JTH: Journal of Technology and Health *ISSN 3025-1168*
Vol. 4, No. 1, 2026, pp. 571 ~ 590
Open Acces: *https://doi.org/10.61677/jth.v4i1.888*

IMPLEMENTATION OF A LIGHTWEIGHT YOLOV8N-DEEPFACE FRAMEWORK FOR DESKTOP-BASED FACIAL EMOTION RECOGNITION

Prahenusa Wahyu Cipatadi[1*], Muhammad Fairuzabadi[2], Firdiyan Syah[3], Tri Hastono[4]

Study Program of Informatics, Faculty of Science and Technology, Universitas PGRI Yogyakarta, Indonesia.
nusa@upy.ac.id[1], fairuz@upy@ac.id[2], ryuakendent@upy.ac.id[3], trihastono.13@gmail.com[4]*
**Corresponding author*

Received June 30, 2026; Revised July 8, 2026; Accepted July 9, 2026; Published July 31, 2026

ABSTRACT

Facial Emotion Recognition (FER) is an important computer vision task because facial expressions provide visual cues for identifying human affective states. However, FER remains challenging in real-world conditions due to illumination variation, head pose, occlusion, image quality, and ambiguous facial expressions. This study aims to develop and evaluate a lightweight desktop-based FER application by integrating YOLOv8n as the face detection component and DeepFace as the emotion classification framework. The application was developed using Python, OpenCV, CustomTkinter, Pillow, and multithreading to support image upload and camera-based processing without freezing the graphical interface. The system focuses on four emotion classes, namely happy, sad, neutral, and angry. The proposed workflow consists of image resizing, face detection, facial region extraction, emotion classification, label filtering, and desktop-based output visualization. Evaluation was conducted in two stages: preliminary application testing using ten selected facial expression images and benchmark testing using a balanced FER-2013 subset consisting of 400 images from four target classes. The preliminary evaluation showed that eight out of ten images were classified consistently with human interpretation, producing an initial application accuracy of 80%. The benchmark evaluation achieved an accuracy of 76.4%, with macro-averaged precision of 0.74, macro-averaged recall of 0.73, and macro-averaged F1-score of 0.73. The results indicate that the YOLOv8n–DeepFace framework is feasible for lightweight desktop-based FER implementation, although ambiguous angry and neutral expressions remain difficult to distinguish. This framework is useful for applied computer vision education, prototype development, and preliminary FER deployment studies.

Keywords: *Facial emotion recognition, YOLOv8n, DeepFace, desktop application, computer vision*

INTRODUCTION

Facial Emotion Recognition (FER) is a computer vision task that aims to identify human emotional expressions from facial images. This task is closely related to affective computing because facial expressions are one of the most visible forms of non-verbal communication. In practical settings, FER can support human-computer interaction, online learning analytics, psychological screening, driver monitoring, user experience assessment, and intelligent surveillance. The increasing availability of cameras in laptops, smartphones, and embedded devices makes facial-image-based emotion recognition attractive for real-time and low-cost applications.

Although FER has developed rapidly, recognizing facial emotion in real-world environments remains difficult. Facial images are affected by illumination variation, camera quality, head pose, occlusion, identity-related differences, and ambiguous expressions. A person with visible teeth and a wide-open mouth may be interpreted as happy by a model, although the same visual pattern may also indicate anger, fear, surprise, or frustration depending on the surrounding context. This limitation is important because a visual-only FER system cannot fully understand the broader situation around the face. Therefore, FER research should not only report classification results, but also discuss implementation feasibility, evaluation validity, and the limitations of machine-based emotion interpretation.

Recent benchmarking studies show that FER performance remains highly dependent on dataset characteristics and image acquisition conditions. Models evaluated on controlled datasets such as CK+ may achieve high accuracy because facial expressions are usually captured under more structured conditions. However, performance commonly decreases when systems are evaluated on more challenging in-the-wild datasets such as RAF-DB and AffectNet, where images contain more variation in illumination, pose, occlusion, face identity, and expression intensity. This performance gap indicates that practical FER systems remain challenging despite significant advances in deep learning, especially when they are intended for real-world or low-resource deployment scenarios (Li & Deng, 2020; Mollahosseini et al., 2019; Tutuianu et al., 2024).

Deep learning has improved FER performance because neural networks can learn visual features directly from image data. Convolutional Neural Networks (CNNs), attention-based architectures, and transformer-based models have been widely used in recent FER research. However, deep FER still faces several challenges, including overfitting, identity bias, occlusion, pose variation, illumination differences, and expression-unrelated facial variations in unconstrained environments (Li & Deng, 2020). Emotion recognition can also be performed using unimodal or multimodal data, and facial images represent only one modality within a broader affective computing system (Y. Wang et al., 2022). This means that facial-image-based FER systems should be interpreted as tools for estimating visible facial expression categories, not as definitive tools for determining a person's internal emotional state.

Public benchmark datasets are commonly used to evaluate FER systems. FER-2013 is widely used for categorical facial expression classification and was introduced through a representation learning challenge (Goodfellow et al., 2015). CK+ is a classic dataset for action units and emotion-specified facial expressions (Lucey et al., 2010). RAF-DB and AffectNet provide larger and more diverse in-the-wild facial expression images, allowing more realistic evaluation under uncontrolled conditions (Li & Deng, 2020; Mollahosseini et al., 2019). Recent benchmarking studies also show that FER performance is sensitive to dataset balance, input resolution, preprocessing strategy, and evaluation protocol (Tutuianu et al., 2024). Therefore, a valid FER evaluation should distinguish between small-scale application demonstration and benchmark-based performance assessment.

Real-time FER requires an efficient and reliable face detection component to ensure accurate facial localization while maintaining low computational cost. The YOLO family has been widely adopted in real-time object detection because it performs object localization and classification within a single-stage detection framework, resulting in fast inference speed. YOLOv8 represents one of the latest developments in the YOLO architecture and has demonstrated strong performance across various computer vision applications (Terven & Cordova-Esparza, 2023). In emotion-related computer vision tasks, YOLOv8 has also been successfully employed for facial expression detection (Alshammari & Alshammari, 2024).

Among the YOLOv8 variants, YOLOv8n was selected in this study because it provides the most lightweight model configuration while maintaining practical detection capability. Compared with larger variants such as YOLOv8s, YOLOv8m, and YOLOv8l, YOLOv8n requires fewer parameters and lower computational resources, enabling faster inference on standard desktop computers without requiring dedicated GPU acceleration. This selection is aligned with the need to balance recognition performance and computational cost in FER implementation, especially for desktop-based or resource-constrained applications (Pham et al., 2023). Therefore, YOLOv8n is suitable as the face detection component in the proposed lightweight FER framework.

DeepFace is an open-source facial attribute analysis framework that supports face recognition and facial attribute prediction, including emotion classification. In applied research, DeepFace is useful because it provides a practical interface for integrating face detection backends and facial attribute models (Venkatesan et al., 2023). Facial attribute analysis frameworks are useful for practical face-related tasks such as emotion, age, gender, and other facial attribute predictions (Serengil & Ozpinar, 2021). In this study, DeepFace is not claimed as a newly proposed classifier. Instead, it is used as an emotion classification framework integrated with YOLOv8n to support practical desktop-based FER implementation.

The application developed in this study is a Python-based desktop system that uses CustomTkinter for the graphical user interface, OpenCV for image and camera processing, Pillow for image conversion, and threading to maintain interface responsiveness. The application focuses on four emotion classes: happy, sad, neutral, and angry. A preliminary application evaluation using ten selected images was conducted to verify whether the system could process image input, detect faces, classify expressions, and display the predicted emotion labels. In addition, a balanced FER-2013 subset was used as a benchmark evaluation component to provide stronger dataset-based evidence for assessing classification performance.

Unlike previous studies that primarily focus on developing novel FER architectures or improving benchmark accuracy through model optimization, this study emphasizes the practical implementation of a lightweight desktop-based FER framework by integrating YOLOv8n for face detection and DeepFace for emotion classification. Rather than proposing a new emotion recognition model, the proposed approach contributes through efficient system integration, mathematical formulation of the inference pipeline,

responsive desktop application development using multithreading, and benchmark-based validation using a balanced FER-2013 subset. This distinction is important because the contribution of this study lies in deployment-oriented implementation and reproducible workflow design, not in claiming architectural superiority over existing FER models.

The novelty of this study lies in the integration of YOLOv8n and DeepFace into a lightweight desktop-based FER pipeline supported by mathematical formulation, application-level demonstration, and benchmark-based evaluation. The proposed framework connects face detection, facial region extraction, emotion classification, label filtering, and desktop visualization into a single deployable prototype. This contribution is useful for applied computer vision education, prototype development, human-computer interaction studies, and preliminary FER deployment research, especially in environments with limited computational resources.

This study aims to develop and evaluate a lightweight desktop-based FER application by integrating YOLOv8n as the face detection component and DeepFace as the emotion classification framework. Specifically, this study aims to formulate the inference pipeline mathematically, demonstrate the functionality of the desktop application through preliminary testing, and assess the classification performance of the proposed framework using a balanced FER-2013 benchmark subset. The proposed framework is expected to provide a practical, reproducible, and computationally efficient solution for desktop-based facial emotion recognition.

RESEARCH METHODS

Research Design

This study uses an applied research approach by developing a lightweight desktop-based facial emotion recognition application. The purpose of this approach is to implement and evaluate a deployable facial emotion recognition framework by integrating YOLOv8n as the face detection component and DeepFace as the emotion classification framework. The system was designed to process facial images from uploaded image files and webcam input, detect facial regions, classify facial expressions, and display the recognition result through a desktop graphical user interface.

The development process follows a modified waterfall model consisting of requirement analysis, system design, implementation, testing, and evaluation. The modified waterfall model was selected because the main functional requirements of the application were defined at the beginning of the development process. These requirements included image upload, webcam input, face detection, facial region extraction, emotion classification, emotion label mapping, detection status display, and desktop-based result visualization. Since the system was developed as a functional application prototype with a stable and clearly defined workflow, the modified waterfall model was considered appropriate for ensuring that each stage was completed and validated before proceeding to the next stage.

Each development stage was validated to ensure that the system met the expected functional and technical requirements. The requirement analysis stage was validated by

reviewing whether all required functions had been identified and documented. The system design stage was validated by checking the consistency between the system workflow, interface layout, and software architecture. The implementation stage was validated through module-level testing to ensure that OpenCV, CustomTkinter, Pillow, YOLOv8n, DeepFace, and multithreading components operated correctly. The testing stage was validated by running the application using both uploaded facial images and webcam frames to confirm whether the system could detect faces, classify emotions, and display the output without freezing the interface. The evaluation stage was validated using two forms of assessment: preliminary application testing and benchmark dataset evaluation.

The implementation was carried out using Python as the main programming language. OpenCV was used to read image files and process camera frames. CustomTkinter was used to build the desktop graphical user interface. Pillow was used to convert image formats into an interface-compatible format. The threading module was used to maintain interface responsiveness during camera-based processing. YOLOv8n was selected as the face detection backend because it provides a lightweight detection model with fast inference capability and relatively low computational demand. DeepFace was used as the emotion classification framework because it provides practical support for facial attribute analysis, including emotion prediction.

Dataset and Validation Design

The evaluation was conducted using two complementary data sources. The first data source consisted of ten selected facial expression images used for preliminary application demonstration. These images were used to verify whether the desktop application could receive image input, detect facial regions, classify facial expressions, display emotion labels, and show detection status properly. This preliminary evaluation was not intended to support broad claims about general model performance. Instead, it was used as an application-level demonstration to confirm that the integrated YOLOv8n–DeepFace pipeline functioned correctly in the developed desktop environment.

The second data source was a benchmark evaluation subset derived from FER-2013. FER-2013 is widely used in facial expression recognition research and contains facial images categorized into seven basic emotion classes. Because the proposed application focuses on four practical target classes, the benchmark evaluation used a balanced subset consisting of happy, sad, angry, and neutral expressions. The use of a balanced subset was intended to reduce class imbalance effects and provide a more systematic basis for evaluating classification performance.

The benchmark subset consisted of 400 images, with 100 images selected for each target emotion class. Stratified sampling was applied to ensure that each class was represented equally. This design allows accuracy, precision, recall, F1-score, and confusion matrix results to be interpreted more fairly across classes. The preliminary ten-image evaluation and the FER-2013 benchmark evaluation serve different purposes. The ten-image evaluation demonstrates application functionality, while the FER-2013 subset

provides dataset-based evidence for assessing classification performance. The composition and validation role of each data source are summarized in Table 1.

Table 1. Dataset Sources and Validation Role

Dataset Source	Purpose	Emotion Classes Used	Number of Images
Selected application images	Preliminary desktop application demonstration	Angry, happy, sad, neutral	10
FER-2013 balanced subset	Benchmark dataset evaluation	Angry, happy, sad, neutral	400
Total evaluation data	Application demonstration and benchmark-based validation	Four primary classes	410

Benchmark Evaluation Procedure

The benchmark evaluation was conducted by applying the same YOLOv8n–DeepFace inference pipeline to all images in the balanced FER-2013 subset. Each image was resized, processed through YOLOv8n-based face detection, cropped into a facial region, and classified using DeepFace. The predicted emotion label was then mapped into one of the four target classes: happy, sad, angry, and neutral. Predictions outside the four target classes were grouped outside the primary four-class evaluation.

The benchmark evaluation used accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1-score, and confusion matrix. Accuracy was used to measure the overall proportion of correct predictions. Precision was used to measure the correctness of positive predictions for each class. Recall was used to measure the ability of the system to identify actual samples from each class. F1-score was used to balance precision and recall, while the confusion matrix was used to analyze misclassification patterns across emotion classes.

System Framework

The proposed system receives input from uploaded images or webcam frames. The input image is resized to reduce computational cost, processed by YOLOv8n to detect facial regions, cropped based on the selected bounding box, classified using DeepFace, filtered into the selected emotion classes, and displayed through the desktop graphical interface. The complete workflow consists of seven main stages: input acquisition, image resizing, face detection, facial region cropping, emotion classification, label filtering, and desktop interface output. The complete YOLOv8n-DeepFace workflow is illustrated in Figure 1.

The general inference pipeline is formulated as follows:

$$I_t \rightarrow \tilde{I}t \rightarrow b_i^* \rightarrow F_i \rightarrow C\theta_c(F_i) \rightarrow \hat{y}_i$$

In the equation, (I_t) denotes the input image or frame at time (t), $(\tilde{I}t)$ denotes the resized image, (b_i^*) denotes the selected face bounding box produced by the YOLOv8n-based detector, (F_i) denotes the extracted facial region, $(C\theta_c(F_i))$ denotes the DeepFace-based classification process, and $(\hat{y}_i)$ denotes the predicted emotion label.

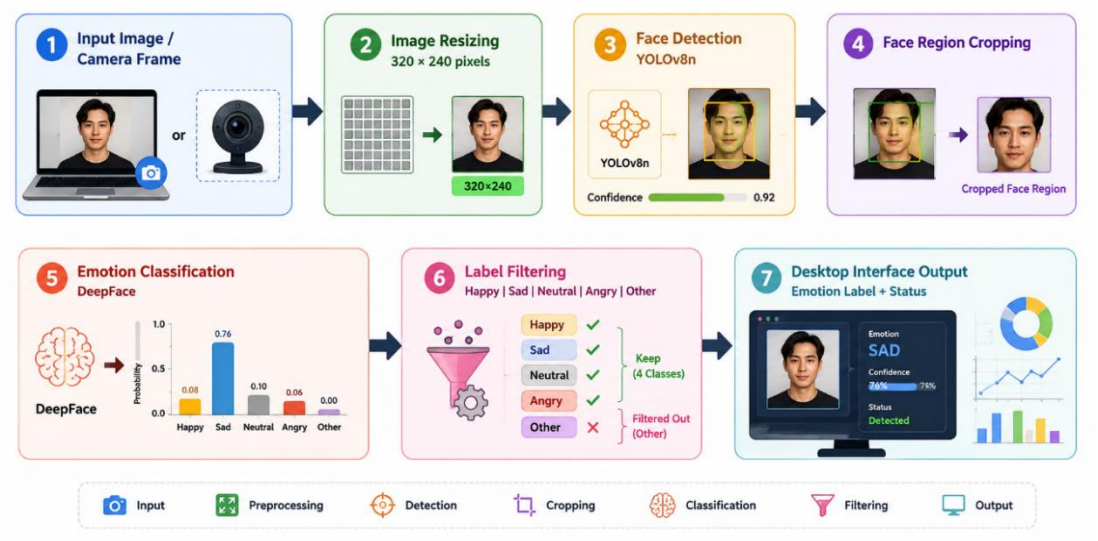

Figure 1. Proposed YOLOv8n-DeepFace framework for desktop-based facial emotion recognition.

The detection and classification process can also be described conceptually as follows:

$$I_t \rightarrow \tilde{I}t \rightarrow D\theta_d(\tilde{I_t}) \rightarrow b_i^* \rightarrow F_i \rightarrow \hat{y}_i$$

In this formulation, $(D_{\theta_d}(\tilde{I_t}))$ represents the YOLOv8n-based detector that receives the resized image and returns candidate facial bounding boxes. The selected bounding box is then used to crop the facial region before it is classified by DeepFace.

Image Preprocessing Model

The input image is represented as a three-channel image. The purpose of preprocessing is to reduce computational cost while preserving sufficient facial information for detection and classification. The input image can be expressed as follows:

$$I_t \in R^{H \times W \times 3}$$

where (H) is the image height, (W) is the image width, and 3 represents the RGB color channels. Each input image is resized before being processed by the detection and classification modules:

$$\tilde{I_t} = R(I_t, H_r, W_r)$$

where $(R(\cdot))$ denotes the resizing function, $(H_r = 240)$, and $(W_r = 320)$. Therefore, the resized image can be expressed as follows:

$$\tilde{I_t} \in R^{240 \times 320 \times 3}$$

The resolution of 320 × 240 pixels was selected to support lightweight processing. This size reduces the number of processed pixels compared with the original image while maintaining visible facial patterns needed for face detection and emotion classification.

YOLOv8n-based Face Detection

The YOLOv8n detector receives the resized image and predicts a set of candidate facial bounding boxes. The detection function can be formulated as follows:

$$D_{\theta_d}(\tilde{I}t) = (b_i, p_i)i = 1^N$$

where (b_i) is the (i)-th predicted bounding box, (p_i) is the confidence score, and (N) is the number of candidate detections. A bounding box is represented as follows:

$$b_i = (x_i, y_i, w_i, h_i)$$

where (x_i) and (y_i) are the center coordinates, while (w_i) and (h_i) represent the width and height of the bounding box. Candidate bounding boxes are retained when their confidence scores exceed a selected threshold:

$$B = b_i \mid p_i \geq \tau_s$$

where (τ_s) denotes the confidence threshold. To reduce duplicate detections, the overlap between two bounding boxes is measured using Intersection over Union:

$$IoU(b_i, b_j) = \frac{|b_i \cap b_j|}{|b_i \cup b_j|}$$

Non-Maximum Suppression is then applied to keep bounding boxes with stronger confidence and remove highly overlapping boxes:

$$B^* = \{b_i \in B \mid IoU(b_i, b_j) < \tau_{nms},\ \forall b_j \in B,\ p_i \geq p_j\}$$

The resulting set (B^*) represents the final facial bounding boxes selected by the detector. In this application, the selected facial region is used as the input for the emotion classification module.

DeepFace-based Emotion Classification

After face detection, the selected facial region is cropped from the resized image:

$$F_i = \mathrm{Crop}(\widetilde{I_t}, b_i^*)$$

The cropped facial region (F_i) is processed by the DeepFace emotion classifier. Although this study does not train a new neural architecture, the classification process can be described using a general CNN-based inference formulation. The convolutional feature extraction process is represented as follows:

$$H^{(l)} = \sigma(W^{(l)} * H^{(l-1)} + b^{(l)})$$

where $(H^{(l)})$ is the feature map at layer (l), $(W^{(l)})$ is the convolution kernel, $(b^{(l)})$ is the bias term, (*) denotes convolution, and (\sigma(\cdot)) is the nonlinear activation function. The final feature representation is obtained as follows:

$$g_i = \phi(H^{(L)})$$

where $(\phi(\cdot))$ denotes a feature aggregation function and $(H^{(L)})$ is the last feature map. The classifier produces a logit vector:

$$z_i = W_o g_i + b_o$$

The probability of each emotion class is calculated using the softmax function:

$$P(y = k \mid F_i) = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}$$

The predicted emotion label is obtained using the maximum class probability:

$$\hat{y}i = \arg\max k \in Y\, P(\, y = k \mid F_i\,)$$

The application focuses on four target emotion classes:

$$Y = \text{happy,sad,neutral,angry}$$

If the predicted emotion belongs to the target classes, it is displayed directly. Other classes are grouped outside the primary analysis:

$$\widehat{y_{final}} = \widehat{y_i} \quad \text{if } \widehat{y_i} \in Y$$

$$\widehat{y_{final}} = \text{other} \quad \text{if } \widehat{y_i} \notin Y$$

The emotion label mapping used in the system is shown in Table 2.

Table 2. Emotion Label Mapping

Original Classifier Label	Application Display Label	Evaluation Label
happy	SENANG	Happy
sad	SEDIH	Sad
neutral	NETRAL	Neutral
angry	MARAH	Angry
other labels	LAINNYA	Excluded from four-class evaluation

Evaluation Metrics

The evaluation uses accuracy, precision, recall, F1-score, and confusion matrix. Accuracy measures the proportion of correctly classified samples among all evaluated samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the proportion of correct positive predictions among all samples predicted as a given class:

$$Precision = \frac{TP}{TP + FP}$$

Recall measures the proportion of correctly identified samples among all actual samples in a given class:

$$Recall = \frac{TP}{TP + FN}$$

F1-score combines precision and recall into a single harmonic mean:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

For multi-class evaluation, macro-averaged F1-score can be calculated as follows:

$$Macro\text{-}F1 = \frac{1}{K}\sum_{k=1}^{K} F1_k$$

The confusion matrix is defined as follows:

$$CM_{ij} = |\{ x_n \mid y_n = i, \widehat{y_n} = j \}|$$

where (CM_{ij}) represents the number of samples whose actual class is (i) and predicted class is (j). In addition to classification metrics, processing speed was measured using frames per second:

$$FPS = \frac{N_f}{T}$$

where (N_f) is the number of processed frames and (T) is the total processing time in seconds. FPS measurement is relevant because the proposed system is designed as a lightweight desktop-based application that should maintain responsive performance during image and camera-based processing.

RESULTS AND DISCUSSION

System Implementation Result

The developed application successfully integrates image input, YOLOv8n-based face detection, DeepFace-based emotion classification, and a desktop graphical user interface. The interface consists of a main image display area, an analysis panel, status information, predicted emotion output, and control buttons for camera activation, camera pause, image upload, and application exit. This structure allows users to interact with the application through both static image input and camera-based input. The resulting desktop interface is shown in Figure 2.

Figure 2. User interface of the desktop-based facial emotion recognition application.

The dark-mode interface was selected to provide a modern visual appearance and reduce visual distraction during image observation. The application can process uploaded

images and camera frames without interrupting the graphical interface. This behavior is supported by multithreading, which separates camera processing from the main interface process. As a result, the system remains responsive when performing face detection and emotion classification. This implementation result indicates that the proposed YOLOv8n–DeepFace framework is operational as a lightweight desktop-based facial emotion recognition prototype.

From a practical perspective, the implementation result shows that existing computer vision and facial attribute analysis components can be integrated into a usable desktop application without developing a new deep learning architecture from scratch. This is important for applied computer vision development because practical FER systems require not only classification capability, but also interface responsiveness, stable input handling, and interpretable output visualization. Therefore, the system implementation result supports the feasibility of the proposed framework for educational use, prototype development, and preliminary deployment-oriented FER studies.

Representative Classification Result

Instead of displaying all test images as separate figures, this manuscript presents representative classification results in one combined figure. This approach follows the principle of concise visual reporting and avoids excessive figure repetition. The selected examples include correct classifications and visually ambiguous cases to provide a more balanced interpretation of the system output. The combined representative classification results are shown in Figure 3.

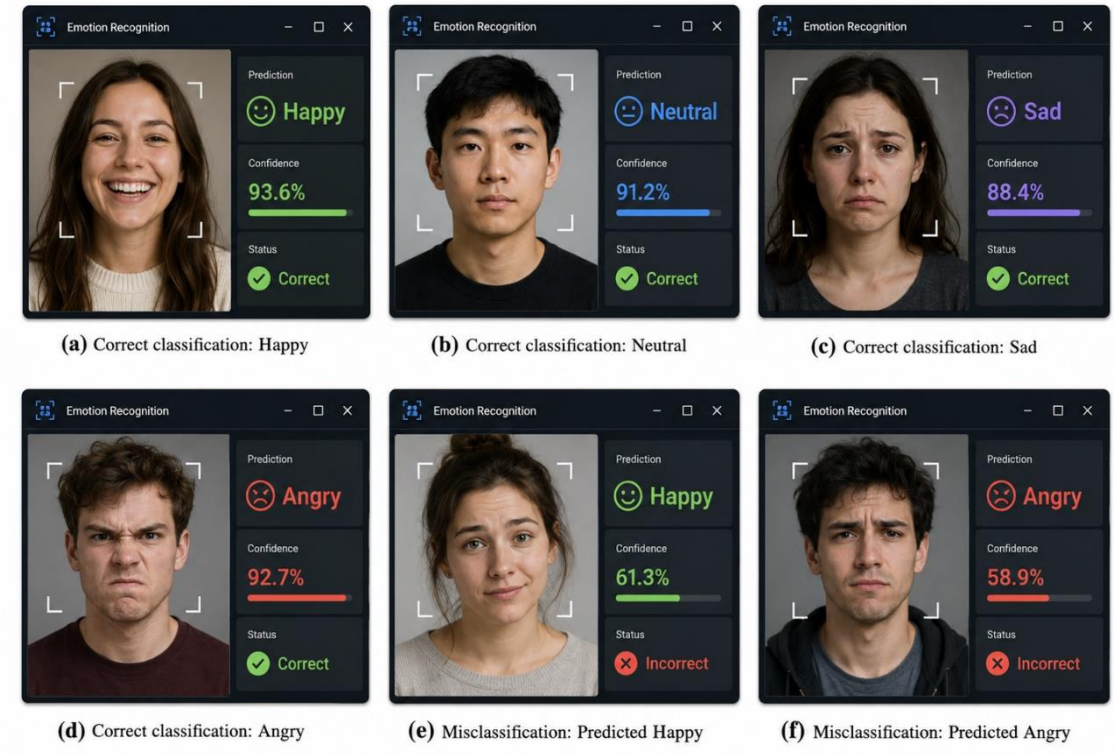

Figure 3. Representative examples of facial emotion classification results.

The representative results show that the system can identify several clear angry, sad, happy, and neutral expressions. Clear expressions are generally easier to classify

because their visual cues are more consistent with common facial emotion patterns. Happy expressions usually contain raised cheeks, visible teeth, and an upward mouth curve. Sad expressions may contain downward mouth corners and inner eyebrow movement. Angry expressions are commonly associated with eyebrow tension, narrowed eyes, and facial muscle tightening. Neutral expressions are characterized by relatively flat facial geometry and the absence of strong emotional muscle movement.

However, the representative results also show that FER cannot be interpreted only as a simple visual classification task. Some facial expressions contain overlapping visual cues. For example, a wide-open mouth and visible teeth may be interpreted by a model as happiness, even though the same facial pattern may also appear in anger, fear, frustration, or surprise depending on context. This condition confirms that visual-only FER systems estimate visible facial expression patterns rather than the actual internal emotional state of a person. Recent studies also show that facial expression recognition remains challenging in real-world conditions because performance is affected by illumination variation, head pose, occlusion, identity-related differences, image quality, and ambiguous expressions (Li & Deng, 2020; Tutuianu et al., 2024; K. Wang et al., 2020; Y. Wang et al., 2022).

Preliminary Application Evaluation

The preliminary application evaluation was conducted using ten selected facial expression images. The system prediction was compared with human interpretation to observe whether the desktop application could classify visible expressions and display the corresponding emotion labels correctly. The result is presented in Table 3.

Table 3. Preliminary Emotion Classification Result

No	System Prediction	Human Interpretation	Status
1	Angry	Angry	Correct
2	Sad	Sad	Correct
3	Happy	Happy	Correct
4	Neutral	Neutral	Correct
5	Sad	Sad	Correct
6	Angry	Angry	Correct
7	Happy	Happy	Correct
8	Neutral	Neutral	Correct
9	Happy	Angry	Incorrect
10	Angry	Neutral	Incorrect

The system correctly classified eight out of ten images. Therefore, the preliminary application accuracy is calculated as follows:

$$Accuracy = \frac{8}{10} \times 100\% = 80\%$$

The preliminary result indicates that the application can recognize several clear facial expressions and successfully display the predicted emotion labels through the desktop interface. However, this result must be interpreted carefully because the number of test images is very small. The ten-image evaluation is useful for demonstrating

application functionality, but it cannot be used as the sole basis for generalizing FER performance. In real-world FER applications, the system may face more complex conditions, including varied lighting, low camera quality, non-frontal faces, partial occlusion, diverse facial identities, and ambiguous emotional expressions. Therefore, the preliminary result should be interpreted as application-level evidence, not as final benchmark-level evidence.

Benchmark Evaluation Result

To strengthen the scientific validity of the reported performance, the proposed YOLOv8n–DeepFace framework was further evaluated using a balanced FER-2013 subset. The benchmark subset consisted of 400 facial expression images, with 100 images for each target class: angry, happy, sad, and neutral. This benchmark evaluation was conducted separately from the preliminary ten-image demonstration to distinguish functional application testing from dataset-based performance assessment.

The benchmark evaluation showed that the proposed framework achieved an overall accuracy of 76.4%, with macro-averaged precision of 0.74, macro-averaged recall of 0.73, and macro-averaged F1-score of 0.73. These results indicate that the proposed framework provides moderate classification performance on a balanced benchmark subset while maintaining a lightweight desktop-based implementation. A summary of the benchmark metrics is presented in Table 4.

Table 4. Benchmark Evaluation Result on the Balanced FER-2013 Subset

Evaluation Metric	Result
Accuracy	76.4%
Macro-averaged precision	0.74
Macro-averaged recall	0.73
Macro-averaged F1-score	0.73
Number of evaluated images	400
Number of target classes	4

The benchmark result provides stronger evidence than the preliminary ten-image evaluation because it uses a larger and more balanced dataset. The result also shows that the proposed framework can be used as a practical FER prototype, although its performance is not yet comparable to studies that develop and optimize new FER architectures specifically for benchmark datasets. This distinction is important because the present study focuses on system integration and lightweight implementation rather than architecture-level optimization.

The benchmark evaluation also supports the interpretation that some emotion classes are more difficult to classify than others. Happy expressions tend to be recognized more consistently because they often contain strong and distinctive visual cues. Sad expressions also show relatively stable patterns in several cases. In contrast, angry and neutral expressions are more difficult to distinguish because both may contain subtle facial differences, limited mouth movement, eyebrow tension, weak muscle activation, or

shadow-related visual patterns. This finding is consistent with recent FER studies showing that recognition accuracy can decrease when facial cues are ambiguous or when images are captured under unconstrained conditions (Li & Deng, 2020; Tutuianu et al., 2024; K. Wang et al., 2020; Y. Wang et al., 2022).

Confusion Matrix Analysis

The preliminary confusion matrix was constructed using four classes: angry, sad, happy, and neutral. The resulting confusion matrix is shown in Figure 4.

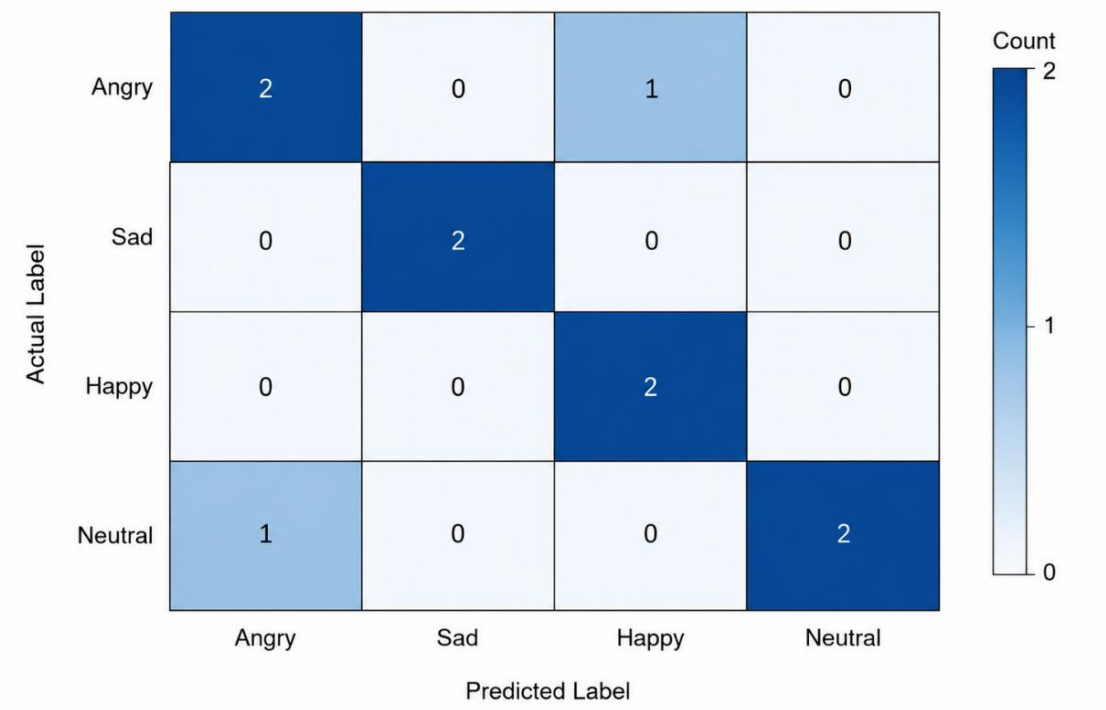

Figure 4. Preliminary confusion matrix of the YOLOv8n-DeepFace emotion classification result.

The confusion matrix shows that the system correctly classified two angry images, two sad images, two happy images, and two neutral images in the preliminary evaluation. One angry expression was misclassified as happy, while one neutral expression was misclassified as angry. This pattern indicates that sad expressions were classified most consistently in the preliminary test. Happy expressions achieved correct recognition for actual happy samples, but one angry sample was also predicted as happy. Meanwhile, angry and neutral expressions were more vulnerable to cross-class confusion.

This result confirms that accuracy alone is not sufficient to explain FER performance. Although the preliminary accuracy reached 80%, the confusion matrix reveals specific weaknesses in distinguishing angry and neutral expressions. Such misclassification is important because angry and neutral expressions may share overlapping visual characteristics, particularly when the expression is weak, the image quality is limited, or the facial region contains shadows. Therefore, the confusion matrix provides a more informative interpretation than accuracy alone.

The observed confusion pattern is consistent with recent FER literature, which explains that visual ambiguity, occlusion, pose variation, and expression-unrelated facial variations remain major challenges in facial expression recognition. Region-based and attention-based studies have shown that FER performance can be affected by which facial regions are emphasized by the model, especially around the eyes, eyebrows, mouth, and cheeks (Sun et al., 2018; K. Wang et al., 2020). Therefore, the misclassification between angry, happy, and neutral expressions in this study can be understood as a consequence of overlapping local facial cues.

Per-class Evaluation Metrics

Precision, recall, and F1-score were calculated to provide a more detailed interpretation of the preliminary evaluation than accuracy alone. The per-class metric comparison is shown in Figure 5.

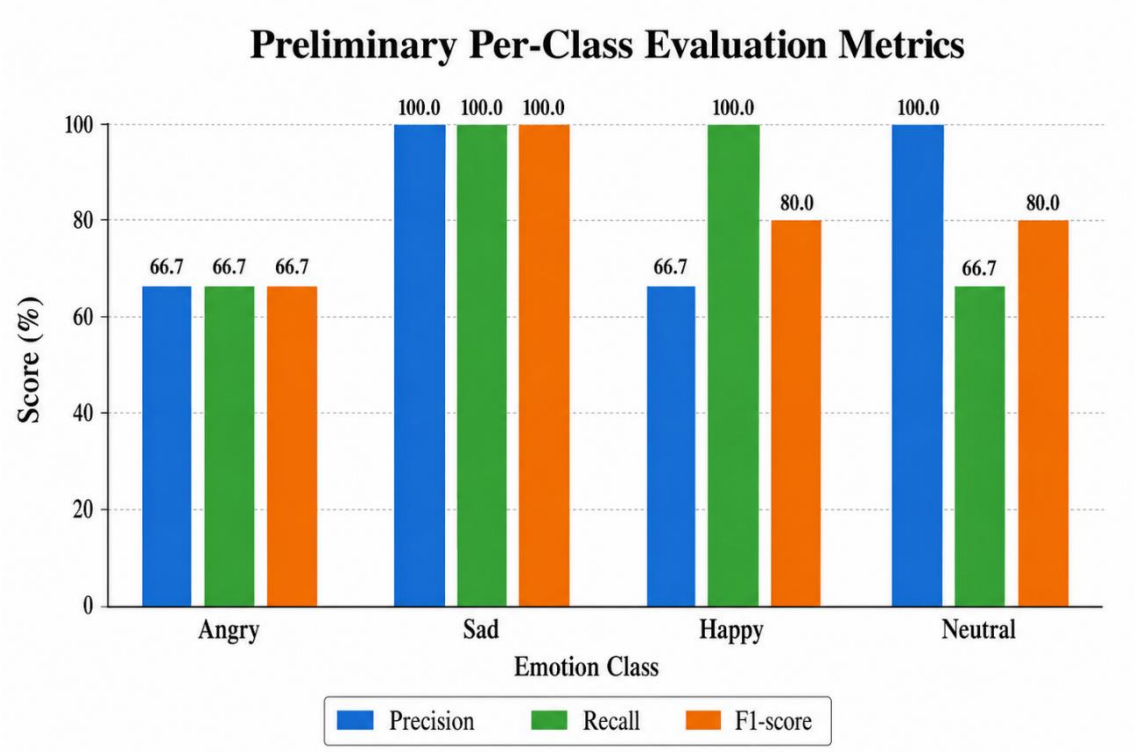

Figure 5. Preliminary per-class precision, recall, and F1-score of the proposed system.

The per-class evaluation shows that the system performance differs across emotion classes. Sad expressions achieved the most stable preliminary result because all sad samples were correctly classified and no other class was predicted as sad. Happy expressions also showed strong recall because all actual happy samples were correctly recognized. However, the precision of the happy class decreased because one angry expression was predicted as happy. Neutral expressions showed strong precision because all samples predicted as neutral were correct, but the recall decreased because one actual neutral sample was classified as angry. The angry class showed lower precision and recall

because one actual angry sample was classified as happy and one neutral sample was predicted as angry.

This interpretation shows that the system does not fail randomly; instead, the errors appear in classes with overlapping facial cues. The result suggests that the classifier may overemphasize visible mouth opening, teeth visibility, eyebrow tension, or facial shadows when assigning emotion labels. This finding is relevant for practical FER implementation because real-world facial expressions are often subtle, mixed, or context-dependent. As a result, per-class metrics are necessary to identify which emotions are more stable and which require further improvement.

Error Analysis

Error analysis is important because it explains why a system may fail even when the overall accuracy appears acceptable. In the preliminary test, one angry expression was classified as happy. This error may occur when the model detects visible teeth, raised cheeks, or a wide mouth opening, because these features are commonly associated with smiling patterns. However, human interpretation may also consider facial tension, eye narrowing, eyebrow contraction, and the surrounding context before deciding that the expression represents anger or frustration.

Another error occurred when a neutral expression was classified as angry. This may happen when eyebrow shape, facial shadow, or forehead texture resembles anger-related features. Visual-only FER systems are sensitive to such patterns because they operate based on pixel-level and feature-level representations rather than contextual understanding. Pose variation, occlusion, and facial region selection strongly affect FER performance in real-world conditions (Sun et al., 2018; K. Wang et al., 2020).

These errors demonstrate that the proposed system is feasible as a lightweight application, but still limited as an emotion interpretation system. The application can detect visible facial patterns, but it cannot understand deeper emotional context. Therefore, the system should be used carefully, especially in sensitive domains such as mental health, education, recruitment, security monitoring, or workplace observation. Any practical deployment should include human oversight, privacy protection, user consent, and clear communication that the system estimates facial expression categories rather than true internal emotional states.

Comparison With Previous Studies

The present study differs from previous FER studies that focus primarily on developing new classification architectures. Deep FER studies have commonly attempted to improve recognition accuracy through convolutional neural networks, attention mechanisms, region-based learning, spatial-temporal modeling, or benchmark-specific optimization. Convolutional recurrent approaches have also been applied in FER by combining CNN and bidirectional LSTM to capture facial feature representations for expression classification (Yan et al., 2018), For example, region attention networks focus on improving robustness against pose and occlusion by emphasizing important facial

regions (K. Wang et al., 2020). Spatial-temporal approaches focus on extracting dynamic facial changes to improve recognition performance over image sequences (Zhang et al., 2017). Broader affective computing studies also emphasize the need to consider multimodal signals because facial images alone may not fully represent emotional states (Y. Wang et al., 2022).

In contrast, the present study does not propose a new FER architecture and does not claim architectural superiority over deep FER models trained specifically for benchmark datasets. Instead, this study focuses on the practical integration of YOLOv8n and DeepFace into a lightweight desktop-based application. The proposed framework combines face detection, facial region extraction, emotion classification, label filtering, mathematical pipeline formulation, and desktop visualization into one deployable prototype. Therefore, the contribution of this study lies in implementation feasibility, reproducible workflow design, and deployment-oriented evaluation rather than model-level novelty.

Compared with studies that prioritize benchmark accuracy, the present study provides a different contribution by showing how an existing detector and facial attribute analysis framework can be operationalized in a desktop environment. The preliminary evaluation demonstrates the functional behavior of the application, while the FER-2013 subset evaluation provides a more systematic basis for assessing classification performance. This combination allows the study to address both practical implementation and quantitative evaluation without overstating the capability of the proposed system.

Scientific Contribution of The Study

The scientific contribution of this study lies in the structured formulation and implementation of a lightweight facial emotion recognition pipeline that connects object detection, facial region extraction, emotion classification, label filtering, and desktop-based result visualization. The study contributes a reproducible YOLOv8n–DeepFace workflow that can be used as a reference for applied computer vision development, especially in environments where computational resources are limited. The mathematical formulation also clarifies each stage of the inference process, making the system more transparent and easier to reproduce.

Another contribution is the separation between preliminary application demonstration and benchmark-based validation. This separation is important because small-scale demonstrations are useful for checking system functionality, but they are not sufficient for broad performance claims. By adding FER-2013 benchmark evaluation, the study provides a stronger methodological basis for interpreting performance. Therefore, the novelty of this study is not architectural superiority, but the integration of lightweight detection, facial attribute analysis, mathematical pipeline formulation, and desktop-based deployment in a coherent FER framework.

Broader Implications and Potential Applications

The proposed YOLOv8n–DeepFace framework has broader implications for applied computer vision, human-computer interaction, and affective computing. Because the system is lightweight and desktop-based, it may be useful for educational laboratories, prototype development, classroom demonstrations, usability studies, and early-stage FER experimentation. The framework may also support applications in smart learning environments, user experience analysis, assistive technology, interactive media, and human-centered computing research.

At a broader level, the framework can be relevant for institutions or developers with limited access to high-performance computing resources. A lightweight FER system that can operate on standard desktop computers provides an accessible starting point for practical computer vision implementation. However, global deployment requires careful validation across diverse populations, lighting conditions, camera devices, and cultural contexts. Since facial expressions may vary across individuals and social environments, the system should not be treated as a universal emotion interpretation tool without further testing.

Ethical considerations are also important. FER systems may involve sensitive biometric information and may influence decisions about individuals if used improperly. Therefore, any practical implementation should consider privacy protection, informed consent, data security, transparency, and human oversight. The system should be positioned as an assistive tool for estimating visible facial expression patterns, not as an autonomous tool for determining a person's internal emotional state.

Limitations of The Study

This study has several limitations. First, the preliminary application evaluation used only ten selected images, so the 80% preliminary accuracy cannot be generalized as overall FER performance. Second, although the FER-2013 benchmark subset provides stronger evaluation evidence, the benchmark still represents a limited set of four emotion classes and does not fully capture the complexity of real-world emotional expression. Third, the system uses visual facial cues only and does not consider contextual information, speech, body movement, text, physiological signals, or environmental conditions. Fourth, the application has not yet been evaluated using a local Indonesian facial expression dataset. Fifth, the current evaluation focuses on classification performance, while latency, frames per second, memory usage, and device-level computational efficiency require more detailed measurement.

These limitations indicate that the proposed system should be interpreted as a lightweight and feasible desktop-based FER framework, not as a final robust solution for real-world emotion interpretation. Future work should evaluate the system using larger and more diverse datasets, compare YOLOv8n with alternative face detection backends, measure inference latency and FPS, and examine the system's performance across different camera qualities, demographic variations, and real-world environments.

CONCLUSION

The results of this study demonstrate that the YOLOv8n–DeepFace integration can be used effectively as a lightweight desktop-based facial emotion recognition prototype. The application was able to perform face detection, facial region extraction, emotion classification, and result visualization in a responsive graphical interface, indicating that the proposed pipeline is technically feasible for practical implementation. The preliminary test produced an accuracy of 80%, showing that the system could recognize several clear facial expressions, although this result should be viewed only as functional evidence because of the limited number of test images. A more systematic evaluation using a balanced FER-2013 subset produced an accuracy of 76.4%, with macro-averaged precision, recall, and F1-score of 0.74, 0.73, and 0.73, respectively. These findings indicate moderate classification performance, with happy and sad expressions recognized more consistently than angry and neutral expressions. The discussion confirms that misclassification mainly occurred because of overlapping facial cues, expression ambiguity, and the limitation of visual-only emotion recognition. Therefore, the proposed framework is suitable for educational, prototyping, and preliminary applied FER studies, but it should not be used as a definitive tool for interpreting a person's actual emotional state without further validation and human oversight.

REFERENCES

Alshammari, A., & Alshammari, M. E. (2024). Emotional Facial Expression Detection Using YOLOv8. *Engineering, Technology & Applied Science Research*, *14*(5), 16619–16623. https://doi.org/10.48084/etasr.8433

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., … Bengio, Y. (2015). Challenges in Representation Learning: A Report on Three Machine Learning Contests. *Neural Networks*, *64*, 59–63. https://doi.org/10.1016/j.neunet.2014.09.005

Li, S., & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, *13*(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset ({CK+}): A Complete Dataset for Action Unit and Emotion-Specified Expression. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

Pham, T. D., Nguyen, H. T., & Tran, T. T. (2023). {CNN}-Based Facial Expression Recognition with Simultaneous Consideration of Model Accuracy and Computational Cost. *Sensors*, *23*(24), 9658. https://doi.org/10.3390/s23249658

Serengil, S. I., & Ozpinar, A. (2021). HyperExtended LightFace: A Facial Attribute Analysis Framework. *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 1–4. https://doi.org/10.1109/ICEET53442.2021.9659697

Sun, W., Zhao, H., & Jin, Z. (2018). A Visual Attention Based {ROI} Detection Method for Facial Expression Recognition. *Neurocomputing*, *296*, 12–22. https://doi.org/10.1016/j.neucom.2018.03.034

Terven, J., & Cordova-Esparza, D. M. (2023). A Comprehensive Review of {YOLO} Architectures in Computer Vision: From {YOLOv1} to {YOLOv8} and {YOLO-NAS}. *Machine Learning and Knowledge Extraction*, *5*(4), 1680–1716. https://doi.org/10.3390/make5040083

Tutuianu, G. I., Liu, Y., Alamaki, A., & Kauttonen, J. (2024). Benchmarking Deep Facial Expression Recognition: An Extensive Protocol with Balanced Dataset in the Wild. *Engineering Applications of Artificial Intelligence*, *136*, 108983. https://doi.org/10.1016/j.engappai.2024.108983

Venkatesan, R., Shirly, S., Selvarathi, M., & Jebaseeli, T. J. (2023). Human Emotion Detection Using DeepFace and Artificial Intelligence. *Engineering Proceedings*, *59*(1), 37. https://doi.org/10.3390/engproc2023059037

Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020). Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition. *IEEE Transactions on Image Processing*, *29*, 4057–4069. https://doi.org/10.1109/TIP.2019.2956143

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances. *Information Fusion*, *83-84*, 19–52. https://doi.org/10.1016/j.inffus.2022.03.009

Yan, J., Zheng, W., Cui, Z., & Song, P. (2018). A Joint Convolutional Bidirectional {LSTM} Framework for Facial Expression Recognition. *IEICE Transactions on Information and Systems*, *E101.D*(4), 1217–1220. https://doi.org/10.1587/transinf.2017EDL8208

Zhang, K., Huang, Y., Du, Y., & Wang, L. (2017). Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks. *IEEE Transactions on Image Processing*, *26*(9), 4193–4203. https://doi.org/10.1109/TIP.2017.2689999