling to fit within 640×640 while padding the remaining area with a constant gray value of 114), using bilinear interpolation. Pixel values were normalized to maintain numerical stability during training and accelerate convergence.

2.3. Data Augmentation

To investigate the impact of data augmentation on model performance, two experimental scenarios were defined:

1) **Scenario 1 (Without Augmentation)**: Models were trained exclusively on the original preprocessed images without any augmentation applied.
2) **Scenario 2 (With Augmentation)**: A combination of augmentation techniques was applied to the training set during training to increase data diversity and improve model robustness. Table 1 lists the augmentation parameters and their values.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

Table 1. Data Augmentation Parameters

Parameter	Value	Description
mosaic	True	Combines four images into one composite
hsv_h	0.015	Hue variation
hsv_s	0.7	Saturation variation
hsv_v	0.4	Brightness variation
flipud	0.5	Vertical flip probability
fliplr	0.5	Horizontal flip probability
scale	0.5	Random scaling factor
translate	0.1	Random translation (up to 10%)
degrees	10	Random rotation (±10°)
shear	2	Random shear distortion (±2°)

These techniques were selected to simulate common real-world variations in facial appearance, including changes in orientation, color conditions, and spatial positioning, thereby encouraging the models to develop more robust and generalizable feature representations.

2.4. Model Selection and Configuration

Three nano-variant YOLO models were selected for comparison, chosen specifically for their lightweight architecture and suitability for resource-constrained deployment:

1) YOLOv8n utilizes a C2f (Cross Stage Partial with two convolutions) backbone and a decoupled anchor-free detection head. It represents a mature and well-optimized architecture with a proven speed–accuracy balance.
2) YOLOv10n introduces a dual-assignment training strategy that eliminates the need for Non-Maximum Suppression (NMS) at inference time. It also features lightweight classification heads and spatial channel decoupled down sampling, targeting reduced inference latency.
3) YOLO11n is the most recent iteration, incorporating C3k2 blocks for efficient feature extraction, a Spatial Pyramid Pooling-Fast (SPPF) module, and a C2PSA (Convolutional block with Parallel Spatial Attention) module for enhanced spatial attention.

To ensure a fair and controlled comparison, identical training hyperparameters were applied across all models and augmentation scenarios, as detailed in Table 2. Early stopping with a patience of 30 epochs was enabled to halt training if validation performance did not improve, thereby mitigating overfitting and unnecessary computation.

Table 2. Training Hyperparameters

Parameter	Value
Maximum Epochs	100
Batch Size	64
Input Image Size	640 × 640 pixels
Optimizer	Adam
Initial Learning Rate	0.0005
Early Stopping Patience	30

2.5. Model Training

Each of the three models was trained independently under both augmentation scenarios, using the hyperparameter configuration described in Section 3.4. During training, performance was monitored on the validation set at each epoch.

2.6. Evaluation Metrics

Model performance was evaluated using the following standard metrics, where TP denotes true positives, FP denotes false positives, and FN denotes false negatives:

1) Precision measures the proportion of positive detections that are correct.

$$P = \frac{TP}{FP+TP} \quad (1)$$

2) Recall measures the proportion of actual positive instances that are correctly detected.

$$R = \frac{TP}{FN+TP} \quad (2)$$

3) F1 Score is the harmonic mean of precision and recall, providing a balanced measure.

$$F1 = 2 \times \frac{P \times R}{P+R} \quad (3)$$

The F1 score was selected as the primary comparison metric because it captures the trade-off between precision and recall in a single value, which is particularly important for facial expression detection where both missed detections and false alarms are undesirable.

2.7. Confidence Threshold Optimization

A systematic confidence threshold analysis was conducted to identify the optimal operating point for each model and augmentation scenario [17]. The procedure was as follows:

1) The best-performing model weights from each training run were loaded.
2) Confidence threshold values were swept from 0.0 to 1.0 in increments of 0.01.
3) For each threshold, the model performed inference on all test set images, with detections filtered by the current threshold.
4) Each detection was classified as a true positive (TP) if it matched the correct expression class and achieved an Intersection over Union (IoU) ≥ 0.5 with the corresponding ground-truth bounding box. Unmatched detections were classified as false positives (FP), and undetected ground-truth boxes as false negatives (FN).

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

5) Precision, recall, and F1 score were computed at each threshold using the accumulated TP, FP, and FN counts across all test images.
6) The threshold yielding the maximum F1 score was identified as the optimal operating point.

This approach provides a detailed view of each model's precision–recall trade-off and enables identification of the confidence level at which each model achieves its best overall balance.

3. RESULTS AND DISCUSSION

This section presents experimental results and discussions, covering overall model performance, the effect of augmentation, confidence threshold behavior, and classification analysis through confusion matrices.

3.1. Overall Model Performance

Table 3 summarizes the detection performance of each model under both augmentation scenarios, along with the corresponding optimal confidence threshold.

Table 3. Model Performance

Model	Augmentation	Precision	Recall	F1 Score (Optimal Threshold)
YOLOv8n	No	0.67	0.85	0.75, (@0.1)
YOLOv8n	Yes	0.67	0.85	0.75, (@0.3)
YOLOv10n	No	0.60	0.79	0.68, (@0.3)
YOLOv10n	Yes	0.69	0.75	0.72, (@0.4)
YOLO11n	No	0.75	0.56	0.65, (@0.6)
YOLO11n	Yes	0.82	0.70	0.72, (@0.5)

Several patterns emerge from these results. First, augmentation improved precision for all models, with YOLO11n showing the most pronounced gain (from 0.75 to 0.82). However, the effect on recall was mixed: YOLOv8n maintained identical recall under both conditions, while YOLOv10n and YOLO11n exhibited slight recall decreases after augmentation. In terms of the F1 score, YOLOv8n was stable at 0.75 regardless of augmentation, while YOLOv10n improved from 0.68 to 0.72 and YOLO11n from 0.65 to 0.72. Notably, the optimal confidence threshold differed substantially across models, ranging from 0.10 (YOLOv8n without augmentation) to 0.60 (YOLO11n without augmentation), indicating that a universal default threshold would be suboptimal.

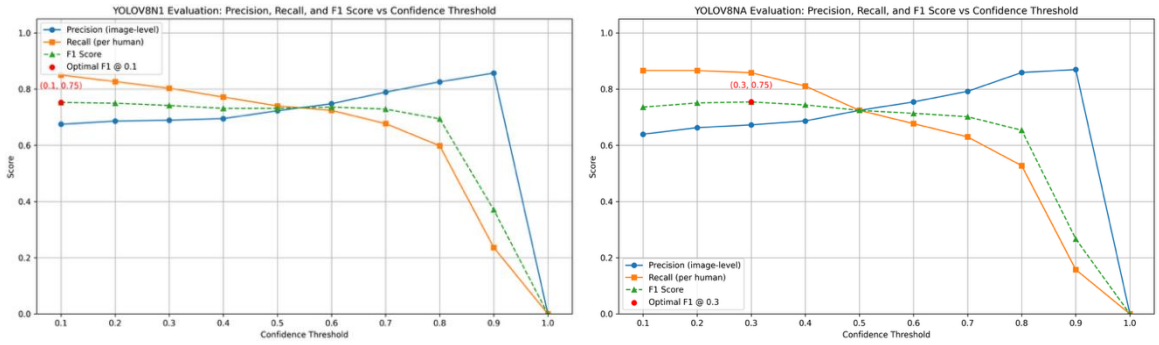
Jurnal Teknologi Informasi dan Pendidikan
Volume 19, No. 2, September 2026
https://doi.org/10.24036/jtip.v19i2.1120

3.2. Confidence Threshold Analysis

In this subsection, confidence threshold analysis is performed on each model: YOLOv8n, YOLOv10n, and YOLOv11n. Figure 4 illustrates the best confidence threshold determination from the curves of models' performance across confidence thresholds.

3.2.1. YOLOv8n

The metric curves of YOLOv8n across confidence thresholds are presented in Figure 2. Without augmentation, the optimal F1 score of 0.75 was achieved at a threshold of 0.10, indicating that the model maintains high recall even at very low confidence levels but with correspondingly lower precision. With augmentation, the identical F1 score of 0.75 was reached at a higher threshold of 0.30, where precision was better balanced against recall. This shift suggests that augmentation improved the model's confidence calibration: the model became more discriminating in its predictions, requiring a higher confidence level to achieve the same performance, which indicates more reliable detections overall.



(a) (b)

Figure 2. YOLOv8n precision, recall, and F1 score vs. confidence threshold: (a) without augmentation, (b) with augmentation

3.2.2. YOLOv10n

Without augmentation, the F1 score of YOLOv10n model peaked at 0.68 with a threshold of 0.30, and the precision–recall curves exhibited a steep trade-off. After augmentation, the peak F1 score improved to 0.72 at a threshold of 0.40, with a notably less steep precision–recall trade-off. The improved precision without a proportional loss in recall indicates that augmentation helped the model learn more robust feature representations, leading to higher-quality detections.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

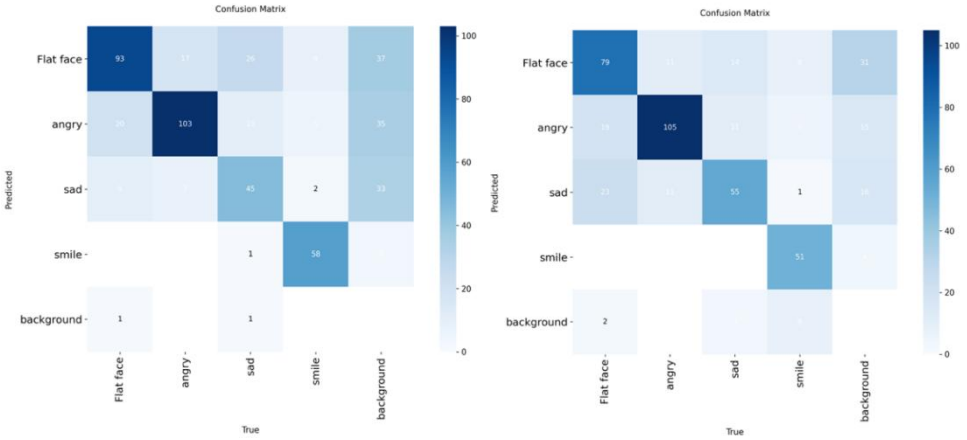
3.2.3. YOLOv11n

Without augmentation, the optimal F1 score of only 0.65 was achieved at a high threshold of 0.60 for YOLOv11n, reflecting a model with high precision but substantially limited recall—indicating that the model frequently failed to detect expressions despite being accurate when it did. With augmentation, the optimal F1 score rose to 0.76 at a threshold of 0.50, with both precision and recall improving and the F1 curve exhibiting greater stability across the threshold range. This represents the largest augmentation-driven improvement among the three models.

3.3. Confusion Matrix Analysis

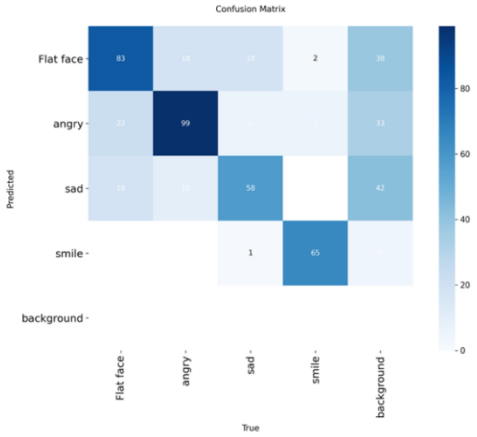
3.3.1. Model With Augmentation

Figure 3 presents the normalized confusion matrices for all three models trained with augmentation. YOLO11n achieved the strongest per-class accuracies, with 76% for flat face, 91% for angry, and 56% for sad, alongside the lowest rate of misclassification to the background class. YOLOv10n showed competitive results for the angry and smile classes but exhibited more frequent confusion for the sad class. YOLOv8n performed adequately overall but displayed a higher tendency to misclassify flat face and sad expressions as background. All three models achieved consistent accuracy of approximately 70% for the smile class. The sad class proved most challenging across all models, likely due to its subtle visual characteristics and potential overlap with the flat face expression.



(a) (b)

Jurnal Teknologi Informasi dan Pendidikan
Volume 19, No. 2, September 2026
https://doi.org/10.24036/jtip.v19i2.1120

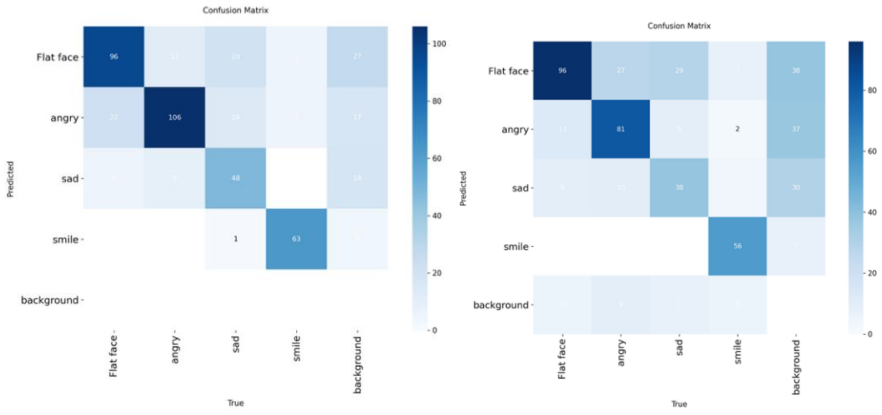


(c)

Figure 3. Normalized confusion matrices (with augmentation): (a) YOLOv8n, (b) YOLOv10n, (c) YOLO11n

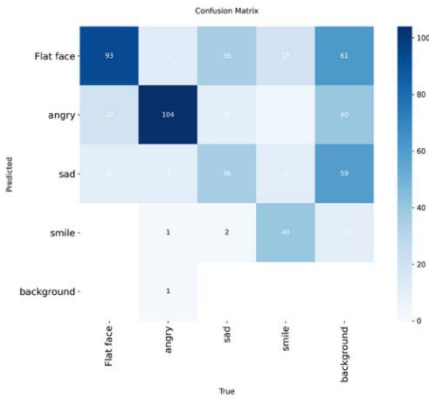
3.3.2. Model Without Augmentation

The confusion matrices for models trained without augmentation are depicted in Figure 4. YOLOv8n correctly classified flat face and smile at rates of 0.75 and 0.89 respectively, though inter-class confusion between sad and angry was observed. YOLOv10n achieved the highest accuracy for the angry class at 0.84 but showed notable misclassification for sad and background. YOLO11n exhibited the most evenly distributed confusion but with the lowest overall per-class accuracies, peaking at only 0.60 for angry. These results highlight that without augmentation; all three models struggled more substantially with expressions that share subtle visual features. Comparing the confusion matrices across augmentation conditions, augmentation consistently reduced background misclassification and improved discrimination between visually similar expression classes, with the most pronounced improvements observed for YOLO11n.



(a) (b)

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id



(c)

Figure 4. Normalized confusion matrices (without augmentation): (a) YOLOv8n, (b) YOLOv10n, (c) YOLO11n

3.4. Comparative Summary

The consolidated comparison is presented in Figure 5. YOLOv8n was the most consistent model, maintaining identical performance regardless of augmentation, but with higher confidence threshold after augmentation. YOLOv10n showed clear improvement with augmentation, particularly in precision. YOLO11n demonstrated the largest performance gap between augmentation conditions and achieved the highest precision overall (0.82) when augmentation was applied, though its recall remained lower than that of YOLOv8n.

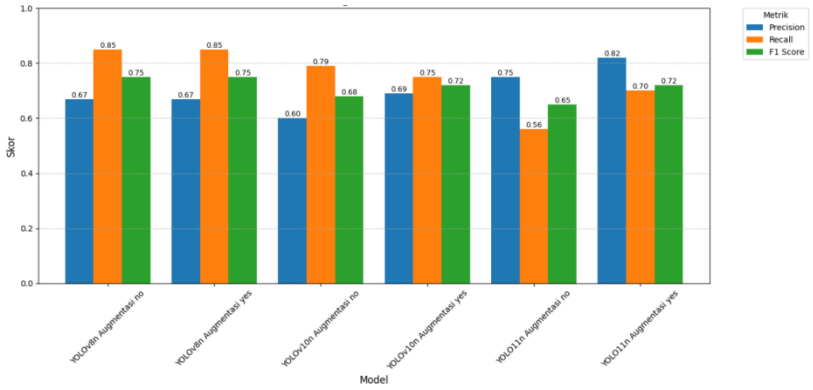


Figure 5. Comparative bar chart of precision, recall, and F1 score across all models and augmentation scenarios

3.5. Discussion

3.5.1. Impact of Data Augmentation

The results demonstrate that data augmentation had a meaningful but model-dependent impact on detection performance, consistent with the broader augmentation literature [12], [14] which established that augmentation generally improves robustness under data-scarce conditions. YOLOv8n exhibited stability across both conditions, achieving an F1 score of 0.75 with and without augmentation. However, the confidence threshold after augmentation is higher which shows that the model became more discriminating against in its predictions, requiring a higher confidence level to achieve the same performance. This indicates more reliable detections overall. The findings suggest that its C2f architecture and anchor-free detection head are sufficiently robust to generalize from the original training data alone.

In contrast, YOLOv10n and YOLO11n both benefited substantially from augmentation, with F1 score improvements of +0.04 and +0.07, respectively. YOLO11n showed the most dramatic precision increase (from 0.75 to 0.82), suggesting that its more complex architecture—incorporating attention mechanisms (C2PSA) and enhanced pooling (SPPF)—has greater capacity to leverage the increased training diversity provided by augmentation, extending the augmentation findings in [13]. However, this comes with a trade-off: without augmentation, YOLO11n exhibited the lowest recall among all models (0.56), indicating that its more complex feature extraction pipeline may require more diverse training data to avoid overly conservative detection behavior.

These findings carry practical implications for industrial deployment. For edge computing in healthcare or assistive technology—such as emotion-aware patient monitoring or communication aids for individuals with autism spectrum disorder—YOLOv8n's data-independent stability makes it a practical default choice, reducing the data engineering burden in settings where collecting expression datasets is costly and ethically constrained. When augmentation can be applied, YOLO11n's highest precision (0.82) makes it well-suited for smart retail emotion monitoring or customer experience analytics, where false positives lead to erroneous sentiment assessments. For real-time human-robot interaction or driver monitoring systems, the per-model threshold optimization framework presented here enables engineers to calibrate the precision-recall trade-off to their specific safety requirements.

3.5.2. Confidence Threshold Behavior

The threshold optimization analysis revealed that each model operates optimally at a distinct confidence level. YOLOv8n achieved its best F1 score at low thresholds (0.10–0.30), indicating that this model assigns moderate confidence scores to correct detections and thus benefits from a permissive threshold. YOLOv10n and YOLO11n required higher thresholds

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

(0.40–0.60), suggesting that these models produce a wider distribution of confidence scores, necessitating more selective filtering to separate true detections from false positives.

A particularly notable finding is that augmentation shifted the optimal threshold for YOLOv8n from 0.10 to 0.30 without changing the peak F1 score. This indicates that augmentation improved the model's confidence calibration—detections became more confidently separated from non-detections—even though the overall accuracy remained constant. This is a practically valuable outcome, as a higher optimal threshold generally implies more reliable individual detections. From a theoretical perspective, the wide threshold divergence across models (0.10 to 0.60) provides empirical evidence that confidence calibration is an architecture-dependent property rather than a universal characteristic, extending the framework proposed in [17] and demonstrating that threshold optimization must be treated as a model-specific design parameter when transitioning between architectures for the same task.

3.5.3. Per-Class Performance

The confusion matrix analysis revealed that the sad expression class was consistently the most challenging across all models and conditions. This is likely attributable to the subtle visual features that distinguish sad from flat face and, to a lesser extent, from angry expressions. The angry class was generally the most reliably detected, possibly due to its more distinctive visual characteristics (e.g., furrowed brows, tense facial muscles), aligning with findings from [1] and [2] who identified similar inter-class confusion patterns in CNN-based facial expression systems. Augmentation reduced background misclassification across all models, with YOLO11n showing the largest improvement, supporting the interpretation that augmentation is particularly beneficial for models with higher architectural complexity.

4. CONCLUSION

This study conducted a comparative evaluation of three lightweight YOLO models—YOLOv8n, YOLOv10n, and YOLO11n—for the detection of four facial expressions (flat face, angry, sad, and smile). The dataset is also augmented with various augmentation techniques to understand the effect of augmentation on these particular YOLO models. Data augmentation had a significant positive effect on model performance, particularly for YOLOv10n and YOLO11n, which showed F1 score improvements of +0.04 and +0.07, respectively. YOLOv8n maintained a stable F1 score of 0.75 under both conditions, indicating inherent robustness to limited data diversity at this dataset scale.

Per-model confidence threshold optimization proved important for maximizing detection performance, with optimal thresholds ranging from 0.10 to 0.60 across models and conditions. Augmentation also improved confidence calibration, shifting optimal

thresholds upward without sacrificing peak performance. Among the three models, YOLO11n with augmentation achieved the highest precision (0.82), making it suitable for applications where minimizing false detections is critical. YOLOv8n offered the most balanced and stable performance across conditions, making it a practical choice for deployment scenarios with limited data preparation resources.

These findings provide empirical guidance for selecting and configuring lightweight YOLO models for facial expression detection. Future work should extend this comparison to larger and more diverse datasets, incorporate inference speed benchmarking, include repeated-trial evaluation with statistical analysis, and explore additional emotion categories and non-YOLO baseline architectures.

Limitations of this study are explained to contextualize the findings appropriately. Firstly, the four expression classes used (flat face, angry, sad, smile) represent a subset of the commonly studied basic emotions. Extending the evaluation to include additional expressions such as surprise, fear, and disgust would provide a more comprehensive assessment of model capabilities. Secondly, this study compares only within the YOLO family. Including comparisons with other lightweight detection or classification architectures (e.g., SSD MobileNet, EfficientDet, or CNN-based FER classifiers) would provide additional context for the reported results.

REFERENCES

[1] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Oct. 2018, doi: 10.1109/TAFFC.2020.2981446.

[2] S. Ullah, J. Ou, Y. Xie, and W. Tian, "Facial expression recognition (FER) survey: a vision, architectural elements, and future directions," *PeerJ Comput. Sci.*, vol. 10, p. e2024, Jun. 2024, doi: 10.7717/PEERJ-CS.2024.

[3] A. Alshammari and M. E. Alshammari, "Emotional Facial Expression Detection using YOLOv8," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16619–16623, Oct. 2024, doi: 10.48084/etasr.8433.

[4] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo Algorithm Developments," *Procedia Comput. Sci.*, vol. 199, pp. 1066–1073, Jan. 2022, doi: 10.1016/j.procs.2022.01.135.

[5] J. Terven, D. M. Córdova-Esparza, and J. A. Romero-González, "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Machine Learning and Knowledge Extraction 2023, Vol. 5, Pages 1680-1716*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.

[6] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using YOLO: challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications 2022 82:6*, vol. 82, no. 6, pp. 9243–9275, Aug. 2022, doi: 10.1007/s11042-022-13644-y.

[7] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient Acceleration of Deep Learning Inference on Resource-Constrained Edge Devices: A Review," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 42–91, Jan. 2023, doi: 10.1109/JPROC.2022.3226481.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
tip.ppj.unp.ac.id

[8] C.-Y. Wang and H.-Y. M. Liao, "YOLOv1 to YOLOv10: The fastest and most accurate real-time object detection systems," *APSIPA Trans. Signal Inf. Process.*, vol. 13, no. 1, pp. 1–38, Aug. 2024, Accessed: Mar. 01, 2026. [Online]. Available: http://arxiv.org/abs/2408.09332

[9] R. Khanam and M. Hussain, "YOLOv11: An Overview of the Key Architectural Enhancements," Oct. 2024, [Online]. Available: https://doi.org/10.48550/arXiv.2410.17725

[10] D. Novaliendry, F. Rizal, M. Anwar, and D. Irfan, "A Web Application for Classification and Detection of Tomato Leaf Diseases Using CNN and Yolo Models," *Jurnal Teknologi Informasi dan Pendidikan*, vol. 19, no. 1, pp. 1179–1192, Feb. 2026, doi: 10.24036/jtip.v19i1.1073.

[11] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, doi: 10.1109/TIM.2023.3243661.

[12] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data 2019 6:1*, vol. 6, no. 1, pp. 60-, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[13] E. Randellini, L. Rigutini, and C. Sacca', "Data Augmentation and Transfer Learning Approaches Applied to Facial Expressions Recognition," Feb. 2024, doi: 10.5121/csit.2021.111912.

[14] N. E. Khalifa, M. Loey, and S. Mirjalili, "A comprehensive survey of recent trends in deep learning for digital images augmentation," *Artificial Intelligence Review 2021 55:3*, vol. 55, no. 3, pp. 2351–2377, Sep. 2021, doi: 10.1007/s10462-021-10066-4.

[15] Rosalinaa, "'Ekspresi Wajah' Kaggle Dataset." [Online]. Available: https://www.kaggle.com/datasets/rosalinaa/ekspresi-wajah/data

[16] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Aug. 2017, doi: 10.1109/TAFFC.2017.2740923.

[17] S. Wenkel, K. Alhazmi, T. Liiv, S. Alrshoud, and M. Simon, "Confidence Score: The Forgotten Dimension of Object Detection Performance Evaluation," *Sensors 2021, Vol. 21, Page 4350*, vol. 21, no. 13, p. 4350, Jun. 2021, doi: 10.3390/s21134350.