APPLIED ARTIFICIAL INTELLIGENCE AND ENGINEERING JOURNAL
E-ISSN: 3124-1212, DOI: 00.000/XXXX.YYYY.II.NNN
Vol. 1 No. 1 2026 pp. 35-47

Facial Emotion Classification Using Transfer Learning with NASNetMobile on the AffectNet Dataset

Livia Safirani[1], Muthiah Rahma Pasaribu[2], Saidah Cilmi Dewi[3*]

[1,2] Program Studi Teknik Informatika, Jurusan Teknologi Informasi dan Komputer, Politeknik Negeri Lhokseumawe, Indonesia
[3] Program Studi Psikologi, Fakultas Psikologi, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia

*Corresponding Author: 230401110247@student-uin-malang.ac.id

Article info: Received 20/03/2026, Revised 28/03/2026, Accepted 05/04/2026

Abstract

Facial expressions represent one of the most important forms of nonverbal communication for conveying human emotions. However, automatic facial emotion classification remains a challenging task due to variations in illumination, facial pose, background complexity, and individual facial characteristics. This study aims to analyze facial emotion classification using a deep learning approach based on Convolutional Neural Networks (CNNs). The experiments were conducted on the AffectNet dataset, a large-scale benchmark containing in-the-wild facial images, with the classification task focused on two emotional categories: happy and fear. The proposed framework consists of several preprocessing stages, including face detection, image resizing, normalization, and data augmentation, followed by CNN training using a transfer learning approach with NASNetMobile as the feature extraction backbone. Experimental results demonstrate that the proposed model achieved a test accuracy of 96%, with F1-scores of 0.97 for the happy class and 0.95 for the fear class. The model exhibited superior performance in recognizing the happy emotion, whereas the fear class produced relatively more misclassifications due to its greater visual complexity and higher intra-class variability. Overall, the findings indicate that the combination of appropriate image preprocessing, data augmentation, and transfer learning significantly enhances the performance and generalization capability of CNN-based facial emotion classification under real-world conditions.

Keywords: Facial Emotion Recognition, Transfer Learning, NASNetMobile, AffectNet, Computer Vision

1. Introduction

Facial expressions constitute one of the most informative forms of nonverbal communication, enabling humans to convey emotions, intentions, and psychological states without verbal interaction. Variations in facial components, such as the eyes, eyebrows, nose, and mouth, provide visual cues that allow individuals to recognize emotional responses during social interactions. Accurate interpretation of facial expressions is therefore essential for effective human communication and plays a significant role in decision-making processes in everyday life. With the rapid advancement of artificial intelligence (AI), the automatic analysis of facial expressions has become an increasingly important research topic, attracting considerable attention from both academia and industry because of its potential to facilitate more natural and intelligent human–computer interaction (HCI) [1]. Automatic Facial Emotion Recognition (FER) has been widely adopted in numerous real-world applications, including intelligent surveillance, healthcare monitoring, driver fatigue detection, online education, customer behavior analysis, social robotics, and affective computing. In healthcare, FER can assist clinicians in monitoring patients with neurological disorders, depression, autism spectrum disorder, and cognitive impairment by providing objective assessments of emotional responses. In educational environments, emotion recognition enables intelligent tutoring systems to evaluate student engagement and learning effectiveness in real time. Likewise, commercial applications exploit facial emotion analysis to understand customer satisfaction and improve personalized services. As these applications continue to expand, the demand for robust and accurate FER systems capable of operating under unconstrained real-world conditions has increased substantially [2].

Despite the remarkable progress achieved in facial emotion recognition, automatic classification of facial expressions remains a challenging computer vision problem. Unlike controlled laboratory environments, real-world facial images exhibit considerable variations in illumination, facial pose, camera viewpoint, image resolution, occlusion, facial accessories, background clutter, and expression intensity. Furthermore, significant differences in facial morphology, age, gender, ethnicity, and individual expression patterns introduce substantial intra-class variability while reducing inter-class separability. These factors often degrade feature discriminability and consequently reduce the classification performance of emotion recognition systems when deployed in practical environments [3]. Traditional facial emotion recognition approaches primarily rely on handcrafted feature extraction techniques combined with conventional machine learning classifiers. Popular descriptors include Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), Gabor filters, and geometric facial landmarks, which are subsequently classified using algorithms such as Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Decision Tree, and Random Forest. Although these methods demonstrate satisfactory performance under controlled conditions, their effectiveness is highly dependent on manually designed features that often fail to capture the complex and nonlinear characteristics of facial expressions. Consequently, handcrafted feature-based approaches exhibit limited robustness when confronted with large appearance variations and complex real-world scenarios [4].

Recent advances in deep learning have fundamentally transformed image recognition by enabling end-to-end feature learning directly from raw image data. Among various deep learning architectures, the Convolutional Neural Network (CNN) has emerged as one of the most successful models for visual recognition because of its ability to automatically learn hierarchical representations ranging from low-level edge information to high-level semantic features. Compared with traditional handcrafted descriptors, CNN-based models produce more discriminative and adaptive feature representations, resulting in superior performance across numerous computer vision applications, including object detection, medical image analysis, plant disease recognition, biometric identification, and facial emotion recognition [5]. Although CNN-based methods significantly outperform conventional machine learning techniques, their performance remains sensitive to the quality and diversity of the training data. Deep neural networks generally require large-scale annotated datasets to achieve strong generalization capability. In facial emotion recognition, additional challenges arise from pose variation, facial occlusion, illumination changes, background complexity, and class imbalance. These challenges often lead to overfitting and performance degradation, particularly when training data are limited or collected under controlled conditions that fail to represent real-world variability [6].

To overcome these limitations, transfer learning has become one of the most effective strategies for developing high-performance deep learning models while reducing computational cost and training time. Rather than training an entire network from scratch, transfer learning utilizes knowledge acquired from large-scale image datasets to initialize feature extraction layers, thereby enabling efficient adaptation to new classification tasks with relatively limited training samples. This approach not only accelerates convergence but also improves feature generalization and reduces the risk of overfitting. Consequently, transfer learning has become a standard practice in numerous image classification applications, particularly those involving specialized or domain-specific datasets [7]. Among the available lightweight convolutional architectures, NASNetMobile represents an attractive solution for facial emotion recognition because it was designed through Neural Architecture Search (NAS), enabling an optimized balance between classification accuracy and computational efficiency. Compared with deeper and computationally expensive CNN architectures, NASNetMobile provides competitive recognition performance while maintaining a relatively small model size and lower computational complexity. These characteristics make it particularly suitable for deployment in real-time applications and edge computing devices, where computational resources and memory are often limited [8].

Another critical factor influencing FER performance is the selection of an appropriate benchmark dataset. The AffectNet dataset has become one of the most comprehensive resources for facial emotion recognition because it contains more than one million facial images collected from the Internet under unconstrained (in-the-wild) conditions. Unlike laboratory datasets such as CK+, JAFFE, or MMI, AffectNet includes substantial variations in facial pose, illumination, background, facial accessories, occlusion, and expression intensity, making it significantly more representative of practical deployment scenarios. Consequently, models evaluated on AffectNet generally exhibit stronger generalization capability for real-world facial emotion recognition [9]. Motivated by these challenges, this study investigates facial emotion classification using a deep learning framework based on Convolutional Neural Networks (CNNs) and transfer learning with NASNetMobile as the feature extraction backbone. The proposed framework incorporates several preprocessing stages, including face detection, image resizing, normalization, and data augmentation, to improve feature consistency and increase the diversity of training samples. The experiments focus on two representative emotional categories, happy and fear, selected from the AffectNet dataset to evaluate the capability of the proposed model under unconstrained real-world conditions.

The main contributions of this study can be summarized as follows:

- A lightweight CNN-based facial emotion classification framework is developed using transfer learning with NASNetMobile to achieve high recognition accuracy while maintaining computational efficiency.

- A comprehensive preprocessing pipeline, including face detection, normalization, resizing, and image augmentation, is implemented to improve feature quality and enhance model generalization.
- Extensive experiments on the AffectNet dataset demonstrate the effectiveness of the proposed approach for recognizing facial emotions captured under challenging real-world conditions.
- Quantitative performance evaluation using accuracy, precision, recall, F1-score, confusion matrix, and qualitative prediction analysis confirms the robustness of the proposed model, achieving a testing accuracy of 96% while maintaining high classification performance for both emotion categories.

The remainder of this paper is organized as follows. Section 2 describes the dataset, preprocessing procedures, CNN architecture, and transfer learning methodology. Section 3 presents the experimental results and discusses the model performance using quantitative and qualitative evaluations. Finally, Section 4 concludes the study and outlines several directions for future research.

2. Methods

The research methodology was designed to achieve the primary objective of this study, namely, to analyze facial emotion classification using a deep learning approach. Since the study focuses on evaluating the performance of a Convolutional Neural Network (CNN) in recognizing facial expressions, each stage of the methodology—from dataset selection and image preprocessing to model training and classification—was systematically designed to extract discriminative visual features associated with human emotions. A comprehensive workflow was adopted to ensure robust feature learning, effective model generalization, and reliable classification performance under real-world conditions. The proposed methodology consists of four sequential stages: (1) data collection, (2) image preprocessing, (3) CNN model training, and (4) facial emotion classification. These stages are closely interconnected, with the output of each stage serving as the input for the subsequent process. The preprocessing stage enhances image quality and increases data diversity, while the transfer learning-based CNN model learns hierarchical feature representations that are subsequently used to classify facial emotions. Finally, the trained model is evaluated using an independent testing dataset to assess its generalization capability and classification performance. The overall research framework is illustrated in Fig. 1, and each stage is described in detail in the following subsections.

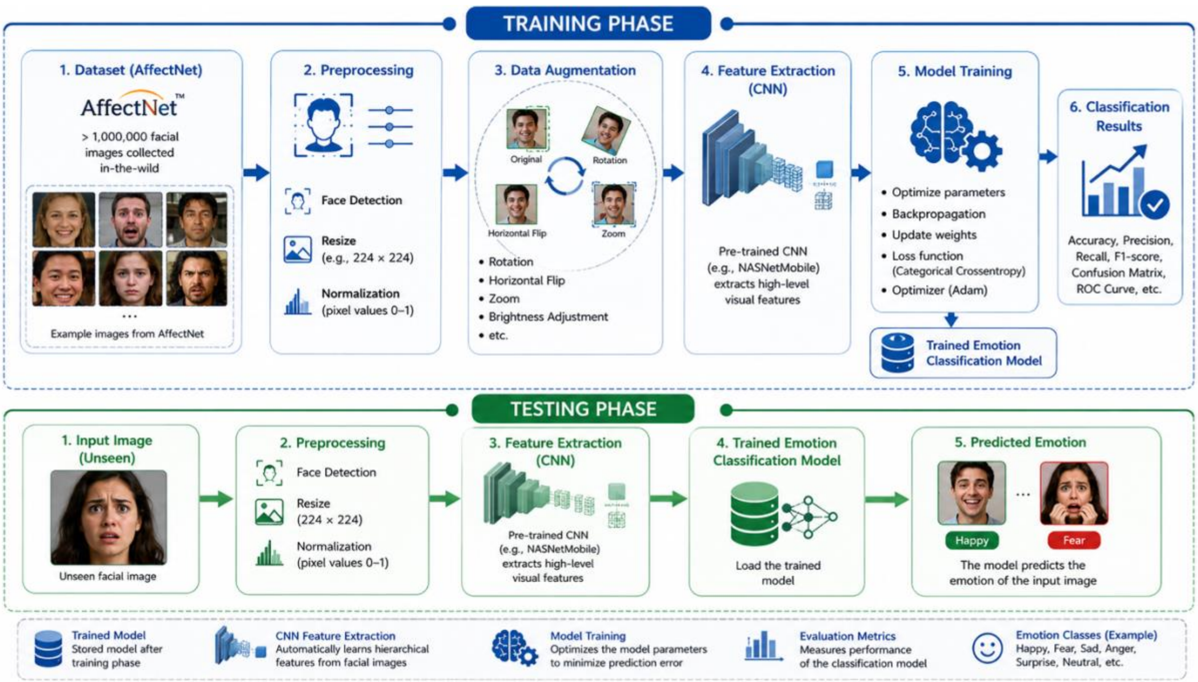


Figure 1. Overall Framework of the Proposed Facial Emotion Recognition System

2.1 Dataset

The performance of a deep learning model is highly dependent on the quality, diversity, and scale of the training dataset. Therefore, selecting an appropriate dataset is a crucial step in developing a robust facial emotion recognition system. In this study, the AffectNet dataset was selected because it is one of the largest and most comprehensive publicly available benchmark datasets for facial emotion recognition under unconstrained (in-the-wild) conditions. Unlike laboratory-controlled datasets, AffectNet contains facial images collected from the Internet, representing substantial variations in illumination, facial pose, occlusion, background, image resolution, ethnicity, age, and facial appearance. These characteristics make the dataset particularly suitable for evaluating the robustness and generalization capability of deep learning models in real-world scenarios [6]. AffectNet contains more than one million facial images, of which approximately 450,000 images have been manually annotated by human experts. Each annotated image is assigned one of eight discrete emotion categories, namely neutral, happy, sad, surprise, fear, disgust, anger, and contempt, together with continuous valence–arousal values that represent the intensity and polarity of human emotions. Owing to its large scale and high annotation quality, AffectNet has become one of the most widely adopted benchmark datasets for evaluating facial emotion recognition algorithms and has been extensively used in recent deep learning studies [6].

This study focuses on a binary emotion classification task involving the happy and fear emotion categories. These two emotions were selected because they exhibit significantly different affective characteristics while presenting varying levels of classification difficulty. The happy class is generally characterized by distinctive facial features, such as raised mouth corners and smiling expressions, making it relatively easier to recognize. In contrast, the fear class often exhibits more subtle and complex facial muscle movements, including widened eyes and tense facial muscles, which increase intra-class variability and make automatic recognition considerably more challenging. Consequently, these two emotion categories provide an appropriate benchmark for evaluating the discriminative capability of the proposed CNN-based classification model. Before model training, all images underwent a preprocessing pipeline consisting of face detection, image resizing, pixel-value normalization, and data augmentation to ensure consistent image quality and improve the generalization capability of the CNN model. The processed images were subsequently divided into training, validation, and testing subsets for model development and performance evaluation. Representative examples of the happy and fear emotion categories from the AffectNet dataset are presented in Table 1.

Table 1. Representative Facial Images of the Happy and Fear Emotion Categories from the AffectNet Dataset

No.	Image	Class
1.		Happy
2.		Fear

The AffectNet dataset has demonstrated excellent effectiveness for training Convolutional Neural Networks (CNNs) in facial emotion recognition because it captures a wide range of facial expression variations under unconstrained (*in-the-wild*) conditions. The diversity of facial poses, illumination, backgrounds, occlusions, and individual facial characteristics enables deep learning models to learn robust and discriminative visual representations. Consequently, models trained on AffectNet exhibit improved generalization capability and achieve more accurate emotion classification when evaluated on real-world facial images [7]. As presented in **Table 1**, the **happy** emotion category contains facial images exhibiting varying smile intensities, different mouth curvatures, slightly narrowed eyes, and raised cheeks, reflecting the natural diversity of positive emotional expressions. These variations enable the CNN model to learn multiple visual patterns associated with happiness rather than relying on a single stereotypical smiling expression. The representative samples illustrate that AffectNet provides rich visual diversity, allowing the model to develop more robust feature representations and improving its ability to recognize happy expressions under different facial poses and environmental conditions.

In contrast, the fear emotion category presents considerably greater visual complexity. Typical facial characteristics include raised eyebrows, widened eyes, a partially open mouth, and increased tension in the upper facial muscles. However, the intensity of these facial movements varies substantially among individuals, ranging from mild surprise-like expressions to intense fear responses. This high intra-class variability makes the fear category more challenging to classify than happy, thereby requiring the CNN model to learn more discriminative hierarchical features. The representative images in Table 1 demonstrate the diversity of fear-related facial expressions contained in AffectNet, which contributes to improving the robustness and generalization capability of the proposed model for real-world facial emotion classification.

2.2 Image Preprocessing

Image preprocessing is a critical stage in facial emotion recognition because it directly affects the quality of the input data used for deep learning model training. The primary objective of preprocessing is to enhance image consistency, reduce irrelevant information, and emphasize facial features that contain emotional cues. Effective preprocessing enables the model to focus on discriminative facial characteristics while minimizing the influence of background clutter, illumination variations, facial pose, and other sources of visual noise. This stage is particularly important when using in-the-wild datasets such as AffectNet, where facial images exhibit substantial variability in acquisition conditions and subject appearance [8]. Unlike images captured in controlled laboratory environments, AffectNet contains facial images collected from the Internet, resulting in considerable variations in lighting conditions, viewing angles, facial orientations, image quality, facial accessories, occlusions, and background complexity. These challenges may introduce redundant visual information that is unrelated to emotional expressions, potentially reducing the effectiveness of feature extraction. Therefore, an appropriate preprocessing pipeline is essential to standardize the input images before they are fed into the Convolutional Neural Network (CNN). In this study, the preprocessing stage consists of four sequential steps: face detection, image resizing, pixel normalization, and image augmentation. Together, these processes improve feature consistency and enhance the generalization capability of the proposed CNN model.

2.2.1 Face Detection

The first stage of the preprocessing pipeline is face detection, which aims to identify and localize the facial region within each image. Since facial emotion recognition relies exclusively on facial characteristics, isolating the face from the surrounding background is essential for improving classification performance. By accurately detecting the facial region, the CNN model is encouraged to learn meaningful emotional features rather than irrelevant visual information originating from the background or surrounding objects. Consequently, face detection improves both the quality of the training data and the discriminative capability of the learned feature representations. Several face detection algorithms have been widely adopted in facial analysis applications, including Multi-task Cascaded Convolutional Networks (MTCNN) and Haar Cascade. These methods are capable of accurately locating facial bounding boxes while maintaining relatively high computational efficiency. In recent facial emotion recognition studies, face detection has been shown to significantly improve classification performance by ensuring that subsequent preprocessing and feature extraction stages operate exclusively on the facial region, where emotional information is primarily expressed [9]. In this study, each facial image from the AffectNet dataset was first processed using a face detection algorithm to identify the facial bounding box.

The detected facial region was then cropped to remove irrelevant background information before being forwarded to the subsequent preprocessing stages. This procedure ensures that the CNN model learns discriminative facial representations while reducing the influence of background noise and non-facial objects. Furthermore, face detection contributes to improving the robustness of the proposed model when handling unconstrained facial images with varying poses and environmental conditions. An example of the face detection process applied to the happy emotion category is presented in Fig. 2, which illustrates the comparison between the original facial image and the corresponding cropped facial region obtained after face detection. The figure demonstrates that the preprocessing stage successfully isolates the facial area while preserving the essential visual characteristics required for accurate emotion classification.

a. Original Image

b. Detected Face Region

Figure 2. Comparison of the Original and the Detected Face Region

2.2.2 Image Resizing

Following the face detection stage, the cropped facial images were subjected to an image resizing process to standardize their spatial dimensions before being used as input to the deep learning model. In this study, all detected facial images were resized to 224×224 pixels, ensuring compatibility with the input requirements of the NASNetMobile architecture employed in the transfer learning framework. Standardizing the image dimensions is an essential preprocessing step because convolutional neural networks require fixed-size inputs to perform efficient feature extraction and batch processing during model training [10]. Image resizing not only ensures architectural compatibility but also improves the consistency of feature learning across the entire dataset. Facial images collected from the AffectNet dataset exhibit substantial variations in image resolution and facial scale due to differences in image acquisition conditions. Without a standardized image size, these variations may introduce inconsistencies in the extracted feature maps and adversely affect the learning process. By resizing all facial images to a uniform resolution, the CNN is able to learn hierarchical facial representations from images with consistent spatial dimensions, thereby improving feature stability and classification performance.

The selected image resolution of 224 × 224 pixels provides an appropriate balance between computational efficiency and feature preservation. Although resizing inevitably modifies the original image dimensions, the process preserves the most important facial characteristics required for emotion recognition while reducing computational complexity and GPU memory consumption during training. Moreover, maintaining a fixed image size facilitates efficient mini-batch optimization and accelerates model convergence. The image resizing procedure was applied after face detection and cropping so that only the facial region was standardized, while irrelevant background information had already been removed. This sequential preprocessing strategy ensures that the resized images retain the essential emotional features, including facial landmarks and muscle movements around the eyes, eyebrows, nose, and mouth, which are critical for accurate facial emotion classification. As illustrated in Fig. 3, the original facial images exhibit considerable variation in image resolution, facial scale, and aspect ratio due to differences in image acquisition conditions. Such inconsistencies may negatively affect the feature extraction process, as the CNN could learn scale-dependent patterns instead of discriminative emotional characteristics. After the resizing process, all facial images were standardized to a fixed resolution of 224 × 224 pixels, resulting in uniform image dimensions and consistent facial proportions across the dataset. This standardization enables the CNN to extract hierarchical facial features more consistently, improves training stability, and enhances the model's ability to learn representative emotional patterns. Consequently, the resizing process contributes to improved classification accuracy and better generalization performance by ensuring that all input images conform to the architectural requirements of the NASNetMobile network.

a. Original Image

b. After Resizing Process

Figure 3. Original and Resized Images

2.2.3 Image Normalization

After the image resizing stage, the facial images were subjected to image normalization to standardize the pixel value distribution before being used for CNN training. Image normalization is a fundamental preprocessing technique that reduces variations in pixel intensity and improves numerical stability during deep learning optimization. Since facial images in the AffectNet dataset were collected under unconstrained (in-the-wild) conditions, they exhibit considerable differences in illumination, contrast, and color intensity. These variations may negatively affect feature extraction and hinder the convergence of the learning process if left untreated. In this study, image normalization was performed using z-score normalization, in which each pixel value was standardized by subtracting the dataset mean ((\mu)) and dividing by the standard deviation ((\sigma)). The normalized pixel value is computed according to Equation (1):

$$\mathrm{xnorm} = \mathrm{x} - \mu/\sigma \tag{1}$$

where (x) denotes the original pixel value, (\mu) represents the mean pixel intensity, and (\sigma) is the standard deviation of the pixel values. This transformation produces pixel distributions with approximately zero mean and unit variance, thereby reducing the influence of extreme intensity values during network optimization.

Image normalization provides several advantages for CNN-based facial emotion recognition. First, it improves the numerical stability of the optimization process by preventing excessively large pixel values from dominating the gradient updates. Second, normalization accelerates model convergence by providing a more consistent input distribution, allowing the optimizer to learn more efficiently. Third, it enhances the robustness of feature extraction by reducing the impact of illumination and contrast variations that commonly occur in real-world facial images. Consequently, the CNN can focus on learning discriminative facial characteristics associated with emotional expressions rather than being influenced by differences in lighting conditions or image intensity [11].

The effect of the normalization process on the facial images is illustrated in Fig. 4, which compares representative happy facial expression images before and after normalization. As shown in the figure, normalization produces a more consistent pixel intensity distribution while preserving the essential facial structures and emotional characteristics required for accurate emotion classification. This preprocessing step contributes to improved training stability, faster convergence, and enhanced classification performance of the proposed CNN model.

(a) Before Normalzation

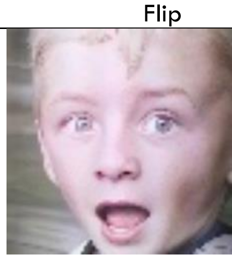
(b) After Normalization

Figure 4. Original and normalized Images

2.2.4 Image Augmentation

Following image normalization, an image augmentation strategy was applied to artificially increase the diversity of the training dataset before model training. Data augmentation is a widely adopted technique in deep learning because it enables the model to learn more robust and generalized feature representations without requiring the collection of additional annotated images. This preprocessing stage is particularly important for facial emotion recognition, where variations in head pose, facial orientation, illumination, and expression intensity frequently occur under real-world (in-the-wild) conditions. By exposing the CNN model to a broader range of facial appearances during training, image augmentation effectively reduces the risk of overfitting and improves the model's ability to generalize to previously unseen facial images [12]. In the proposed preprocessing pipeline, image augmentation was performed after the resizing and normalization stages to ensure that all transformed images possessed consistent spatial dimensions and standardized pixel distributions. Consequently, the augmentation process introduced only realistic geometric and photometric variations while preserving the intrinsic emotional characteristics of each facial expression. This sequential preprocessing strategy enables the CNN to learn invariant facial features that are robust to changes in viewing conditions rather than memorizing specific image patterns.

Tabel 2. Augmentation Results

No.	Image	Class
1.	 Original  Rotate  Bright  Flip	Happy
2.	 Original  Rotate  Bright  Flip	Fear

Several augmentation techniques were employed to simulate common variations encountered in practical facial emotion recognition scenarios. Rotation was applied to emulate slight head tilts and camera orientation changes that naturally occur during image acquisition. Horizontal flipping generated mirror images of facial expressions, allowing the model to learn facial symmetry and effectively increasing the number of training samples. In addition, brightness adjustment was introduced to simulate different illumination conditions, thereby improving the model's robustness to lighting variations frequently observed in unconstrained environments. Collectively, these geometric and photometric transformations encourage the CNN to focus on discriminative facial features that remain stable despite variations in pose, orientation, and illumination. The integration of these augmentation techniques substantially enhances the diversity of the training data while preserving the semantic information associated with each emotion category. As a result, the CNN learns more generalized feature representations, improves classification robustness, and exhibits greater resilience when processing facial images captured under diverse environmental conditions. Furthermore, data augmentation contributes to faster convergence and improved generalization performance by reducing the likelihood of the model memorizing the original training samples. Representative examples of the augmented facial images for the happy and fear emotion categories are presented in Table 2. The table illustrates the visual transformations produced by each augmentation technique while demonstrating that the essential emotional characteristics of the facial expressions are preserved. These examples confirm that the augmentation process successfully increases the visual diversity of the dataset without altering the underlying emotion labels, thereby providing richer training samples for the proposed CNN-based facial emotion classification model.

2.3 Convolutional Neural Network (CNN)

Following the image preprocessing stage, the next step in the proposed framework is facial emotion classification using a Convolutional Neural Network (CNN). CNN has become one of the most successful deep learning architectures for image recognition because of its ability to automatically learn hierarchical feature representations directly from raw pixel data. Unlike conventional machine learning approaches that depend on handcrafted features, CNN performs end-to-end learning by simultaneously extracting discriminative visual features and optimizing the classification model. This capability enables CNN to effectively capture subtle facial characteristics associated with different emotional expressions, making it particularly suitable for facial emotion recognition tasks [13]. Rather than training a CNN model from scratch, this study adopts a transfer learning strategy to improve classification performance while reducing computational cost and training time. Transfer learning allows knowledge learned from large-scale image datasets to be transferred to a target task with relatively limited training data. Since pretrained models have already learned rich low-level and high-level visual representations, they can be efficiently adapted to facial emotion classification with only minor architectural modifications and parameter optimization. Consequently, transfer learning significantly accelerates convergence, improves feature generalization, and minimizes the risk of overfitting [13].

2.3.1 Transfer Learning

The transfer learning framework employed in this study utilizes NASNetMobile as the backbone feature extractor. NASNetMobile is a lightweight convolutional neural network architecture developed using Neural Architecture Search (NAS) and pretrained on the ImageNet dataset, which contains more than one million natural images across one thousand object categories. Through pretraining on ImageNet, NASNetMobile has learned rich hierarchical visual representations that can be transferred to downstream computer vision tasks, including facial emotion recognition. Compared with conventional CNN architectures, NASNetMobile provides an excellent balance between classification accuracy and computational efficiency. Its relatively small number of parameters and optimized network structure make it suitable for deployment in real-time and resource-constrained environments while maintaining competitive recognition performance. For facial emotion recognition, transfer learning enables the pretrained convolutional layers to extract generic facial features such as edges, textures, contours, and facial structures, whereas the newly added classification layers are optimized to learn emotion-specific representations. Previous studies have demonstrated that transfer learning using pretrained CNN models consistently outperforms models trained entirely from scratch, particularly when the available training dataset is relatively limited [14].

2.3.2 CNN Model Architecture

The proposed facial emotion recognition model was developed by employing NASNetMobile as the base network with the include_top parameter set to False, allowing the original ImageNet classification layer to be removed. This configuration enables the pretrained convolutional layers to function solely as a feature extraction backbone while permitting the addition of customized classification layers specifically designed for facial emotion recognition. Following the NASNetMobile feature extractor, several additional layers were incorporated to improve feature learning and classification performance. A Conv2D layer containing 32 convolutional filters with a 3×3 kernel size, ReLU activation function, and same padding was first added to further refine the extracted feature maps. Subsequently, a MaxPooling2D layer with a 2×2 pooling window was employed to reduce the spatial dimensions of the feature maps while preserving the most discriminative information. To minimize overfitting during training, a Dropout layer with a dropout rate of 0.5 was introduced by randomly deactivating a subset of neurons in each training iteration. The resulting feature maps were then transformed into a one-dimensional feature vector using a Flatten layer before being forwarded to the fully connected classification layer. Finally, a Dense output layer equipped with the Softmax activation function generated the probability distribution over the target emotion classes, namely happy and fear. This architectural design effectively combines the strong feature extraction capability of NASNetMobile with lightweight task-specific classification layers, thereby achieving an appropriate balance between computational efficiency and recognition accuracy [15].

2.3.3 Model Compilation

After constructing the CNN architecture, the model was compiled prior to the training process. Since the classification task involves categorical emotion labels, the categorical cross-entropy loss function was employed to quantify the discrepancy between the predicted probability distribution and the corresponding ground-truth labels. This loss function is widely adopted for multi-class image classification because it provides stable gradient updates and facilitates efficient optimization. Model optimization was performed using the Adam optimizer with an initial learning rate of 0.001. Adam was selected because it adaptively adjusts the learning rate for each trainable parameter based on the first and second moments of the gradients, resulting in faster convergence and improved optimization stability compared with conventional stochastic gradient descent algorithms. Throughout the training process, classification accuracy was used as the primary evaluation metric to monitor the model's ability to correctly predict facial emotion categories. The combination of categorical cross-entropy and the Adam optimizer has been extensively validated as one of the most effective optimization strategies for CNN-based image classification tasks [16].

2.3.4 Model Training

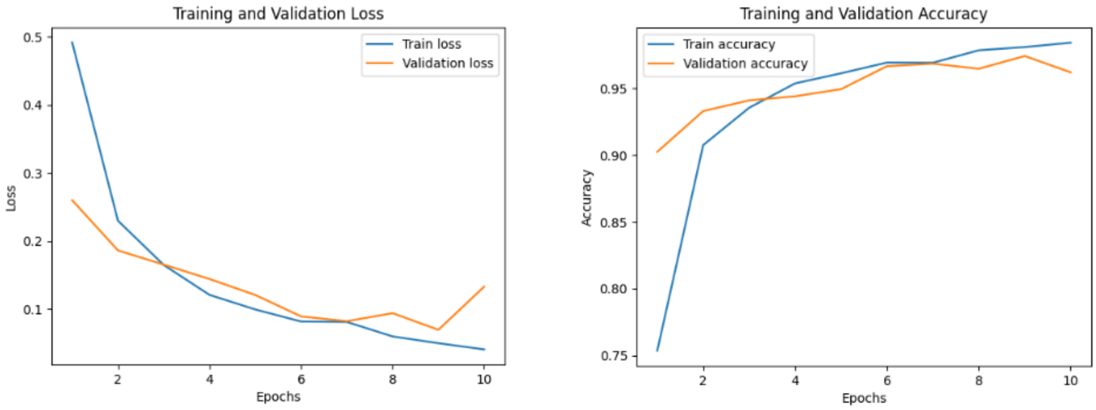
The proposed CNN model was trained for 30 epochs using a batch size of 64. During the training stage, real-time image augmentation was applied exclusively to the training dataset to increase sample diversity and improve the model's generalization capability. In contrast, the validation and testing datasets underwent only image rescaling to preserve their original visual characteristics and ensure unbiased model evaluation. Model training and validation were performed using the flow_from_dataframe data generator, which efficiently loads images directly from the prepared dataframe while generating mini-batches during training. This approach reduces memory consumption and enables seamless integration with online image augmentation. Throughout the optimization process, the CNN continuously updated its trainable parameters by minimizing the categorical cross-entropy loss through backpropagation and gradient-based optimization. The selection of training hyperparameters, including the number of epochs, batch size, optimizer, and data augmentation strategy, was carefully designed to achieve stable convergence while minimizing overfitting. During training, both training accuracy, validation accuracy, training loss, and validation loss were continuously monitored to evaluate learning progress and assess the generalization capability of the proposed facial emotion recognition model. Appropriate hyperparameter selection and efficient data generation have been shown to substantially improve CNN robustness and classification performance in facial emotion recognition applications [17].

3. Result and Discussions

This section presents the experimental results of the proposed Convolutional Neural Network (CNN) model for facial emotion classification using the AffectNet dataset. The objective of the experiments is to evaluate the effectiveness of the proposed transfer learning framework in recognizing facial expressions under unconstrained (in-the-wild) conditions. Model performance is analyzed comprehensively through the training process, validation performance, and final evaluation using an independent testing dataset that was not involved during model training. Several quantitative evaluation metrics, including accuracy, precision, recall, F1-score, confusion matrix, and classification report, are employed to assess the classification capability of the proposed model. In addition, the learning behavior of the CNN is examined through training and validation curves to evaluate model convergence, learning stability, and potential overfitting. The experimental findings are organized into several subsections to provide a comprehensive analysis of the proposed facial emotion recognition system.

3.1 CNN Model Training Results

The CNN model was trained using the preprocessed and augmented AffectNet dataset following the methodology described in the previous section. During the training process, the pretrained NASNetMobile network functioned as the feature extraction backbone, while the newly added classification layers were optimized specifically for the facial emotion recognition task. Image augmentation was applied exclusively to the training dataset to increase sample diversity and improve the model's generalization capability, whereas the validation dataset underwent only image rescaling to provide an unbiased assessment of model performance. Training was conducted for several epochs until the optimization process reached convergence. Throughout the training procedure, the model continuously updated its parameters using backpropagation and the Adam optimization algorithm to minimize the categorical cross-entropy loss. Simultaneously, the values of training accuracy, validation accuracy, training loss, and validation loss were monitored at each epoch to evaluate the learning progress and determine whether the model exhibited signs of overfitting or underfitting. Overall, the experimental results demonstrate a consistent improvement in model performance as the number of training epochs increased. This behavior is indicated by the gradual increase in both training and validation accuracy, accompanied by a corresponding decrease in training and validation loss. Such learning characteristics indicate that the proposed CNN model successfully learned discriminative facial representations while maintaining good generalization capability on unseen validation data. Furthermore, the relatively small gap between the training and validation performance suggests that the adopted preprocessing strategy, image augmentation, transfer learning framework, and regularization techniques effectively reduced the risk of overfitting.



(a) Loss (b) Accuracy

Figure 5. Training Loss dan Accuracy

As illustrated in Fig. 5, the proposed CNN model exhibited stable and consistent learning behavior throughout the training process. The training loss decreased substantially from approximately 0.48 in the initial epoch to around 0.05 by the tenth epoch, indicating that the network progressively minimized prediction errors while learning increasingly discriminative facial representations. Similarly, the validation loss declined from approximately 0.26 to 0.13, although minor fluctuations were observed during the final training epochs. Such fluctuations are commonly encountered in deep learning optimization and generally reflect natural variations in the validation data rather than indicating unstable model convergence. The accuracy curves further demonstrate the effectiveness of the proposed training strategy. The training accuracy increased steadily from approximately 75% in the first epoch to nearly 99% at the end of training, while the validation accuracy improved from approximately 88% to 97%. The continuous improvement in both training and validation accuracy indicates that the CNN successfully learned discriminative facial features associated with the happy and fear emotion categories. Furthermore, the relatively small gap between the training and validation curves suggests that the model maintained strong generalization capability without memorizing the training samples.

The simultaneous decrease in loss and increase in accuracy indicate that the optimization process converged effectively and that the selected hyperparameters—including transfer learning with NASNetMobile, image preprocessing, data augmentation, and the Adam optimizer—contributed positively to model performance. Moreover, the absence of a substantial divergence between the training and validation metrics suggests that the proposed regularization strategy successfully mitigated overfitting, enabling the model to generalize well to previously unseen facial images. These findings are consistent with the study reported by Li et al. [18], which demonstrated that comprehensive image preprocessing and data augmentation improve training stability, accelerate CNN convergence, and enhance the robustness of facial emotion recognition models. Overall, the learning curves confirm that the proposed CNN framework achieved efficient optimization and established a solid foundation for the subsequent evaluation on the independent testing dataset.

3.2 Confusion Matrix Analysis

To further evaluate the classification performance of the proposed CNN model, a confusion matrix was employed to provide a detailed analysis of the prediction results for each emotion category. Unlike overall accuracy, which summarizes the model performance using a single metric, the confusion matrix reveals the distribution of correct and incorrect predictions across individual classes. This analysis provides valuable insight into the model's ability to distinguish between happy and fear facial expressions while identifying specific misclassification patterns that may occur during the recognition process. The confusion matrix enables the computation of several important evaluation metrics, including precision, recall, and F1-score, by quantifying the numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each emotion category. Consequently, it serves as an effective tool for assessing not only the overall classification performance but also the robustness and discriminative capability of the proposed facial emotion recognition model. In this study, the confusion matrix was generated using the validation dataset to examine the classification performance before evaluating the model on the independent testing dataset. The matrix illustrates the correspondence between the ground-truth emotion labels and the predicted labels generated by the CNN model, thereby highlighting both successful classifications and potential sources of classification errors. An ideal confusion matrix is characterized by large values along the main diagonal and minimal values in the off-diagonal elements, indicating that most facial images are assigned to their correct emotion categories.

The confusion matrix obtained from the validation dataset is presented in Fig. 6. It provides a comprehensive visualization of the classification results for the happy and fear emotion categories and serves as the basis for the subsequent analysis of precision, recall, and F1-score. Through this evaluation, the effectiveness of the proposed transfer learning framework in distinguishing facial emotions under unconstrained (in-the-wild) conditions can be thoroughly assessed.

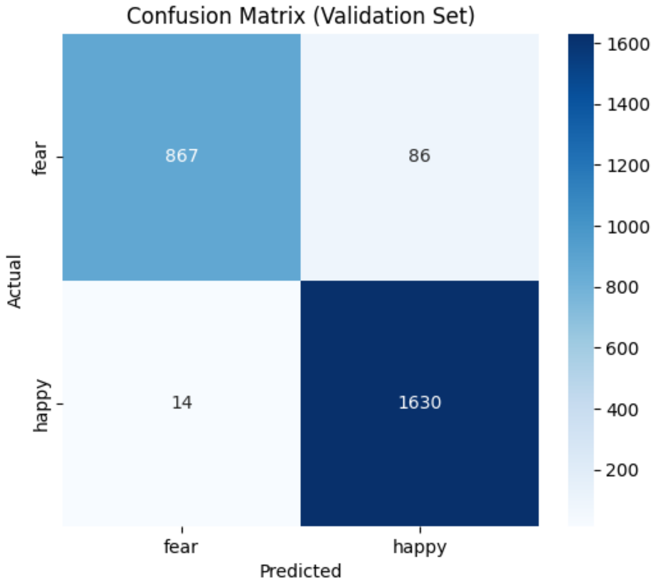


Figure 6. Confusing Matrix dari Dataset Affectnet Happy dan Fear

3.3 Model Evaluation on the Testing Data

To assess the generalization capability of the proposed CNN model, performance evaluation was conducted using an independent testing dataset that was not involved during either the training or validation stages. Evaluating the model on previously unseen data provides an objective measure of its ability to recognize facial emotions in practical scenarios and demonstrates whether the learned feature representations can be effectively generalized beyond the training samples. The classification performance was analyzed using a classification report, which summarizes four widely used evaluation metrics: precision, recall, F1-score, and support for each emotion category. Precision measures the proportion of correctly predicted samples among all predictions assigned to a particular class, whereas recall evaluates the model's ability to identify all samples belonging to the corresponding class. The F1-score, which represents the harmonic mean of precision and recall, provides a balanced assessment of classification performance, particularly when differences exist in the classification difficulty among emotion categories.

Table 3. Classification Report of the Proposed Model on the Testing Data

Class	Precision	Recall	F1-Score	Support
Fear	0.98	0.91	0.95	953
Happy	0.95	0.99	0.97	1644
Accuracy			0.96	2597
Macro avg	0.97	0.95	0.96	2597
Weighted avg	0.96	0.96	0.96	2597

As summarized in Table 3, the proposed CNN model achieved excellent classification performance on both emotion categories. For the fear class, the model obtained a precision of 0.98, recall of 0.91, and an F1-score of 0.95, based on 953 testing images. These results indicate that the model was highly reliable when predicting the fear emotion, as reflected by its high precision. However, the comparatively lower recall suggests that a small proportion of fear expressions were incorrectly classified as happy, indicating that certain fear-related facial characteristics remain challenging to distinguish because of their high visual variability. The happy emotion category achieved even stronger classification performance. The model obtained a precision of 0.95, recall of 0.99, and an F1-score of 0.97 from 1,644 testing images. The exceptionally high recall demonstrates that the CNN successfully recognized nearly all happy facial expressions, indicating that this emotion possesses distinctive visual characteristics that can be effectively captured by the NASNetMobile-based feature extractor. Typical facial features such as smiling lips, raised cheeks, and narrowed eyes contribute to the high separability of the happy class from other emotional categories.

4. Conclusion

This study successfully developed a Convolutional Neural Network (CNN)-based facial emotion recognition model using a transfer learning approach with the NASNetMobile architecture for classifying happy and fear emotions from the AffectNet dataset. The experimental results demonstrate that the proposed model achieved stable and effective learning throughout the training process, as evidenced by the reduction of the training loss from approximately 0.48 to 0.05 and the validation loss from 0.26 to 0.13. Simultaneously, the training accuracy increased to nearly 99%, while the validation accuracy reached 97%, indicating good convergence and no significant signs of overfitting. Evaluation on an independent testing dataset further confirmed the robustness of the proposed model, achieving an overall classification accuracy of 96% on 2,597 testing images. The model obtained an F1-score of 0.97 for the happy emotion and 0.95 for the fear emotion, demonstrating consistently high classification performance. The happy class achieved superior recognition accuracy because of its distinctive facial characteristics, whereas the fear class exhibited relatively lower performance due to greater intra-class variability and visual similarity with other facial expressions. Overall, the results demonstrate that the integration of comprehensive image preprocessing, data augmentation, transfer learning, and the NASNetMobile architecture substantially improve the effectiveness of CNN-based facial emotion recognition under unconstrained (in-the-wild) conditions. The proposed framework exhibits strong potential for practical applications in human–computer interaction, affective computing, intelligent surveillance, healthcare monitoring, and emotion-aware intelligent systems. Future research should extend the proposed framework by incorporating additional emotion categories available in the AffectNet dataset, investigating advanced attention-based and transformer-based deep learning architectures, and evaluating the model on more diverse real-world datasets to further enhance robustness, generalization capability, and practical deployment performance.

REFERENCES

[1]. A. K. S. Canal, T. A. S. R. S. M. G. P. G., and A. L. C. M. S. A. S., "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Information Sciences*, vol. 582, pp. 593–617, 2022, doi: 10.1016/j.ins.2021.10.005.

[2]. M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–31, 2023, doi: 10.1109/TIM.2023.3243661.

[3]. A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges," *Information*, vol. 13, no. 6, Art. no. 268, 2022, doi: 10.3390/info13060268.

[4]. N. Rathour, R. Singh, A. Gehlot, S. V. Akram, A. K. Thakur, and A. Kumar, "The decadal perspective of facial emotion processing and recognition: A survey," *Displays*, vol. 75, Art. no. 102330, 2022, doi: 10.1016/j.displa.2022.102330.

[5]. R. Kaur, S. S. Bhatia, and others, "Facial emotion recognition: A comprehensive review," *Expert Systems*, 2024, doi: 10.1111/exsy.13670.

[6]. N. N. Ali and A. M. Abdulazeez, "Facial emotion recognition based on deep learning: A review," *International Journal of Research and Applied Technology (INJURATECH)*, vol. 4, no. 1, pp. 21–34, 2024.

[7]. M. C. Gursesli, L. Sara, 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.

[8]. I. Na, A. Aldrees, A. Hakeem, L. Mohaisen, M. Umer, D. A. AlHammadi, S. Alsubai, N. Innab, and I. Ashraf, "FacialNet: Facial emotion recognition for mental health analysis using UNet segmentation with transfer learning model," *Frontiers in Computational Neuroscience*, vol. 18, Art. no. 1485121, 2024, doi: 10.3389/fncom.2024.1485121.

[9]. S. Porcu, A. Floris, and L. Atzori, "Evaluation of data augmentation techniques for facial expression recognition systems," *Electronics*, vol. 9, no. 11, Art. no. 1892, 2020, doi: 10.3390/electronics9111892.

[10]. Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, no. 4, pp. 3510–3519, 2021, doi: 10.1609/aaai.v35i4.16465.

[11]. J. R. Lee, L. Wang, and A. Wong, "EmotionNet Nano: An efficient deep convolutional neural network design for real-time facial expression recognition," *Frontiers in Artificial Intelligence*, vol. 3, Art. no. 609673, 2021, doi: 10.3389/frai.2020.609673.

[12]. D. Gera and S. Balasubramanian, "Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition," *Pattern Recognition Letters*, vol.

145, pp. 58–66, 2021, doi: 10.1016/j.patrec.2021.01.029.
[13]. S. Umer, R. K. Rout, C. Pero, and M. Nappi, "Facial expression recognition with trade-offs between data augmentation and deep learning features," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 2, pp. 721–735, 2021, doi: 10.1007/s12652-020-02845-8.
[14]. H. Siqueira, S. Magg, and S. Wermter, "Efficient facial feature learning with wide ensemble-based convolutional neural networks," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 34, no. 4, pp. 5800–5809, 2020, doi: 10.1609/aaai.v34i04.6037.
[15]. Y. Cai, J. Gao, G. Zhang, and Y. Liu, "Efficient facial expression recognition based on convolutional neural network," *Intelligent Data Analysis*, vol. 25, no. 1, 2021, doi: 10.3233/IDA-194965.
[16]. W. G. Colares, M. G. F. Costa, and C. F. F. Costa Filho, "Enhancing emotion recognition: A dual-input model for facial expression recognition using images and facial landmarks," in *Proc. IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–5, doi: 10.1109/EMBC53108.2024.10782924.
[17]. M. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computation in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2019, doi: 10.1109/TAFFC.2017.2740923.
[18]. "Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets," *Information*, vol. 15, no. 3, Art. no. 135, 2024, doi: 10.3390/info15030135.
[19]. M. I. Maulana, K. Aeni, and F. Fathulloh, "Pengenalan ekspresi wajah menggunakan Convolutional Neural Network (CNN)," *Indonesian Journal of Informatics and Research*, vol. 4, no. 2, 2023, doi: 10.58436/ijir.v4i2.1905.
[20]. A. S. Guntoro, E. Julianto, and D. Budiyanto, "Pengenalan ekspresi wajah menggunakan Convolutional Neural Network," *Jurnal Informatika Atma Jogja*, vol. 3, no. 2, pp. 155–160, 2022, doi: 10.24002/jiaj.v3i2.6790.
[21]. S. Sutarti and F. Syaqialloh, "Klasifikasi dan pengenalan emosi dari ekspresi wajah menggunakan CNN-BiLSTM dengan teknik data augmentation," *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 5, no. 1, 2025, doi: 10.51454/decode.v5i1.1038.