

Comparison and Implementation of CNN Facial Emotion Recognition Model with Hyperparameter Analysis on Multiple Datasets

Xaviera Valentina Tandianto^{1*}, Dwi Hosanna Bangkalang², Nina Setiyawati³

^{1,2}Department of Information System, Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Central Java, Indonesia

³Department of Informatics Engineering, Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Central Java, Indonesia

E-mail: ^{1*}682022005@student.uksw.edu, ²dwi hosanna.bangkalang@uksw.edu, ³nina.setiyawati@uksw.edu

(Received: 20 Jan 2026, revised: 25 Feb 2026, accepted: 27 Feb 2026)

Abstract

This study presents a systematic comparison and implementation of a Convolutional Neural Network (CNN) for Facial Emotion Recognition (FER) across multiple public datasets, namely FER-2013, FER+, RAF-DB, and AffectNet. Unlike previous studies that focused on a single dataset or different model architectures, The main contributions of this research consist of three aspects. First, a five-layer integrated CNN architecture is used to enable fair cross-dataset evaluation within a consistent training and testing framework. Second, structured hyperparameter tuning is performed, including variations in learning rate, batch size, filter configuration, and dropout rate, resulting in a stable and reproducible model configuration. Third, an in-depth analysis was conducted to explore the impact of annotation quality and dataset complexity on model performance. The experimental results show that FER+ achieved the highest accuracy and weighted F1 score thanks to better label consistency, followed by RAF-DB, while FER-2013 and AffectNet experienced a decline in performance due to label noise and higher pose and lighting variations. Further confusion matrix analysis shows that happy and neutral expressions are classified more reliably, while negative emotions such as anger, fear, and disgust remain challenging. To validate practical application, the best-performing model was implemented in a webcam-based facial expression recognition prototype using Python and OpenCV, demonstrating reliable frame-level emotion inference under controlled real-time conditions.

Keywords: Convolutional Neural Network, Facial Emotion Recognition, Hyperparameter Analysis, Webcam Implementation

I. INTRODUCTION

Facial emotion recognition (FER) plays an important role in human-computer interaction by enabling intelligent systems to recognize and respond to human emotional states. In applications such as social robots, healthcare assistants, and interactive systems, the ability to recognize facial expressions allows machines to adapt their responses more naturally and contextually. Without reliable FER mechanisms, human-machine interaction remains rigid and lacks emotional awareness, limiting its effectiveness in real-world scenarios [1].

Facial expressions are one of the most important channels of nonverbal communication [2]. Classic studies by Ekman and Friesen identified six basic emotions, namely happiness, sadness, fear, disgust, surprise, and anger, that are considered universal across cultures. However, recent studies show that

facial expression recognition performance is influenced by additional factors such as age, gender, ethnicity, and cultural context [3]. In addition, real-world conditions present significant challenges, including variations in pose, lighting, occlusion, and spontaneous expressions, which are often not present in controlled datasets.

Convolutional Neural Networks (CNNs) have demonstrated strong capability in addressing these challenges by automatically learning hierarchical feature representations from facial images [4]. Despite this progress, existing FER studies often evaluate models on a single dataset or use different architectures and training strategies, making it difficult to objectively analyze the influence of dataset characteristics on model performance. Furthermore, many studies emphasize architectural complexity while providing limited analysis of how key hyperparameters and annotation quality affect generalization across datasets.

To overcome these limitations, this study examines the performance of integrated CNN architectures on various public FER datasets with diverse characteristics, namely FER-2013, FER+, RAF-DB, and AffectNet. The main contributions of this study lies in (1) conducting a systematic hyperparameter analysis using a single CNN architecture, (2) evaluating the impact of annotation quality and dataset complexity on recognition performance, and (3) validating the best-performing model through deployment in a webcam-based FER prototype. This approach allows for a fair and critical comparison between datasets while demonstrating the practical feasibility of the proposed model.

II. RESEARCH METHODOLOGY

This study compares the performance of Convolutional Neural Network (CNN) models on several dataset and hyperparameter configurations. Figure 1 illustrates the overall flow of this research, which begins with the collection of public datasets (FER2013, FER+, RAFDB, and Affectnet), followed by data preprocessing and simple data augmentation, hyperparameter tuning, CNN model training on each dataset with the best configuration obtained, and model evaluation and implementation on a webcam-based facial emotion recognition system.

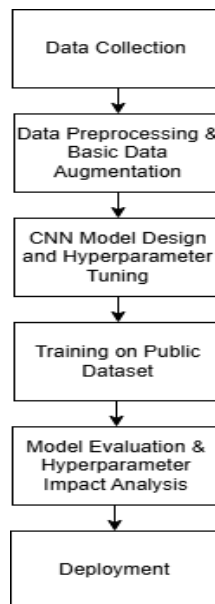


Figure 1. Research Flow

A. Data Description

The research process began with the collection of datasets to be used, namely FER-2013, FER+, RAF-DB, and AffectNet (Table 1). These datasets were selected based on the diversity of data characteristics in each dataset so that they could support a more in-depth analysis of model performance. FER-2013 was selected because it is a standard dataset that is often used as an evaluation standard in facial recognition research [5][6]. FER+ is a re-annotated version of FER-2013 with better quality so that it can be used as a

comparison [7]. RAF-DB was used because it presents color (RGB) facial images with variations in pose and lighting in accordance with real conditions [8][9]. Meanwhile, AffectNet is one of the largest datasets with a wide variety of expressions, so this dataset can be used to provide an overview of model performance in large-scale data scenarios [9].

Table 1. Comparison of FER Datasets Used in Research

	FER-2013	FER+	RAF-DB	Affect-Net
Image Format	Grayscale	Grayscale	RGB	RGB
Image Size	48x48	48x48	100x100	Variation
Number of Emotion Classes	7 classes: <i>anger, disgust, fear, happiness, sadness, surprise, neutral</i>	8 classes: <i>anger, disgust, fear, happiness, sadness, surprise, neutral, contempt</i>	7 classes: <i>anger, disgust, fear, happiness, sadness, surprise, neutral</i>	8 classes: <i>anger, disgust, fear, happiness, sadness, surprise, neutral, contempt</i>
Number of Images	35,887 images	35,485 images	15,339 images	30,626 images

By combining these four datasets, this study covers a wide spectrum of FER data characteristics, ranging from controlled grayscale images to complex real-world RGB images, thereby enabling a fair and systematic evaluation of CNN performance across heterogeneous datasets.

B. Data Preprocessing and Basic Data Augmentation

At this stage, all data undergoes consistent preprocessing and is also adjusted to the characteristics of each dataset. In general, this preprocessing stage involves image size adjustment, color conversion, pixel value normalization, data division, and basic data augmentation on the training data from each dataset.

1. Preprocessing on the FER-2013 and FER+ datasets

The public dataset already provides cropped face images, so there is no need for additional face detection. At the beginning of the preprocessing stage, the image size for the FER-2013 and FER+ datasets is adjusted to 48x48 in accordance with the original resolution of the dataset [10]. This size is also used as an initial reference because it is commonly used in various FER studies with CNN. In addition, no color conversion is performed, so grayscale image format is used in training with the FER-2013 and FER+ datasets. Then, all data is normalized to the range [0,1] to ensure training process stability and improve model performance consistently.

2. Preprocessing on the RAF-DB and AffectNet datasets

Unlike the FER-2013 and FER+ datasets, RAF-DB and AffectNet provide colored (RGB) facial images with more complex variations in pose and illumination. Therefore, the RGB format is preserved during training to maintain the original data characteristics. While RAF-DB provides images with a fixed resolution of 100×100 pixels and AffectNet contains images with varying resolutions due to web-based collection, all facial images are resized to 48×48 pixels in this study. This resizing is performed to standardize the input resolution across all datasets, enabling the use of a consistent CNN architecture and reducing computational complexity.

Although resizing RGB images from RAF-DB and AffectNet to 48×48 pixels may result in the loss of fine-grained facial details, this resolution is intentionally selected to ensure fair cross-dataset comparison and architectural consistency. Previous FER studies have demonstrated that discriminative emotional features can still be effectively captured at lower resolutions using CNN-based models. Moreover, the use of a uniform input size allows observed performance differences to be primarily attributed to dataset characteristics rather than variations in input resolution or network design.

3. Data division and basic augmentation

This study follows the provided data separation, which is to use the entire train section as training and validation data, while the dataset's default test folder is used as test data. In the train section, 85% of the data is used as training data and 15% as validation data. Then, all data is normalized to the range [0,1] to ensure the stability of the research process and improve model performance consistently.

Basic augmentation such as random rotation up to 10°, horizontal flip, and small width and height shifts were applied to the training data using ImageDataGenerator. This augmentation technique aims to increase the diversity of training images without changing the main characteristics of the face, while also helping the CNN model to be more robust against minor variations in the data. No additional augmentation was performed on the validation and test data in order to represent the model's performance with the original data without modification.

C. CNN Model Architecture and Hyperparameters

To ensure fair and controlled cross-dataset evaluation, a single CNN architecture and uniform hyperparameter configuration were consistently applied across all experiments. This design choice is in line with the goal of isolating the influence of dataset characteristics on model performance, rather than introducing variability from architecture modifications or pre-trained representations. The detailed hyperparameter settings used to implement this unified CNN architecture are summarized in Table 2.

Table 2 presents the hyperparameters used in developing the CNN model, including input size, image format, hidden units, number of convolutional layers, dropout rate, batch size, learning rate, and number of training epochs. These parameters were selected to balance model capacity, training

stability, and computational efficiency while maintaining architectural consistency across all datasets.

Table 2. CNN Model Hyperparameter

Hyperparameter	Value
Input Size	48x48
Image Format	Grayscale/RGB
Hidden Units	64-512
Number of Convolution Layers	5 convolutional blocks + 1 fully connected
Dropout Rate	0.3
Batch Size	64
Learning Rate	0.0001 (with exponential decay)
Number of Training Epochs	100

The CNN model architecture consists of five multi-level convolution blocks with a number of filters ranging from 64 to 512. Each block consists of several Conv2D layers with ReLU activation, BatchNormalization, MaxPooling, and Dropout of 0.3 to prevent overfitting [11][12]. This hierarchical structure enables progressive feature abstraction while maintaining a lightweight design suitable for cross-dataset analysis. The final layer is a fully connected Dense layer with Softmax activation to produce class probability distributions for facial emotion recognition [13]. Through several experiments analyzing the performance of the evaluation, the batch size used in training was determined to be 64 and trained for 100 epochs using the Adam optimizer and an initial learning rate of 0.0001, which was gradually adjusted through exponential decay. Table 2 summarizes the final hyperparameter configuration used across all experiments. These values represent the selected configuration applied consistently to all datasets in order to ensure a fair cross-dataset comparison.

To further prevent overfitting, early stopping was also set to prevent overfitting to the training data. Early stopping will terminate training if there is no improvement in accuracy or decrease in loss on the validation data for 15 consecutive epochs. To save the best weights during training, a model checkpoint was also applied, so that the final model evaluated was the best model.

D. Hyperparameter Tuning Procedure

Hyperparameter tuning was performed using a controlled, iterative experimental strategy. At each stage, one hyperparameter, such as batch size, learning rate, or convolutional filter configuration, was changed while the others remained constant, allowing the impact of each parameter to be analyzed independently.

Unlike exhaustive grid search or Bayesian optimization methods, which require significant computational resources, this approach was chosen to improve readability and computational efficiency. Model performance was evaluated based on validation accuracy and loss stability across the dataset.

1. Initial configuration selection

The initial configuration used in the training refers to the general pattern often used in CNN-based FER research, namely input sizes that follow the original FER-2013 format of 48×48, Adam optimizer with an initial learning rate in the range of 1e-3, batch sizes between 32 and 64, several multi-level convolution blocks, and the use of dropout to reduce overfitting. Initial testing was first conducted on public datasets to obtain baseline performance and training curve patterns before further variations were made.

2. Hyperparameter variation experiments

In the next stage, a series of controlled experiments were conducted by changing one to two hyperparameters at a time to be assessed and further evaluated for their effect on the performance of the resulting model. The variations tested were as follows:

- **Learning rate:** With the aim of determining the effect of learning speed on training stability, tests were conducted on learning rates of 1e-3 and 1e-4 with an exponential decay scheme [14]. The results showed that using a learning rate of 1e-4 produced better performance than using a higher learning rate.
- **Batch size:** A comparison of model performance was conducted using batch sizes of 32 and 64. The experiment showed that the model performance with a batch size of 64 produced more stable and efficient results. This is also in line with deep learning literature recommendations, which often use medium batch sizes for image data.
- **Configuration of the number of filters per convolution block:** In this study, two filter configurations were tested, namely 64–128–256–256–512 and 32–64–128–128–256. Based on the comparison results, it was finally decided to use the higher filter, 64–128–256–256–512, in this study.
- **Use and value of dropout:** In each convolution block, dropout with a value of 0.3 was used to reduce the risk of overfitting. The use of dropout was determined from initial experiments comparing the performance results of architectures with and without dropout.

3. Final configuration determination

Based on the experimental results, the final hyperparameter combination that provided the best performance and efficiency was determined. This combination was then used as the final configuration for each group of public datasets. The aim was to evaluate the generalization ability of the model with the resulting tuned configuration.

E. Experimental Scenarios

To analyze the influence of dataset characteristics and hyperparameter configurations on CNN performance, several controlled experimental scenarios were designed using the same CNN architecture and final hyperparameter configuration. It is important to note that these scenarios aim to perform cross-dataset performance evaluation, rather than

domain generalization testing, as each dataset was trained and evaluated independently within its own data split.

1. Scenario 1

In the first scenario, the CNN model was trained and evaluated using the FER-2013 and FER+ datasets with an input size of 48×48 and grayscale format. This scenario serves as a baseline evaluation to analyze model behavior on datasets with relatively simple image characteristics and controlled annotation formats.

2. Scenario 2

Training in the second scenario was conducted using the RAF-DB and AffectNet datasets, which have more complex variations in pose and lighting. The image format is maintained in its original format, which is color (RGB). Although these datasets originally provide higher image resolutions, all images were resized to 48×48 to maintain architectural consistency. This scenario aims to examine how increased dataset complexity affects CNN performance under identical model configurations.

By designing these scenarios, performance differences can be attributed primarily to dataset properties rather than architectural or resolution factors.

F. Evaluation Metrics

The evaluation was conducted using test data that was separate from the training and validation sets. Model performance was evaluated using several standard classification metrics. Accuracy was used to measure the overall proportion of facial expressions that were recognized correctly. In addition, model performance was also evaluated using precision, recall, and F1 score. Precision indicates how accurate the model's predictions are for a particular class, recall represents the model's ability to correctly identify examples from each class, and F1 score reflects the balance between precision and recall. Mathematically, the evaluation metrics used in this study are defined as follows. Suppose TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

Accuracy is defined as the ratio of correctly classified samples to the total number of samples, as shown in Equation (1):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision for class k measures the proportion of correctly predicted samples of class k among all samples predicted as class k , as defined in Equation (2):

$$Precision_k = \frac{TP_k}{TP_k+FP_k} \quad (2)$$

Recall for class k represents the model's ability to correctly identify samples belonging to class k , as shown in Equation (3):

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \quad (3)$$

The F1-score for class k is the harmonic mean of precision and recall, which balances both metrics, as defined in Equation (4):

$$F1_k = \frac{Precision_k \cdot Recall_k}{Precision_k + Recall_k} \quad (4)$$

To determine the evaluation of model performance, a confusion matrix is also used to show the distribution of predictions and areas where the model often experiences confusion between classes [15]. This procedure is carried out consistently in all scenarios, both in training the CNN model on the FER-2013, FER+, RAF-DB, and AffectNet datasets.

G. Webcam Deployment

To demonstrate practical application, the best CNN model was integrated into a webcam-based facial emotion recognition prototype using Python and OpenCV. This prototype performs real-time emotion prediction at the frame level under controlled conditions, where subjects are placed directly in front of the camera with stable lighting.

This implementation is intended as a functional proof of concept rather than a comprehensive real-world evaluation. Measurements of latency, testing for resilience to extreme lighting changes, obstructions, or multi-face scenarios are not included in the scope of this study. Future work will focus on evaluating system latency, resilience to environmental variations, and performance stability in uncontrolled real-world environments.

III. RESULTS AND DISCUSSION

This section presents the results of the implementation and evaluation of the CNN model on all datasets. In all scenarios, all models were trained using the same CNN architecture and final hyperparameter configuration, with adjustments to the size and images according to the characteristics of the dataset, so that the assessment could be carried out fairly.

A. CNN Results on Public Datasets

The experiments described below are designed to further explore how different dataset characteristics affect CNN model performance. This approach allows the evaluation of the results of CNN training to show the influence of dataset characteristics on CNN performance.

The results shown in Figure 2 and Table 3 reveal consistent performance trends across various evaluation metrics, indicating that dataset characteristics have a systematic influence on the performance of CNN-based facial emotion recognition. Rather than relying solely on accuracy, the use of macro and weighted average F1 scores provides additional insight into the classification behavior per class under varying dataset conditions.

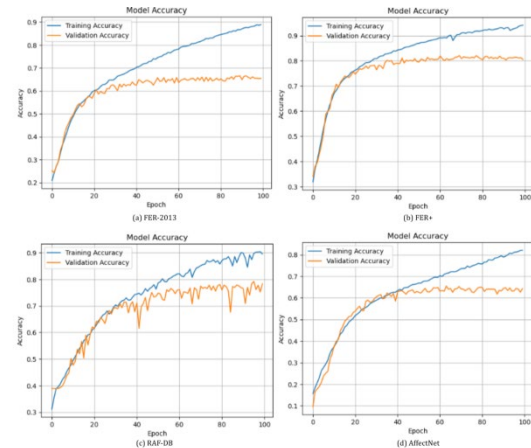


Figure 2. Comparison of CNN Accuracy on Public Datasets

Figure 2 contains a comparison of the accuracy graphs from the training results. Evaluation through this graph was carried out to observe the dynamics of model learning and to observe the potential for overfitting. To provide a clearer picture, Table 3 below contains a comparison of the accuracy, precision, recall, F1-score, and weighted F1-score results from training with each dataset.

Table 3. CNN Model Evaluation Results on Public Datasets

	FER-2013	FER+	RAF-DB	Affect-Net
Accuracy	65.94%	81.86%	76.63%	63.92%
Precision (Macro)	65.94%	74.75%	67.77%	61.17%
Recall (Macro)	63.13%	63.47%	68.26%	59.66%
F1-score (Macro)	64.06%	66.63%	67.55%	59.31%
Weighted F1-score	65.94%	81.54%	76.99%	62.55%

The FER+ dataset achieved the highest accuracy (81.86%) and weighted F1 score (81.54%), but showed a significant difference between the macro F1 score (66.63%) and the weighted F1 score. This difference indicates that the model performs well on dominant emotion classes, which is consistent with the refined annotation process in FER+ that reduces label ambiguity inherited from FER-2013. The re-annotation strategy in FER+ improves the separation between classes, allowing CNN to learn more stable feature representations, especially for visually distinct emotions.

On the other hand, the FER-2013 dataset shows lower and more uniform performance across metrics, with both accuracy and weighted F1 score at 65.94%. The relatively small difference between the macro and weighted F1 scores indicates that classification errors are spread across several classes rather than being dominated by class imbalance alone. This behavior supports the presence of annotation noise and label inconsistencies in FER-2013, which can impair feature learning and reduce overall classification reliability.

The RAF-DB dataset shows a balanced performance profile, with accuracy (76.63%), macro F1 score (67.55%), and weighted F1 score (76.99%) indicating stable recognition of both frequent and infrequent emotion classes. Despite higher variation in poses and lighting, RAF-DB maintains strong performance thanks to high-quality annotations and a controlled emotion labeling process. This shows that annotation consistency can partially compensate for increased visual complexity in CNN-based FER models.

AffectNet is the most challenging case, with the lowest accuracy (63.92%) and lowest macro F1 score (59.31%), despite being the largest dataset in terms of image count. The relatively low macro and weighted F1 scores indicate that increasing the dataset size alone does not guarantee improved generalization. High intra-class variation due to uncontrolled facial expressions, lighting changes, and pose diversity causes overlapping feature distributions between negative emotion classes such as anger, fear, and disgust. Without additional augmentation strategies or domain adaptation, CNNs struggle to learn robust and discriminative features under these conditions.

Overall, these results indicate that annotation reliability and expression consistency play a more critical role in the performance of CNN-based facial emotion recognition systems than dataset size alone. When a unified architectures and hyperparameter configurations are applied, performance variations can be systematically attributed to specific dataset properties rather than architectural biases. This analysis highlights the importance of dataset selection and characterization as fundamental steps in developing reliable facial emotion recognition systems. Although causal inference cannot be fully proven, the consistent performance patterns observed across various evaluation metrics provide strong empirical evidence that dataset annotation quality and intra-class variability significantly affect the performance of CNN-based FER.

To assess the model's performance relative to other studies, Table 4 below presents a comparative review of several previous studies using the same public datasets, namely FER-2013, FER+, RAF-DB, and AffectNet.

Table 4. Comparative Review of Related Works

Study	Year	Architecture	Dataset	Accuracy
[16]	2020	CNN backbone + Region Attention	AffectNet	59.50%
[17]	2021	CNN, ELM, ViolaJones Algorithm	FER-2013	62.5%
[18]	2021	ResNet-50	AffectNet	59.72%
[19]	2022	BaseDCNN	RAF-DB	73.19%
[20]	2022	ResNet-50+Adam+0.001	FER+	73.05%

Study	Year	Architecture	Dataset	Accuracy
[21]	2023	VGG16 pre-trained model	FER+	67.33%
[22]	2024	CLCM	FER-2013, AffectNet	63%, 54%
[23]	2025	LFARNet model	AffectNet	57.03%

From these results, it can be seen that several previous studies, such as the CNN method with ELM and Viola-Jones on FER-2013, ResNet-50 on AffectNet, BaseDCNN on RAF-DB, and VGG16 and ResNet-50 on several public datasets, still produce lower accuracy compared to the model proposed in this study. This confirms that the proposed CNN approach is capable of improving performance while validating the effectiveness of the model design and hyperparameter configuration on the four public datasets.

B. Hyperparameter Influence Analysis

In addition to looking at the performance of CNNs on several datasets, this study also analyzes the effect of hyperparameter variations on CNN performance. In this study, the final hyperparameters obtained through the analysis of the dataset variation experiment were used, namely learning rate, batch size, configuration of the number of filters per convolution block, and the use of dropout.

In the learning rate test, an initial experiment was conducted using a smaller learning rate, namely $1e-4$ with an exponential decay scheme. Next, a comparison experiment was conducted using a learning rate of $1e-3$ with an exponential decay scheme. The test results showed relatively similar performance. The learning rate of $1e-3$ produced a slightly higher test accuracy (82.87%) but with a higher loss value (0.6017). Conversely, the learning rate of $1e-4$ provided an accuracy of 82.01% with a lower loss (0.5572) and a more stable training curve. Therefore, the learning rate of $1e-4$ was used throughout the study, considering the trade-off between accuracy, convergence stability, and probability calibration.

Batch size testing was also conducted by comparing the test results from experiments using batch sizes of 32 and 64. The test results also produced similar accuracy, namely 82.01% for batch size 64 and 82.59% for model test accuracy with batch size 32. However, the test loss produced by the model with a batch size of 32 was very large compared to 64, namely 4.7103, while the test loss in the batch size 64 model was only 0.5572. Therefore, a batch size of 64 was used in all training because using a batch size of 32 could result in an unstable training process and poor probability calibration.

For the configuration of the number of filters per convolution block, a comparison was made between the use of smaller 32-64-128-128-256 filters and 64-128-256-256-512 filters. The lighter configuration resulted in slightly higher test accuracy, namely 82.52% with a higher loss value (0.6487). Meanwhile, the higher configuration resulted in a slightly lower test accuracy of 82.02% with a lower loss value (0.5572). Considering the very small difference in

accuracy and the lower loss value, the filter configuration 64-128-256-256-512 as the final architecture for the entire experiment due to the need for feature representation capacity to deal with expression variations in several public datasets.

In addition, an analysis of the effect of using dropout was also conducted in this study. The architecture without dropout resulted in a test accuracy of 81.56% with a loss value of 0.8265, while using dropout 0.3 increased the accuracy to 82.01% and simultaneously reduced the loss to 0.5572. This proves the function of dropout in reducing overfitting and producing models with better generalization.

Although the hyperparameter tuning process was performed using a controlled iterative strategy, this study did not aim to perform statistical significance tests between different configurations. Instead, parameter selection is based on consistent validation performance trends, convergence stability, and loss behavior throughout the training rounds. Therefore, the observed performance differences are interpreted as empirical tendencies rather than statistically conclusive improvements.

C. In-Depth Evaluation of the Best Model

Based on the evaluation results on the entire public dataset, the CNN model trained using the FER+ dataset showed the best performance because it produced the best accuracy and F1-score values, namely 81.86% and 81.54%. In addition, a confusion matrix analysis was also carried out to see the distribution of predictions for each emotion class. Figure 3 is the confusion matrix from training the CNN model using the FER+ dataset.

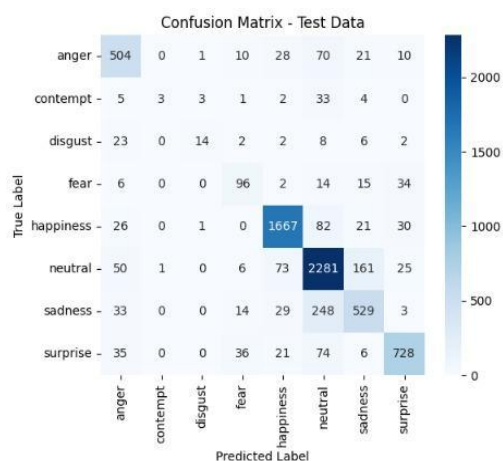


Figure 3. Confusion Matrix of the Best Model with FER+

The confusion matrix shows that the CNN model consistently achieves high positive accuracy rates for expressions of happiness and neutral, which have distinctive and stable facial features even on low-resolution grayscale inputs. On the other hand, systematic classification error patterns are observed in the negative emotion classes, particularly anger, fear, and disgust. These emotions have similar facial muscle activations and subtle visual

differences, causing overlapping feature representations learned by the CNN.

Furthermore, classification errors between sadness and fear further indicate that the model has difficulty distinguishing emotions with similar expression intensities and shape variations. This behavior is reflected in the evaluation metrics, where the macro precision and macro recall values are lower than the weighted F1 score. This discrepancy indicates that although the model performs well on dominant classes, recognizing less frequent or visually ambiguous emotions remains challenging. These results suggest that the observed classification errors are driven by intrinsic visual similarities between emotion classes rather than random prediction noise.

The FER+ model was selected for further evaluation not only based on its highest accuracy and F1-weighted score, but also because it demonstrated the most stable and balanced performance profile across various evaluation metrics. The relatively consistent difference between the macro average metric and the weighted metric indicates an improvement in class separation ability and a decrease in sensitivity to annotation ambiguity compared to other datasets.

This performance stability indicates that the refined annotation process in FER+ allows CNNs to learn more discriminative and consistent emotion representations under a unified architecture and hyperparameter configuration. Furthermore, from an implementation perspective, the use of 48×48 grayscale images together with the selected hyperparameters (learning rate 1e-4, batch size 64, dropout rate of 0.3, and filter configuration of 64-128-256-256-512) provides an effective balance between recognition accuracy, convergence stability, and computational efficiency. Therefore, the FER+ model is considered the most representative candidate for integration into webcam-based facial emotion recognition systems.

D. Webcam System Evaluation

In the final stage of this research, an implementation was carried out in a webcam-based facial expression recognition system to demonstrate the practical applicability of the proposed CNN model (Figure 4). The implementation was developed using the Python programming language, while OpenCV was utilized to access the webcam video stream and perform image processing tasks [24].

Each frame captured from the webcam is converted to grayscale, followed by face detection using the Haar Cascade algorithm. This method was selected due to its computational efficiency and its ability to detect multiple faces in a single frame with sufficient accuracy for demonstration purposes [25]. The detected face region is then cropped, resized to a resolution of 48×48 pixels, and normalized using the same preprocessing scheme applied during the training phase. The processed facial image is subsequently fed into the trained CNN model to generate emotion class probability predictions. This implementation is intended to demonstrate that the proposed model is not limited to offline evaluation but can also be deployed in interactive, frame-based applications using live webcam input.

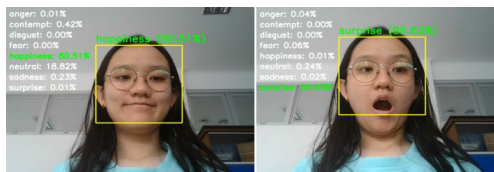


Figure 4. Webcam-Based FER System Display

IV. CONCLUSION

This study examines the performance of a single CNN architecture with fixed hyperparameter configurations on various public facial emotion recognition datasets with different characteristics. The experimental results show that, when trained with identical architecture and training settings, CNN models exhibit different levels of performance depending on the quality of dataset annotation, class balance, and data complexity.

The results indicate that datasets with more detailed annotations and more consistent label distributions, such as FER+ and RAF-DB, tend to produce higher accuracy and F1-scores compared to datasets with higher label noise or greater variation, such as FER-2013 and AffectNet. This indicates that dataset characteristics play a significant role in influencing CNN-based FER performance, even when the same model configuration is applied.

From the hyperparameter influence analysis, the selected configuration—learning rate of $1e-4$ with exponential decay, batch size of 64, dropout rate of 0.3, and filter configuration of 64–128–256–256–512—provided a stable balance between accuracy, loss convergence, and training stability across the dataset. Based on these findings, the CNN model trained on the FER+ dataset achieved the most consistent performance and was selected for further implementation.

A webcam-based facial expression recognition system was implemented as a functional demonstration to show the feasibility of applying the trained model in interactive scenarios. This implementation aimed to demonstrate deployment feasibility rather than provide a quantitative evaluation of real-time system performance.

The main contribution of this research lies in the systematic analysis of CNN performance on various public FER datasets using uniform architectures and hyperparameter settings, as well as highlighting the influence of dataset characteristics on model behavior. Future research will focus on the application of advanced data augmentation techniques or domain adaptation, as well as conducting comprehensive real-world performance evaluations to improve model robustness and generalization.

REFERENCES

- [1] J. Shen, J. Yu, H. Zhang, M. A. Lindsey, and R. An, “Artificial intelligence-powered social robots for promoting physical activity in older adults: A systematic review,” Dec. 01, 2025, *Elsevier B.V.* doi: 10.1016/j.jshs.2025.101045.
- [2] T. Kujani and V. D. Kumar, “Emotion Understanding from Facial Expressions using Stacked Generative Adversarial Network (GAN) and Deep Convolution Neural Network (DCNN),” *International Journal of Engineering Trends and Technology*, vol. 70, no. 10, pp. 98–110, 2022, doi: 10.14445/22315381/IJETT-V70I10P212.
- [3] D. de M. Fabrício, B. L. C. Ferreira, M. A. Maximiano-Barreto, M. Muniz, and M. H. N. Chagas, “Construction of face databases for tasks to recognize facial expressions of basic emotions: a systematic review,” *Dementia e Neuropsychologia*, vol. 16, no. 4, pp. 388–410, 2022, doi: 10.1590/1980-5764-DN-2022-0039.
- [4] N. F. Alkhaleeli, “Facial Expression Recognition and Emotion Detection with CNN methods And SVM Classifiers,” *Journal of Millimeterwave Communication, Optimization and Modelling*, vol. 4, no. January 2023, 2024.
- [5] E. Mansouri-Benssassi and J. Ye, “Generalisation and robustness investigation for facial and speech emotion recognition using bio-inspired spiking neural networks,” *Soft comput.*, vol. 25, no. 3, pp. 1717–1730, 2021, doi: 10.1007/s00500-020-05501-7.
- [6] C. Duan, “A survey of facial expression recognition in the wild,” *Applied and Computational Engineering*, vol. 6, no. 1, pp. 98–106, 2023, doi: 10.54254/2755-2721/6/20230760.
- [7] Y. Fan *et al.*, “FER-PCVT: Facial Expression Recognition with Patch-Convolutional Vision Transformer for Stroke Patients,” *Brain Sci.*, vol. 12, no. 12, 2022, doi: 10.3390/brainsci12121626.
- [8] Z. Y. Huang *et al.*, “A study on computer vision for facial emotion recognition,” *Sci. Rep.*, vol. 13, no. 1, pp. 1–13, 2023, doi: 10.1038/s41598-023-35446-4.
- [9] J. Wang, “Evaluation and analysis of visual perception using attention-enhanced computation in multimedia affective computing,” *Front. Neurosci.*, vol. 18, 2024, doi: 10.3389/fnins.2024.1449527.
- [10] M. I. Abdi and B. Ergen, “Facial Expression Based Emotion Recognition,” *Journal homepage: Journal of Electrical Engineering and Computer (JEECOM)*, vol. 7, no. 1, 2025, doi: 10.33650/jeeecom.v4i2.
- [11] D. Pruthviraja, U. M. Kumar, S. Parameswaran, V. G. Chowdary, and V. Bharadwaj, “Deep convolutional neural network architecture for facial emotion recognition,” *PeerJ Comput. Sci.*, vol. 10, pp. 1–20, 2024, doi: 10.7717/peerj-cs.2339.
- [12] B. H. Pansambal and A. B. Nandgaokar, “Integrating Dropout Regularization Technique at Different Layers to Improve the Performance of Neural Networks,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, pp. 716–722, 2023, doi: 10.14569/IJACSA.2023.0140478.
- [13] M. Ramdhany Edy, “Deteksi Emosi dari Ekspresi Wajah dengan Deep Learning,” *INTEC Journal: Information Technology Education Journal*, vol. 3, no. 2, 2024.

- [14] M. Smaida, S. Yaroshchak, and A. Y. Ben Sasi, "Learning Rate Optimization in CNN for Accurate Ophthalmic Classification," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 3075, no. 4, pp. 211–216, 2021, doi: 10.35940/ijitee.B8259.0210421.
- [15] A. R. Fadillah, C. Sri, and K. Aditya, "Gfpgan Upscaling for Human Facial Expression Classification Using Vgg19 Architecture," vol. 11, no. 1, pp. 18–26, 2025, doi: 10.33480/jitk.v11i1.6588.INTRODUCTION.
- [16] and Y. Q. Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 4057–4069, 2020.
- [17] Y. K. Bhatti, A. Jamil, N. Nida, M. H. Yousaf, S. Viriri, and S. A. Velastin, "Facial Expression Recognition of Instructor Using Deep Features and Extreme Learning Machine," *Comput. Intell. Neurosci.*, vol. 2021, 2021, doi: 10.1155/2021/5570870.
- [18] F. Y. Rahadika, N. Yudistira, and Y. A. Sari, "Facial Expression Recognition using Residual Convnet with Image Augmentations," *Jurnal Ilmu Komputer dan Informasi*, vol. 14, no. 2, pp. 127–135, Jul. 2021, doi: 10.21609/jiki.v14i2.968.
- [19] W. Yu and H. Xu, "Co-attentive multi-task convolutional neural network for facial expression recognition," *Pattern Recognit.*, vol. 123, p. 108401, 2022, doi: 10.1016/j.patcog.2021.108401.
- [20] H. B. Khoirullah, N. Yudistira, and F. Abdurrahman Bachtiar, "Facial Expression Recognition Using Convolutional Neural Network with Attention Module," *INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION*, Dec. 2022, doi: 10.30630/joiv.6.4.963.
- [21] A. Y. Nawaf and W. M. Jasim, "A pre-trained model vs dedicated convolution neural networks for emotion recognition," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 1123–1133, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1123-1133.
- [22] M. C. Gursesli, S. Lombardi, M. Duradoni, L. Bocchi, A. Guazzini, and A. Lanata, "Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model: Performance Evaluation in Public Datasets," *IEEE Access*, vol. 12, pp. 45543–45559, 2024, doi: 10.1109/ACCESS.2024.3380847.
- [23] L. Duongthang, "Driver facial emotion tracking using an enhanced residual network with weighted fusion of channel and spatial attention," *Sci. Rep.*, vol. 15, no. 1, pp. 1–10, 2025, doi: 10.1038/s41598-025-97451-z.
- [24] G. J. P. Hakim *et al.*, "Real-Time Facial Emotion Detection Application with Image Processing Based on Convolutional Neural Network (CNN)," *International Journal of Electrical Engineering, Mathematics and Computer Science*, vol. 1, no. 4, pp. 27–36, Nov. 2024, doi: 10.62951/ijeemcs.v1i4.123.
- [25] O. Fassya Enbiva and R. Toyib, "Penerapan Metode Haar Cascade Untuk Deteksi Pola Wajah Berdasarkan Perkembangan Usia," *Jurnal Informatika Teknologi dan Sains*, Aug. 2025.