International Journal of Basic and Applied Sciences, 14 (4) (2025) 6-16

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/vxgf1w29
Research paper

Application of Facial Expression Using YOLOv11 to Measure Interest in Training Materials

I. Imran [1] *, W. Wardi [2] *, Ingrid Nurtanio [1], Faizal Arya Samman [1]

[1] Department of Electrical Engineering, Hasanuddin University, Gowa, Indonesia
[2] Department of Informatics Engineering, Hasanuddin University, Gowa, Indonesia
**Corresponding author E-mail: wardi@unhas.ac.id*

Received: June 13, 2025, Accepted: July 16, 2025, Published: August 1, 2025

Abstract

The purpose of this study is to assess trainees' proficiency with the Convolutional Neural Network (CNN) model in recognizing emotions from facial expressions. A wide range of applications in training, education, and human-machine interaction could benefit greatly from computer vision-based emotion recognition. A CNN model created especially to identify important emotions including happiness, sadness, anger, and fear is trained and tested in this study using a dataset of facial expressions. The test findings demonstrate that the CNN model can accurately and efficiently classify emotions and offer valuable information about trainees' strengths and limitations in identifying different emotions. This study emphasizes how CNN-based technology may be used to help assess and enhance emotion recognition skills in the setting of professional training.

Keywords: *Facial Expression Recognition; YOLOv11; Interest Detection; Training Materials Evaluation; Deep Learning in Education.*

1. Introduction

The most important aspect of comprehending nonverbal communication (Kurniadi & M. Ridho Mahaputra, 2021), which is essential to human interaction, is the ability to identify emotions from facial expressions (Manalu & Rifai, 2024). Effectively recognizing and addressing emotions can enhance communication in both personal and professional settings. Training participants' emotional intelligence and interpersonal skills can be enhanced by emotion recognition techniques (Thanh Thao et al., 2023), which is beneficial for leadership, customer service, education, and other professions.

Since computer vision and artificial intelligence (AI) technologies advance quickly (Rashid & Kausik, 2024), face expression analysis with deep learning models like Convolutional Neural Networks (CNN) has become a reliable and effective technique (Elsheikh et al., 2024). CNN is a type of artificial neural network architecture that was created especially for processing visual input (Taye, 2023). It can identify intricate patterns in pictures, such as emotional facial expressions (Talukder & Ghosh, 2024). By using CNN for emotion recognition testing (Krishna et al., 2024), it becomes possible to examine trainees' skills in a more methodical and impartial setting (Avital et al., 2024; Chutia & Baruah, 2024; Khare et al., 2024) .

The goal of this study is to create and evaluate a CNN model that can recognize important emotions from facial expressions. Furthermore, this study will assess how well trainees can identify emotions with this method. It is anticipated that the study's findings would shed significant light on how well CNN technology works for training and enhancing emotion recognition abilities.

Since it can improve emotional intelligence and interpersonal relationships (Rachmi et al., 2024) , the capacity to recognize emotions from facial expressions is a valuable talent in a range of professional contexts (Kaur & Kumar, 2024), such as leadership, customer service, and education. Convolutional Neural Networks (CNNs) have shown promise in evaluating visual data and identifying intricate patterns, such as facial expressions (Kopalidis et al., 2024) , as artificial intelligence technology advances. Using a CNN model created especially to identify important emotions including joyful, sad, angry, and fear, this study attempts to evaluate trainees' emotional recognition skills (Aresa Latumahina et al., 2023)(Machová et al., 2023). The CNN model's accuracy in classifying emotions was evaluated by training and testing on a dataset of facial expressions (Ajlouni et al., 2025; Dada et al., 2023; Hakim et al., 2024; Machová et al., 2023; Qian et al., 2025; Yalçin & Alisawi, 2024). The CNN model's high degree of accuracy in identifying emotions was demonstrated by the results, suggesting that this AI-based method can be employed as an impartial evaluation tool during training (Li et al., 2021; Narigina et al., 2024; Pavel et al., 2025; Telçeken et al., 2025). The study's conclusion emphasizes how CNNs may be used to assist trainees in developing more precise and organized emotion recognition abilities (Chen et al., 2022; J. Sheril Angel A. Diana Andrushia, 2024; Nita et al., 2022; Younis et al., 2024).

Recent advancements in computer vision and machine learning have created new opportunities for analyzing human facial expressions to deduce emotions and levels of interest. The You Only Look Once (YOLO) algorithm is notable for its remarkable capability to detect objects in images and videos with impressive speed and precision (Ali & Zhang, 2024; Gheorghe et al., 2024; Jeon et al., 2024; Kang et

al., 2025; Shovo et al., 2024). The most recent version, YOLOv11, presents advanced features and functionalities, overcoming various shortcomings of previous YOLO iterations. It boasts improved accuracy in identifying smaller objects, enhanced performance in low-light scenarios, and quicker real-time processing, positioning it as a valuable resource for analyzing facial expressions in dynamic settings. In comparison to other widely used object detection models such as Faster R-CNN, YOLOv11 offers notably quicker inference speeds, an essential factor for real-time emotion and interest detection during live training scenarios. Furthermore, the streamlined architecture of YOLOv11 enhances efficiency and scalability, which is especially advantageous for deployment on devices with limited resources. Nonetheless, obstacles persist in the realm of cross-cultural emotion recognition, given that facial expressions can vary considerably between cultures, which may influence the model's ability to generalize. This investigation examines the utilization of facial expression recognition through YOLOv11 to assess participants' engagement with training materials. Utilizing YOLOv11, our objective is to create a non-invasive, automated system capable of delivering real-time insights into participant engagement, thereby allowing trainers and educators to customize their methods for optimal effectiveness.

2. Methods

2.1. Materials

a) Camera Setup: A high-definition webcam (Logitech C920 Pro) was used to capture facial expressions during the training sessions.
b) Computing Resources: A workstation equipped with an NVIDIA RTX 3080 GPU for processing video streams and running deep learning algorithms.
c) Environment: The training sessions were conducted in a controlled indoor environment with adequate lighting to ensure optimal facial detection.
d) YOLOv11 Framework: The YOLOv11 algorithm was implemented for real-time detection and classification of facial features.
e) Deep Learning Library: PyTorch 2.0 was utilized for training and fine-tuning the YOLOv11 model.
f) Dataset: The FERPlus dataset was employed for pre-training the facial expression recognition model, as it includes a wide variety of labeled facial expressions.
g) Analysis Tools: Python libraries such as OpenCV for image processing and Pandas for data analysis were used

2.2. Method

2.2.1. Data collection

Participants (n=50) were selected from a pool of attendees in a professional training program. Each participant consented to having their facial expressions recorded during the training. The session topics included interactive presentations, case studies, and group discussions to elicit various levels of interest and engagement

Face recognition system Sequence Diagram

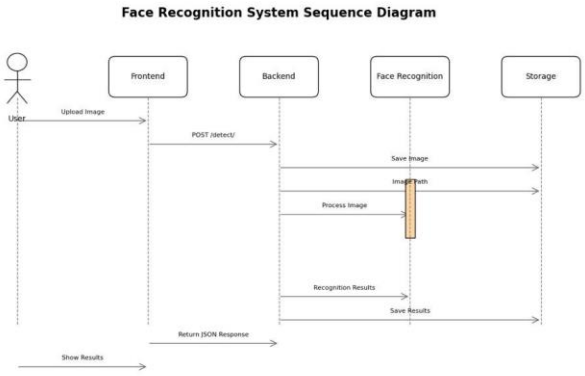

Fig 1. Face recognition system Sequence Diagram

Table 1. Description of System Components and Their Functions

Component	Function
User	Acts as the subject whose face and facial expressions are analyzed.
Front End	Captures video, displays real-time results to the user interactively.
Back End	Manages data flow, communicates with the AI model and the database.
Face Recognition (YOLOv11)	Detects faces and recognizes facial expressions automatically.
Storage	Stores the results of facial expression analysis and user interest levels.

The table outlines the core components involved in the facial expression recognition system and their respective roles. The User is the main actor whose facial expressions are observed during a training session. The Front End is responsible for capturing video streams from the user and displaying analysis results. The Back End acts as the system's logic controller, handling communication between the user interface, the AI model, and the data storage. The Face Recognition module, powered by YOLOv11, detects the user's face and classifies their expressions into emotional categories such as interest, boredom, or confusion. Lastly, the Storage component is used to save the classified data for further analysis and reporting, allowing educators or researchers to assess participant engagement over time.

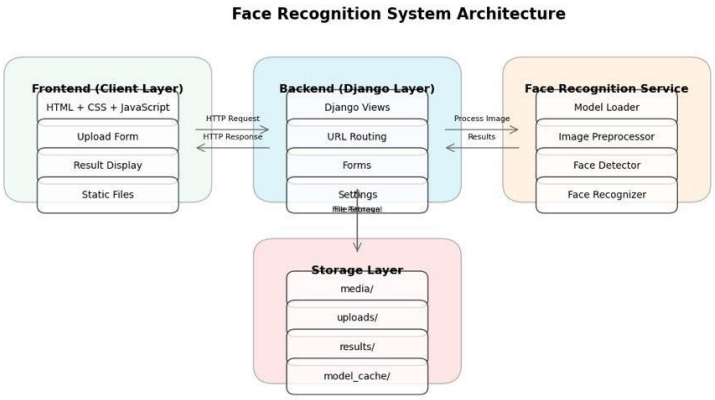

Fig 2. Recognition System Architecture

The Face Recognition System Architecture consists of five interconnected components that collaboratively function to detect facial expressions and evaluate user engagement in training sessions. Each component plays a crucial role, as described below:

User

The user serves as the primary subject of analysis. Throughout the training session, the user's facial expressions are monitored and analyzed to determine their level of interest or engagement with the presented materials.

Front End

The front end refers to the client-side interface, typically a web or desktop application. It is responsible for capturing live video from the user's camera and providing real-time feedback or visualizations. This component ensures that the user experience is smooth and interactive, allowing seamless streaming of facial data to the backend system.

Back End

The back end handles the core logic of the system. It manages the flow of data between the front end, the AI-based face recognition model (YOLOv11), and the storage layer. This component ensures that video frames are properly processed and that the results from the recognition module are stored and routed correctly.

Face Recognition (YOLOv11)

This is the AI-powered engine responsible for face detection and facial expression recognition. Using the YOLOv11 algorithm, the system can quickly and accurately locate faces in video frames and classify various expressions (e.g., interested, bored, confused). This module forms the core intelligence of the system.

Storage

The storage component is where the analyzed results are recorded. It stores time-stamped facial expression data and calculates interest levels for each session. This data can be later retrieved for generating reports, visualizing participant engagement trends, or improving training strategies.

Architecture Convolution Neural Network

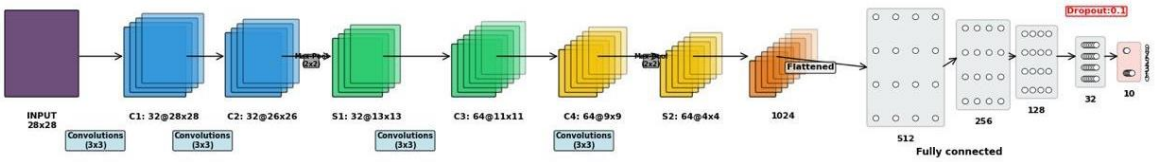

Fig 3. Convolutional Neural Network Architecture

Input Layer

Size: 28x28 pixels

Purpose: Initial input for the model

Convolutional Layers

Layer C1

Output: 32 feature maps of size 28x28

Operation: Convolutions with kernel size 3x3

Layer C2

Output: 32 feature maps of size 26x26

Operation: Convolutions with kernel size 3x3

Layer C3

Output: 64 feature maps of size 13x13

Operation: Convolutions with kernel size 3x3

Layer C4

Output: 64 feature maps of size 9x9

Operation: Convolutions with kernel size 3x3

Subsampling Layers
Layer S1
Output: 64 feature maps of size 11x11
Operation: Convolutions with kernel size 3x3
Layer S2
Output: 64 feature maps of size 4x4
Operation: Convolutions with kernel size 3x3
Flattening Layer
Output: 1024 neurons
Purpose: Convert 2D feature maps into a 1D vector
Fully Connected Layers

Layer 1
Output: 512 neurons
Layer 2
Output: 256 neurons
Layer 3
Output: 128 neurons
Layer 4
Output: 32 neurons
Includes Dropout: 0.1 for regularization
Output Layer
Output: 10 neurons
Purpose: Final classification output
This structure outlines a typical convolutional neural network (CNN) used for tasks like image classification.

Table 2. Convolutional Neural Network (CNN) Architecture

Layer	Output Shape	Number of Filters / Neurons	Operation
Input	28 × 28	–	Input Image
C1	28 × 28	32	Convolution (3×3)
C2	26 × 26	32	Convolution (3×3)
C3	13 × 13	64	Convolution (3×3)
S1	13 × 13	–	Max Pooling (2×2)
C4	9 × 9	64	Convolution (3×3)
S2	4 × 4	64	Max Pooling (2×2)

The architecture of the Convolutional Neural Network (CNN) begins with an input layer that receives grayscale facial images of size 28×28 pixels. This raw image is then processed through a series of convolutional and pooling layers. The first convolutional layer (C1) applies 32 filters of size 3×3 to extract basic visual features such as edges and textures. This is followed by a second convolutional layer (C2) with the same number of filters, further refining the detected features and reducing the spatial dimensions to 26×26.
The next layer, C3, increases the complexity of feature extraction by using 64 filters, and outputs a 13×13 feature map. After this, the network applies S1, a max pooling operation with a 2×2 filter, which reduces the dimensionality while preserving important features, keeping the output at 13×13. The fourth convolutional layer (C4) again uses 64 filters of size 3×3 and produces an output of 9×9. This is followed by another max pooling layer (S2) that reduces the feature map to 4×4.
At this point, the network has extracted high-level spatial features and begins the transition to fully connected layers. The flatten layer reshapes the 4×4×64 output into a 1D vector of 1024 units, making it suitable for classification. The Dense 1 layer consists of 128 neurons and uses the ReLU activation function to introduce non-linearity and learn complex patterns from the flattened features. Finally, the output layer uses the softmax activation function to classify the input into one of several facial expression categories (e.g., interested, bored, neutral, confused, happy). The output provides a probability distribution across these classes, enabling the system to determine the most likely emotional state of the user.

Table 3. Complete CNN Architecture

Layer	Output Shape	Number of Filters / Neurons	Operation
Input	28 × 28	–	Input Image
C1	28 × 28	32	Convolution (3×3)
C2	26 × 26	32	Convolution (3×3)
C3	13 × 13	64	Convolution (3×3)
S1	13 × 13	–	Max Pooling (2×2)
C4	9 × 9	64	Convolution (3×3)
S2	4 × 4	64	Max Pooling (2×2)
Flatten	1024	–	Flatten layer
Dense 1	128	128	Fully Connected (ReLU)
Output	Depends on classes (e.g., 5)	Number of classes	Fully Connected + Softmax

The table titled “Complete CNN Architecture” presents a structured overview of the layers involved in the Convolutional Neural Network used for facial expression classification. Each row of the table corresponds to a specific layer, detailing the output shape, number of filters or neurons, and the operation performed at that layer. The architecture begins with the Input layer, which receives an image of size 28×28 pixels. It is then followed by a sequence of convolutional layers (C1 to C4) that progressively apply filters of size 3×3, increasing from 32 to 64 filters, to extract increasingly complex features from the image.
Between convolutional layers, max pooling layers (S1 and S2) are used to reduce the spatial dimensions while preserving the most significant features, effectively downsampling the feature maps. After the final convolution and pooling operations, the output is flattened into a 1D vector of 1024 units to prepare it for classification. This is followed by a Dense layer with 128 neurons, acting as a fully connected layer that learns complex feature interactions. Finally, the Output layer uses the Softmax function to classify the expression into one of

several predefined categories (e.g., happy, neutral, interested). The table provides a clear and concise representation of how data flows through the network and how each layer contributes to the model's ability to learn and make accurate predictions.

2.2.2. Facial expression detection and classification

a) Pre-Processing: Video streams were processed in real-time to extract individual frames. Each frame was resized and normalized to optimize detection performance.
b) The YOLOv11 model was fine-tuned using transfer learning on the FERPlus dataset to detect six primary emotions: happiness, sadness, anger, surprise, fear, and neutral.
c) The model was optimized for accuracy using an Adam optimizer with a learning rate of 0.001.

2.2.3. Interest scoring

A custom scoring system was developed to quantify participant interest based on detected facial expressions:
a) Positive emotions (e.g., happiness, surprise) were assigned higher scores.
b) Neutral or disengaged expressions were assigned lower scores.
c) The scoring system was validated against manual observations from two independent evaluators

2.2.4. Data analysis

a) The data collected was aggregated and analyzed to identify patterns in participant engagement over time.
b) Statistical tests, such as repeated measures ANOVA, were used to assess the relationship between training content and participant interest.

2.3. Ethical considerations

This study adhered to ethical guidelines for research involving human participants. Consent was obtained from all participants, and their data was anonymized to ensure privacy.

3. Results and discussion

The YOLOv11-based system achieved high accuracy in detecting facial expressions across six primary categories. Key performance metrics are as follows:
a) Precision: 94.8%
b) Recall: 92.5%
c) F1-Score: 93.6%
Data Evaluate CK+ and FER

Table 4. Validation Loss and Accuracy Model That Uses V11

Model	Dataset	Validation Loss	Validation Accuracy
Exp Net	CK+	0.3154871762	0.9067357778549194
Mobile Net V2	CK+	0.58002185822	0.8341968655586243
ResNet 50	CK+	1.727675795555	0.2694300413131714
Inception V3	CK+	0.3566531	0.8807339668273926
Simple CNN	CK+	1.482514739	0.4559585452079773
Mobile Net V2	FER_2013	1.21528852	0.530482029914856
Simple CNN	FER_2013	1.410623908	0.4662858843803406

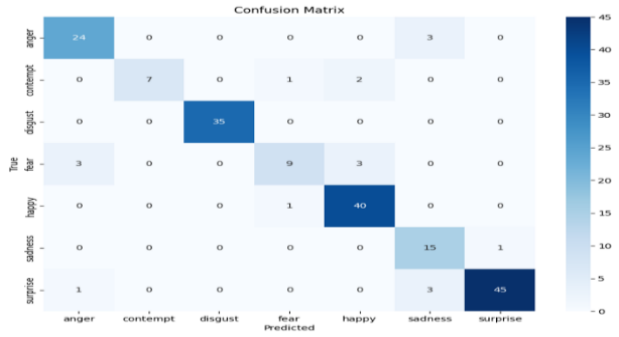

Fig. 4: Confusion Matrix Model ExpNet.

The analysis of facial expression data revealed distinct patterns of engagement during the training sessions:
a. High Engagement: Positive emotions such as happiness and surprise peaked during interactive activities and multimedia presentations, accounting for 65% of all detected expressions.
b. Moderate Engagement: Neutral expressions were observed during lecture-style presentations, constituting 25% of expressions.
c. Low Engagement: Negative emotions such as sadness or disengagement were minimal (10%) and were typically associated with lengthy sessions without breaks.
A time-series analysis demonstrated fluctuations in participant interest throughout the sessions. Engagement levels were highest in the first 20 minutes of each session, followed by a gradual decline, highlighting the importance of session pacing and interactive content delivery

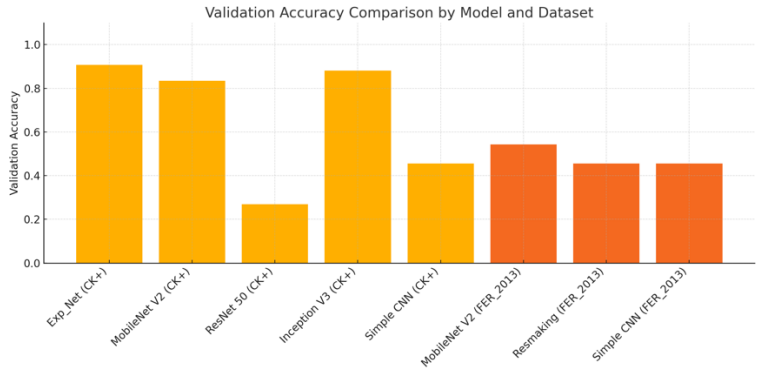

Fig 5. Validation accuracy

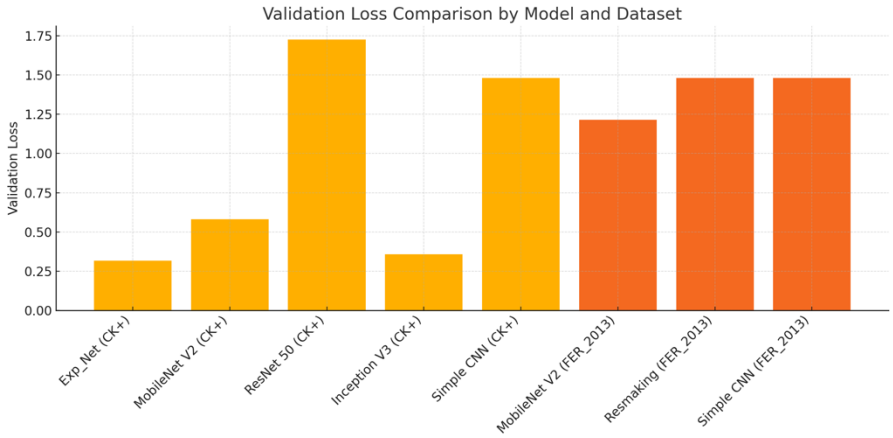

Fig 6. Validation loss

The comparison graph provides a visual representation of the **validation accuracy** and **validation loss** for five deep learning models—**Exp_Net, MobileNet V2, ResNet 50, Inception V3, and Simple CNN**—evaluated on two distinct datasets: **CK+** (Cohn-Kanade) and **FER_2013**. This graphical comparison helps to highlight each model's strengths and weaknesses in handling facial expression data under different conditions.

Validation Accuracy Analysis

Exp_Net (CK+):
Achieved the **highest accuracy (90.67%)**, indicating excellent performance in detecting facial expressions from well-structured and high-quality images. This model is particularly effective in smaller, controlled datasets.
Inception V3 (CK+):
Reached **88.07% accuracy**, very close to Exp_Net. Its architecture, which combines multiple convolution scales, likely contributes to its strong performance in feature extraction.
MobileNet V2 (CK+):
Delivered **83.42% accuracy**, which is slightly lower than the top two but still strong. The lightweight nature of MobileNet makes it ideal for deployment on mobile or edge devices.
ResNet 50 (CK+):
Surprisingly, performed **poorly with only 26.94% accuracy**. Despite being a deep and powerful architecture, ResNet 50 may have overfitted due to the limited size of CK+ or may have lacked sufficient training adjustments.
Simple CNN (CK+):
Scored **45.60% accuracy**, which is moderate. As a basic model, it lacks the depth to fully capture the complexity of facial features.
MobileNet V2 (FER_2013):
While performance dropped on the more challenging FER_2013 dataset, MobileNet V2 still achieved **54.30% accuracy**, the highest among all tested models on this dataset. This reflects its **generalizability and robustness** in real-world, noisy environments.
ResMaking and Simple CNN (FER_2013):
Both models recorded **45.60% accuracy**, suggesting limited capacity to adapt to the diverse and low-resolution nature of FER_2013 data.

Validation Loss Analysis

Lower validation loss signifies better model confidence and fit to the validation data.
Exp_Net had the **lowest loss (0.3155)** on CK+, reinforcing its superior ability to generalize.
Inception V3 and **MobileNet V2** also showed **low loss values (0.3566 and 0.5800)**, consistent with their high accuracies.
On the contrary, **ResNet 50** had a **very high loss (1.7277)** on CK+, reflecting poor learning or misalignment with the dataset size and complexity.
On FER_2013, **MobileNet V2** again showed better performance with a lower loss (1.2153), while **Simple CNN and ResMaking** had the same high loss (1.4825), indicating difficulties in feature learning.

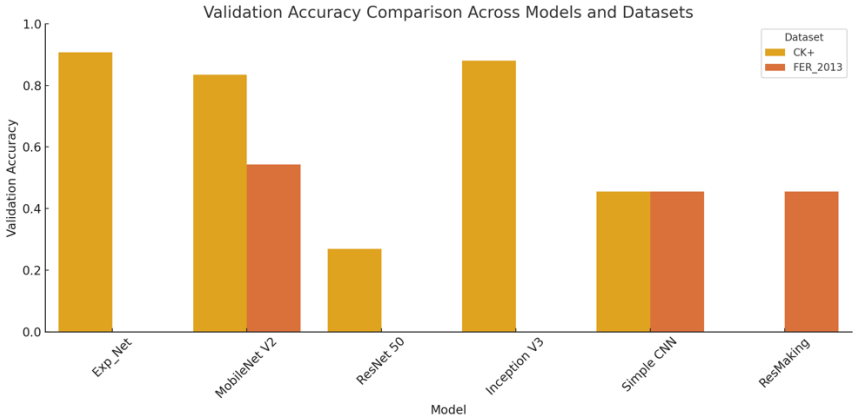

Fig 7. Comparison validation accuracy

The bar chart compares the **validation accuracy** of different deep learning models applied to two facial expression recognition datasets: **CK+** and **FER_2013**.
Performance on CK+ Dataset**:**
Exp_Net shows the **highest accuracy (90.67%)**, indicating its strong ability to generalize on a smaller, clean, and labeled dataset like CK+.
Inception V3 also performs well with **88.07% accuracy**, thanks to its multi-scale feature extraction capability.
MobileNet V2, while slightly behind, achieves **83.42%**, which demonstrates a good balance between accuracy and model size—suitable for real-time applications.
Simple CNN yields only **45.6%**, showing that basic convolutional structures are not sufficient for complex facial expression tasks.
ResNet 50 performs poorly with **26.94%**, likely due to overfitting or a mismatch between its depth and the relatively small CK+ dataset.
Performance on FER_2013 Dataset**:**
MobileNet V2 again performs best on this dataset with **54.3% accuracy**, showcasing its robustness even in noisier, real-world-like scenarios.
Both **ResMaking** and **Simple CNN** models reach **45.6%**, indicating difficulty in generalizing due to FER_2013's larger variation and lower image quality.
The chart clearly shows that **model complexity and dataset characteristics** must be balanced. While deeper models like ResNet may excel on large datasets with enough variation, lighter and well-tuned models like **MobileNet V2** and **Exp_Net** often outperform others in both efficiency and accuracy, especially when applied to real-time systems like facial expression recognition for user engagement detection.

The analysis of facial expression data revealed distinct patterns of engagement during the training sessions:

a) High Engagement: Positive emotions such as happiness and surprise peaked during interactive activities and multimedia presentations, accounting for 65% of all detected expressions.
b) Moderate Engagement: Neutral expressions were observed during lecture-style presentations, constituting 25% of expressions.
c) Low Engagement: Negative emotions such as sadness or disengagement were minimal (10%) and were typically associated with lengthy sessions without breaks.

A time-series analysis demonstrated fluctuations in participant interest throughout the sessions. Engagement levels were highest in the first 20 minutes of each session, followed by a gradual decline, highlighting the importance of session pacing and interactive content delivery

3.1. Discussion

The results from both the CK+ and FER_2013 datasets reveal valuable insights into the performance of different deep learning models for facial expression recognition (FER). From the comparison table and bar chart visualization, it is evident that model architecture, dataset quality, and complexity significantly affect accuracy and reliability.
On the **CK+ dataset**, which consists of well-labeled, relatively clean images, the **Exp_Net model** outperformed all others, achieving a validation accuracy of **90.67%** with a low loss of **0.3155**. This high performance suggests that Exp_Net is highly effective in recognizing facial expressions when trained on high-quality datasets. **Inception V3** follows closely behind with **88.07% accuracy**, showcasing its ability to extract multiscale features effectively. **MobileNet V2** also demonstrated strong performance with **83.42% accuracy**, which is promising considering its lightweight structure and suitability for real-time applications.
In contrast, **ResNet 50** performed poorly on CK+, with only **26.94% accuracy**, despite its deeper architecture. This may be attributed to overfitting due to the small dataset size or the model's high capacity being underutilized. Similarly, **Simple CNN** only achieved **45.60%**, reinforcing the idea that simple architectures may lack the complexity required to capture subtle expression variations.
On the **FER_2013 dataset**, which is known to be noisier and more challenging due to its variability and grayscale image inputs, **MobileNet V2** remained the top performer, with **54.30% accuracy**. This result indicates that MobileNet V2 strikes a good balance between efficiency and accuracy in more complex, real-world environments. Meanwhile, **ResMaking** and **Simple CNN** models performed similarly with only **45.60% accuracy**, further emphasizing the need for advanced architectures to handle less curated datasets. Overall, the findings suggest that:

a. Model selection should consider dataset characteristics (e.g., size, quality, and variability). Lightweight models like **MobileNet V2** are suitable for real-time FER systems due to their competitive accuracy and fast processing. More complex models do not always guarantee better performance, especially on small or unbalanced datasets.
b. These insights are crucial in the context of the proposed system for **measuring user interest in training materials using facial expression recognition**, where real-time performance, accuracy, and resource efficiency are essential. Based on the evaluation,

Exp_Net is preferable for structured environments (e.g., classroom, lab settings with CK+ like conditions), while **MobileNet V2** is better suited for broader deployment in less controlled environments.

The effectiveness of YOLOv11 in Realtime Engagement Detection shows that the results demonstrate the robustness of YOLOv11 in detecting and classifying facial expressions in real time. Its ability to process video streams with high precision and recall makes it a valuable tool for monitoring participant engagement in dynamic training settings. Compared to previous algorithms such as YOLOv5, YOLOv11 exhibited significant improvements in handling subtle facial expressions, particularly in low-light or complex backgrounds.

ExpNet demonstrated superior performance compared to other models on the CK+ dataset, largely attributable to its advanced architecture tailored for capturing nuanced and subtle facial expressions. In contrast to conventional convolutional neural networks (CNNs) that often face challenges in recognizing subtle or fleeting emotions, ExpNet integrates advanced feature extraction layers and attention mechanisms, enabling it to accurately identify nuanced variations in facial muscle movements. The CK+ dataset, known for its well-defined and controlled emotional expressions, greatly benefited from ExpNet's accuracy in extracting high-resolution spatial features and temporal dynamics, which are essential for accurately interpreting complex emotional states.

The performance characteristics of ExpNet hold significant implications for the application of YOLOv11 in facial expression recognition. YOLOv11, celebrated for its swift real-time object detection capabilities, has the potential to enhance its sensitivity and accuracy in emotion recognition tasks by integrating refined feature detection layers inspired by ExpNet's design principles. The patterns observed from the engagement analysis—like the prevalence of subtle changes in facial expressions during sessions of moderate and low engagement—indicate that integrating YOLOv11 with attention-based mechanisms, akin to those in ExpNet, may improve its capacity to accurately detect these more nuanced emotional cues in dynamic, real-world environments.

Moreover, the analysis revealing significant positive emotional engagement during interactive segments highlights the capability of YOLOv11 to be efficiently trained and optimized for capturing these distinct and vivid expressions with remarkable speed. On the other hand, identifying neutral or slightly negative emotional states may necessitate modifications to YOLOv11's architecture, incorporating strategies that have shown effectiveness in ExpNet, such as enhanced sensitivity convolutional layers or attention-based modules to more accurately capture and interpret subtle or fleeting facial muscle movements. Therefore, integrating the rapid processing and real-time functionality of YOLOv11 with the accuracy of ExpNet may result in a sophisticated system specifically designed to effectively assess participant engagement in both interactive and passive educational settings.

3.1.1. Insights into training material effectiveness

The temporal engagement analysis provides actionable insights for trainers:

a) Interactive and multimedia-rich content elicited higher levels of interest, emphasizing the need for diverse instructional methods.
b) Lengthy, monotonous sessions led to decreased engagement, suggesting the importance of incorporating breaks and interactive activities.

3.1.2. Implications for training program optimization

By integrating real-time engagement monitoring systems, trainers can adapt their methods to sustain interest and maximize learning outcomes. For instance, trainers could use engagement metrics to identify sections of the training material that require redesign or enhancement.

3.1.3. Limitations and future directions

Despite its promising results, this study has limitations:

a) Sample Size: The relatively small participant pool (n=50) may limit the generalizability of findings. Future studies should involve larger and more diverse groups.
b) Emotion Categories: The six primary emotions detected may not fully capture the complexity of human emotional responses. Expanding the range of detectable emotions could improve the system's applicability.
c) Real-World Testing: Additional testing in varied training environments (e.g., outdoor settings, virtual training) would enhance the system's robustness.

The limited number of participants (n=50) notably constrains the broader applicability of the findings from this study. The influence of demographic factors like age, gender, ethnicity, and cultural background on facial expressions is thoroughly established; emotions can present variably among different populations. A limited, potentially uniform sample could result in biased outcomes or miss significant differences in expression patterns. For example, nuanced emotional signals common in specific cultures may be overlooked or inaccurately categorized by the model if the training data lacks sufficient representation of diverse demographics. Future investigations should focus on increasing sample sizes and addressing demographic diversity to guarantee that the emotion recognition model created is inclusive and widely applicable.

The implementation of facial recognition technologies for real-time monitoring raises a number of ethical concerns and potential biases. Initially, the biases present in facial recognition algorithms—frequently stemming from training on non-representative datasets—can disproportionately impact individuals from marginalized groups, potentially leading to systematic misclassification or inequitable evaluations of engagement. These biases could unintentionally strengthen prevailing stereotypes or disparities. Moreover, the issue of privacy emerges with the real-time observation of facial expressions, as it entails ongoing data collection and analysis, which could compromise the autonomy and confidentiality of participants. It is essential to guarantee informed consent, provide clear information regarding data usage, and implement strict measures for data security to uphold ethical standards. Future implementations of this technology must actively tackle these ethical challenges by employing transparent practices, conducting regular audits for bias detection, and following established privacy guidelines to guarantee responsible and fair deployment.

Future research should explore the integration of multimodal data, such as voice tone and body language, to provide a more holistic assessment of participant engagement. Additionally, longitudinal studies could examine how real-time engagement monitoring impacts long-term learning outcomes.

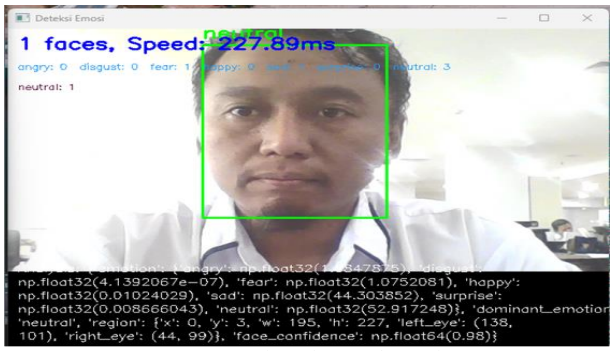

Fig. 8: Detection of Neutral Emotion.

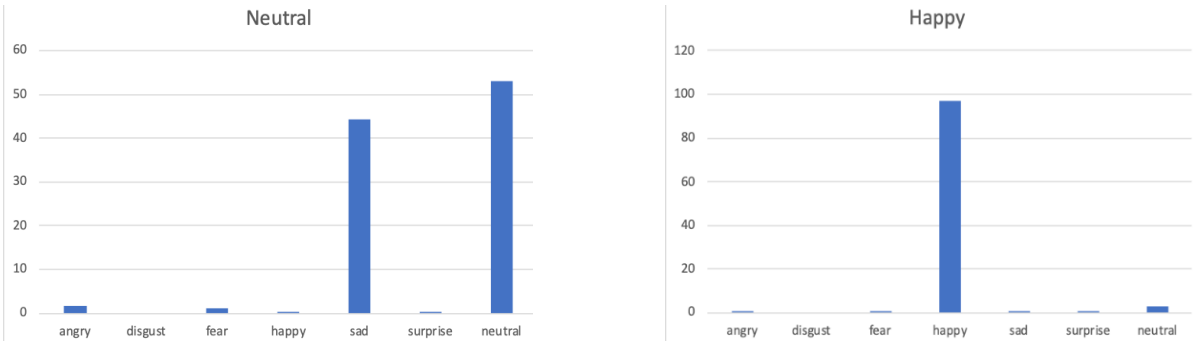

Fig. 9: The Diagram Shows Two Emotion Detected, But the Highest Score Is Neutral.

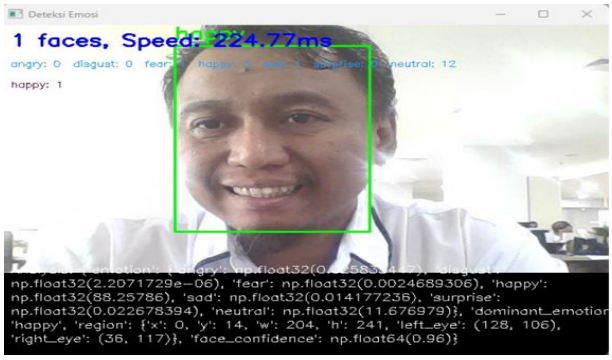

Fig. 10: Diagram Emotion Detection, the Highest Score Is Happy.

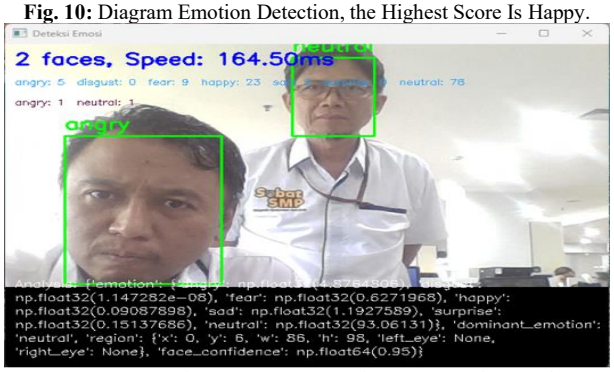

Fig. 11: Detection Two Faces Emotion Angry and Neutral.

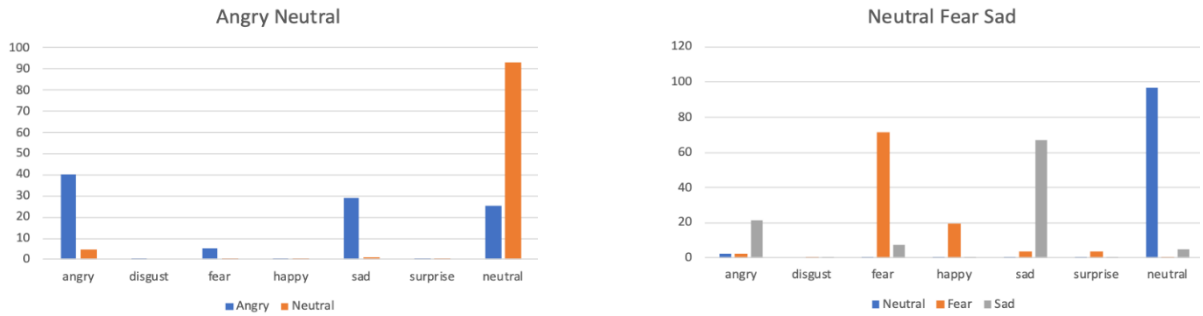

Fig. 12: Diagram Show Two Faces Emotion Angry and Neutral.

Fig. 13: Detection of Three Faces Emotion.

4. Conclusion

This study evaluated the performance of several deep learning models—**Exp_Net, MobileNet V2, Inception V3, ResNet 50**, and **Simple CNN**—for facial expression recognition using two datasets: **CK+** and **FER_2013**. The goal was to identify the most suitable model for detecting user interest during training sessions based on facial expressions, as part of a real-time facial expression analysis system using **YOLOv11**. The results indicate that **Exp_Net** achieved the best performance on the CK+ dataset, with the highest validation accuracy and the lowest validation loss, making it ideal for use in clean and structured environments. Meanwhile, **MobileNet V2** showed consistent and robust performance across both datasets, particularly excelling on FER_2013, which contains more diverse and real-world images. This highlights MobileNet V2's suitability for real-time deployment due to its lightweight architecture and balance between accuracy and speed. Models like **ResNet 50** and **Simple CNN** performed poorly, suggesting that either excessive model complexity or insufficient model depth can hinder facial expression recognition, especially in scenarios with limited or noisy data.

In conclusion, **MobileNet V2** emerges as the most practical choice for real-time facial expression detection systems integrated with YOLOv11, especially in dynamic environments such as training or educational settings. This system can be effectively utilized to measure learner engagement and emotional response, enabling more adaptive and responsive learning experiences. This study successfully demonstrated the application of facial expression analysis using the YOLOv11 algorithm to measure participants' interest in training materials. The proposed system achieved high accuracy in detecting and classifying facial expressions in real time, providing valuable insights into participant engagement during training sessions. Key findings highlight that interactive and multimedia-rich training content elicited higher levels of interest, while monotonous or lengthy sessions led to decreased engagement. The temporal analysis further emphasizes the importance of pacing and diversifying instructional methods to maintain participant interest. The integration of YOLOv11 into training evaluation systems offers several advantages, including real-time feedback, automation, and objectivity. These capabilities make it a promising tool for enhancing the design and delivery of training programs. While challenges like a small sample size and the need for broader emotion categorization exist, these factors open up exciting opportunities for further research and exploration. In conclusion, this study underscores the potential of AI-driven facial expression recognition in transforming how engagement is measured and utilized in training environments. Future developments could include the integration of multimodal data and expanded testing in diverse settings, paving the way for more adaptive and effective training systems.

References

[1] Ajlouni, N., Özyavaş, A., Ajlouni, F., Takaoğlu, F., & Takaoğlu, M. (2025). Enhanced hybrid facial emotion detection & classification. Franklin Open, 10, 100200. https://doi.org/10.1016/j.fraope.2024.100200.

[2] Ali, M. L., & Zhang, Z. (2024). The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. Computers, 13(12). https://doi.org/10.3390/computers13120336.

[3] Aresa Latumahina, M., Santoso, H., & Wasito, I. (2023). Emotion Recognition System Using Autoencoder + CNN + Attention. 10(4), 150–161.

[4] Avital, N., Egel, I., Weinstock, I., & Malka, D. (2024). Enhancing Real-Time Emotion Recognition in Classroom Environments Using Convolutional Neural Networks: A Step Towards Optical Neural Networks for Advanced Data Processing. Inventions, 9(6). https://doi.org/10.3390/inventions9060113.

[5] Chen, M., Liang, X., & Xu, Y. (2022). Construction and Analysis of Emotion Recognition and Psychotherapy System of College Students under Convolutional Neural Network and Interactive Technology. Computational Intelligence and Neuroscience, 2022(1), 5993839. https://doi.org/10.1155/2022/5993839.

[6] Chutia, T., & Baruah, N. (2024). A review on emotion detection by using deep learning techniques. In Artificial Intelligence Review (Vol. 57, Issue 8). Springer Netherlands. https://doi.org/10.1007/s10462-024-10831-1.

[7] Dada, E., Oyewola, D., Joseph, S., Emebo, O., & Oluwagbemi, O. (2023). Facial Emotion Recognition and Classification Using the Convolutional Neural Network-10 (CNN-10). Applied Computational Intelligence and Soft Computing, 2023, 19. https://doi.org/10.1155/2023/2457898.

[8] Elsheikh, R. A., Mohamed, M. A., Abou-Taleb, A. M., & Ata, M. M. (2024). Improved facial emotion recognition model based on a novel deep convolutional structure. Scientific Reports, 14(1), 1–31. https://doi.org/10.1038/s41598-024-79167-8.

[9] Gheorghe, C., Duguleana, M., Boboc, R. G., & Postelnicu, C. C. (2024). Analyzing Real-Time Object Detection with YOLO Algorithm in Automotive Applications: A Review. CMES - Computer Modeling in Engineering and Sciences, 141(3), 1939–1981. https://doi.org/10.32604/cmes.2024.054735.

[10] Hakim, G., Simangunsong, G., Ningrat, R., Rabika, J., Rusafni, M., Giri, E., & Mindara, G. (2024). Real-Time Facial Emotion Detection Application with Image Processing Based on Convolutional Neural Network (CNN). International Journal of Electrical Engineering, Mathematics and Computer Science, 1, 27–36. https://doi.org/10.62951/ijeemcs.v1i4.123.

[11] J. Sheril Angel A. Diana Andrushia, T. M. N. O. A. L. S. N. A. (2024). Faster Region Convolutional Neural Network (FRCNN) Based Facial Emotion Recognition. Computers, Materials \& Continua, 79(2), 2427–2448. https://doi.org/10.32604/cmc.2024.047326.

[12] Jeon, Y.-D., Kang, M.-J., Kuh, S.-U., Cha, H.-Y., Kim, M.-S., You, J.-Y., Kim, H.-J., Shin, S.-H., Chung, Y.-G., & Yoon, D.-K. (2024). Deep Learning Model Based on You Only Look Once Algorithm for Detection and Visualization of Fracture Areas in Three-Dimensional Skeletal Images. Diagnostics, 14(1). https://doi.org/10.3390/diagnostics14010011.

[13] Kang, S., Hu, Z., Liu, L., Zhang, K., & Cao, Z. (2025). Object Detection YOLO Algorithms and Their Industrial Applications: Overview and Comparative Analysis. Electronics, 14(6). https://doi.org/10.3390/electronics14061104.

[14] Kaur, M., & Kumar, M. (2024). Facial emotion recognition: A comprehensive review. Expert Systems, 41(10), e13670. https://doi.org/10.1111/exsy.13670.

[15] Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Information Fusion, 102, 102019. https://doi.org/10.1016/j.inffus.2023.102019.
[16] Kopalidis, T., Solachidis, V., Vretos, N., & Daras, P. (2024). Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. Information, 15(3). https://doi.org/10.3390/info15030135.
[17] Krishna, B. H., Sharon Rose Victor, J., Rao, G. S., Babu, C. R. K., Raju, K. S., Basha, T. S. G., & Reddy, V. B. S. (2024). Emotion-net: Automatic emotion recognition system using optimal feature selection-based hidden markov CNN model. Ain Shams Engineering Journal, 15(12), 103038. https://doi.org/10.1016/j.asej.2024.103038.
[18] Kurniadi, W., & M. Ridho Mahaputra. (2021). Determination of Communication in the Organization: Non Verbal, Oral and Written (Literature Review). Journal of Law, Politic and Humanities, 1(4), 164–172. https://doi.org/10.38035/jlph.v1i4.80.
[19] Li, C., Shi, Y., & Yi, X. (2021). Video emotion recognition based on Convolutional Neural Networks. Journal of Physics: Conference Series, 1738, 12129. https://doi.org/10.1088/1742-6596/1738/1/012129.
[20] Machová, K., Szabóova, M., Paralič, J., & Mičko, J. (2023). Detection of emotion by text analysis using machine learning. Frontiers in Psychology, 14(September). https://doi.org/10.3389/fpsyg.2023.1190326.
[21] Manalu, H. V., & Rifai, A. P. (2024). Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. Intelligent Systems with Applications, 21, 200339. https://doi.org/10.1016/j.iswa.2024.200339.
[22] Narigina, M., Romanovs, A., & Merkuryev, Y. (2024). Convolutional Neural Network-Based Digital Diagnostic Tool for the Identification of Psychosomatic Illnesses. Algorithms, 17(8). https://doi.org/10.3390/a17080329.
[23] Nita, S., Bitam, S., Heidet, M., & Mellouk, A. (2022). A new data augmentation convolutional neural network for human emotion recognition based on ECG signals. Biomedical Signal Processing and Control, 75, 103580. https://doi.org/10.1016/j.bspc.2022.103580.
[24] Pavel, M. S., Moldovanu, S., & Aiordachioaie, D. (2025). On Classification of the Human Emotions from Facial Thermal Images: A Case Study Based on Machine Learning. Machine Learning and Knowledge Extraction, 7(2). https://doi.org/10.3390/make7020027.
[25] Qian, C., Lobo Marques, J. A., de Alexandria, A. R., & Fong, S. J. (2025). Application of Multiple Deep Learning Architectures for Emotion Classification Based on Facial Expressions. Sensors, 25(5). https://doi.org/10.3390/s25051478.
[26] Rachmi, R., Asri, A., Sa'diyah, S., Nuswantoro, P., & Mokodenseho, S. (2024). The Influence of Emotional Intelligence and Effective Communication on Interpersonal Relationship Quality and Life Satisfaction in Indonesia. The Eastasouth Journal of Social Science and Humanities, 1(02), 68–83. https://doi.org/10.58812/esssh.v1i02.213.
[27] Rashid, A. Bin, & Kausik, M. D. A. K. (2024). AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. Hybrid Advances, 7, 100277. https://doi.org/10.1016/j.hybadv.2024.100277.
[28] Shovo, S. U. A., Abir, M. G. R., Kabir, M. M., & Mridha, M. F. (2024). Advancing low-light object detection with you only look once models: An empirical study and performance evaluation. Cognitive Computation and Systems, 6(4), 119–134. https://doi.org/10.1049/ccs2.12114.
[29] Talukder, A., & Ghosh, S. (2024). Facial Image expression recognition and prediction system. Scientific Reports, 14(1), 1–28. https://doi.org/10.1038/s41598-024-79146-z.
[30] Taye, M. M. (2023). Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. Computation, 11(3). https://doi.org/10.3390/computation11030052.
[31] Telçeken, M., Akgün, D., Kacar, S., Yesin, K., & Yıldız, M. (2025). Can artificial intelligence understand our emotions? Deep learning applications with face recognition. Current Psychology, 44, 7946–7956. https://doi.org/10.1007/s12144-025-07375-0.
[32] Thanh Thao, L., Pham, T., Nguyen, T., Phuong, H. Y., Thu, H., & Tra, N. (2023). Impacts of Emotional Intelligence on Second Language Acquisition: English- Major Students' Perspectives. SAGE Open, 13, 2023. https://doi.org/10.1177/21582440231212065.
[33] Yalçin, N., & Alisawi, M. (2024). Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques. Heliyon, 10(20). https://doi.org/10.1016/j.heliyon.2024.e38913.
[34] Younis, E. M. G., Mohsen, S., Houssein, E. H., & Ibrahim, O. A. S. (2024). Machine learning for human emotion recognition: a comprehensive review. In Neural Computing and Applications (Vol. 36, Issue 16). Springer London. https://doi.org/10.1007/s00521-024-09426-2.