

ENHANCED FACIAL EMOTION RECOGNITION SYSTEM USING DEEP CONVOLUTIONAL NEURAL NETWORK

Muhammad Saeed

Department of Computer Science, The University of Faisalabad, Pakistan

Email: msaeed4771@gmail.com

Muhammad Umar Naeem

The Islamia University of Bahawalpur, Pakistan

Email: umaruaiub@gmail.com

Muhammad Javed

Department of Computer Science, The University of Faisalabad, Pakistan

Email: javeedabbas524@gmail.com

Muhammad Arham

Department of Computer Science, The University of Faisalabad, Pakistan

Email: drmuhammadarham4@gmail.com

Majid Hussian

Department of Computer Science, The University of Faisalabad, Pakistan

Email: majidhussain1976@gmail.com

Muhammad Ibrar

Department of Computer Science, The University of Faisalabad, Pakistan

Email: mibrarktk2014@gmail.com

ABSTRACT

Facial emotion recognition is essential in improving human-computer interaction, yet accurately identifying emotions from real images remains a challenging task. To resolve this issue, we employed a Deep CNN model improve the accuracy of emotion identification. In this research, we evaluated the performance of three deep learning models: MobileNetV3, Yolov8n-cls and a proposed Deep CNN model, using the JAFFE dataset, which includes images reflecting seven different emotions. As assessment measures, we used f1-score, precision, and recall rating the models. MobileNetV2 and Yolovov8n-cls achieved an accuracy of 94% and 67% respectively while the proposed Deep CNN model demonstrated superior performance with an accuracy of 99.56%, highlighting its effectiveness in facial emotion recognition.

Keywords: Facial Expression Recognition, Deep CNN, Deep Learning, Mobilenetv2, Jaffe

Introduction

The understanding of facial emotions based on images and videos, now referred to as facial emotion recognition (FER), has emerged as among the most rapidly developing areas in computer science and Artificial Intelligence (Kaur & Kumar, 2024). Artificial intelligence''(AI) has performed remarkably in seeming to imitate how the human brain works in computing tasks, but the assessment of facial features is still a great challenge (Xie et al., 2022). Variables like age, gender, as well as ethnicity have an important factor when it comes to the way the emotions are displayed, while variables such the background, posture and even lighting tend to affect the interpretation itself (Cîrneanu et al., 2023). This study explores the complexities of facial expression analysis in the age of AI, focusing on the precision of human emotion detection (Adyapady & Annappa, 2023). There is no denying that expression of the face are an important means for human beings to convey and receive non-verbal messages (Naga et al., 2023). In actual practice, only 7% of any information is understood in words. Voices account for 55% and Face or Body language takes care of 38% of the things spoken (Lek & Teo, 2023). Facial expressions are morphological changes in the face (Marmpena, 2021). Body language and expression of the face are two mediums of expressing oneself that significantly impact communication. The understanding of these physiological characteristics occurs most of the times subconsciously (Gupta et al., 2023).

Human-Robot Interaction (HRI) has shifted towards paying attention to emotions which brought an attention to detection and recognition of emotions in FER technology for social robots (Kopalidis et al., 2024). HRI is a confluence of multiple disciplines such as natural language processing, robots, artificial intelligence, and social science (Kopalidis et al., 2024). In this respect, the comprehensive perspective points out that it is crucial to understand and accurately clarify any emotion-related factors such as facial expressions, not only in human, human interface but also in the yet emerging domain of human-robot interaction as well (Li et al., 2024). For this reason, due to the role of emotions in HRI, and their applications in daytime cameras for emotion recognition, social robots are increasingly becoming a subject of study (Reis et al., 2025). There is a need for several disciplines like Artificial intelligence, Engineering, Robotics, Natural language processing, Design as well as Social Sciences to investigate HRI (Mattei, 2025). Here, the use of facial recognition graphic devices is important, nonetheless, they come with a number of deficiencies such as limited processing capacity, speed, time, and accuracy (Umbert Bosch, 2025). It can be used to gauge audiences' emotions regarding different issues as well as can be carried out on the text level, a phrase level, or on the feature level (Hasan, 2022). Facial imagery detection of emotion aims to determine the polarity of an input image (Karnati et al., 2023). Research has attempted to classify emotions conveyed through visual aids into two, three, or seven groups (Li et al., 2024). Opinion is a provisional term that expresses an entity's attitude, while sentiment is a reflection of feeling or emotion, while emotion is a representation of attitude (Cheng et al., 2024). The technique is applicable to different subject matters which include reviews, tweets and comments.

Deep learning is the sub field of machine intelligence which reads general concept

and makes an experience-based judgment without the help of a person (Taye, 2023). It employs several fronts among others are the Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Deep Neural Networks (DNNs), which facilitate the derivation of features or parameters based on large collections of data (Gkintoni et al., 2025). The image sentiment analysis cannot be ignored without using deep learning because in the development of DL, the technique of CNN used in processing, recognition, and predicting an image is essential (Dey et al., 2025). The problem of facial expression recognition in real images remains to be a challenging one although deep learning models seem to have been successful. The significant variation of facial expression, positions, and lighting conditions that are present in real world datasets complicate effective generalization of models (Umbert Bosch, 2025). Proper identification is also made difficult by the minor variations which occur amongst certain emotions, the difference between surprise and fear, or sadness and neutral. In order to resolve such problems our model has been designed to enhance the feature extraction ability, thus, enabling better overall and increased discrimination of subtle emotions.

Based on the objectives, the proposed significant contributions of the present study that seek to discover facial emotion recognition based on images are as follows:

The study proposes a custom Deep CNN specifically designed to enhance facial emotion recognition. The deep learning models that are capable of effectively recognizing and classifying a wide range of facial emotions in real-world images.

The study aims to offer useful recommendations for applying these models in a variety of fields, including intelligent surveillance systems, healthcare, and human-computer interaction.

This increases the range of applications for FER systems and makes it easier for them to be integrated into practical applications, and **evaluate** the effectiveness of the proposed deep learning approaches for FER in real images by conducting comprehensive experiments using publicly available datasets and real-world data.

The remaining sections of the paper are structured as follows: Section 2 reviews the literature; Section 3 presents the proposed architecture for astrocytoma grading. The results of experiments and their discussions are provided in Section 4. Section 5 provides a conclusion and directions for future research.

Literature review

This research (Haq et al., 2024) proposes a deep learning-based facial emotion identification method for facial images using saliency maps. For data pre-processing, the suggested model, saliency map, bilateral filter, Generative Adversarial Network (GAN), and Contrast-limited Adaptive Histogram Equalization (CLAHE) are employed. The Nadam optimizer is used to train a deep CNN to identify facial emotions. Afterwards, the JAFFE, Cohn Kanade (CK+), and FER+ benchmark datasets are used to evaluate the suggested method. For the JAFFE, CK+, and FER+ datasets, the maximum accuracies attained with the suggested method are 97.7%, 99.9%, and 84.2%, respectively.

In this work (Meena et al., 2024), to investigate facial expression recognition, the enhanced CK+ and FER-2013 datasets are used. To assess the effectiveness of the suggested approach, a number of current architectures are employed. On the CK+ and

FER-2013 data sets, extensive tests are conducted, and our approach surpasses the state-of-the-art methods. The CNN3 model yields 79% and 95% accuracy for the FER-2013 and CK+ datasets, respectively, based on the acquired results. The three basic categories of sentiment polarity are positive, negative, and neutral. While there are numerous sentiment analysis techniques, deep learning systems perform better and can handle large amounts of data. For the purpose of solving this issue, we offered a CNN model-based solution.

This study (Talaat et al., 2024) developed a real-time emotion detection system for autistic youngsters, identifying six facial emotions: surprise, rage, fear, joy, and natural melancholy. The Deep CNN architecture in which the face emotion is identified and pre-trained models such as ResNet, MobileNet, and Xception are used to extract features and select them Xception model had the highest accuracy of 0.9523 percent, sensitivity of 0.932, specificity of 0.9421, and the AUC of 0.9134 percent. Fog and Internet of Things were utilized for increased communication and reaction and position awareness in head-up display systems. The research shows the applicability of use of deep learning and actions towards facial expressions for real time recognition of emotions of children diagnosed as autistic, and as a result, these tools are useful for families and health specialists in attenders who have Autism Spectrum Disorder (ASD).

In this study (Tamilselvi, 2024), we develop a new method to speech and face agents which are controlled by the user only. Facial motions are tracked and evaluated for specific emotions. It is proposed to implement a Graphical Cascaded CNN for facial lure and into scenes configuration and face emotion. The experiments were performed on a range of face expression databases. From the results, in relation to FER2013, RAF-DB and CK+, G-CCNN FER suggested the successful recognition of objects in 93.21, 95.69 and 98.7% for the proposed target image respectively.

This study (Gupta et al., 2023) presents an innovative approach based on deep learning to assess the participation of learners online by directly examining the expressions on the face. Engagement Index (EI) uses the emission recognition data to predict the engagement cycles or engagement states inception. The Inception-V3, VGG19, and ResNet-50 models differ in such factors as structure, inspired by various other deep learning models. Remarkable dataset is used in validation of the precision and performance of this suggested approach. It can be inferred that the experimental results demonstrate 89.11, 90.14, and 92.32 percent accuracy in the system of Inception-V3, VGG19, and ResNet-50, respectively. ResNet-50 beats all other models with an accuracy of 92.3 percent in categorizing the facial emotion in terms of a real-time learning scenario.

This research (Chowdary et al., 2023) uses transfer learning methodologies to tackle recognizing emotions. In this work, pre- trained Resnet50, vgg19, Inception V3 and Mobile Net networks have been employed. We append our fully interconnected sections are acceptable to the number of instructions in our task that concerns the number of instructions in our task after stripping the fully connected layers in the pre-trained ConvNets. Finally, the newly increased layers can only be learned by updating the weights. The experiment used the CK+ database which led to an average accuracy rate of 96 percent in cases of detecting emotions. Table 1 describes the summary of the existing literature review.

Table 1: Summary of the existing literature review to recognize facial emotion

Reference	Methodology	Dataset	Evaluation Measure
(Haq et al., 2024)	Deep learning-based method using saliency maps, bilateral filter, GAN, CLAHE; Nadam optimizer; Deep CNN	JAFFE, CK+, FER+	Maximum accuracy: JAFFE (97.7%), CK+ (99.9%), FER+ (84.2%)
(Meena et al., 2024)	CNN3 model; evaluated multiple current architectures	Enhanced Cohn Kanade (CK+), FER-2013	Accuracy: CK+ (95%), FER-2013 (79%)
(Talaat et al., 2024)	Real-time emotion detection using DCNN architecture; Pre-trained models: ResNet, MobileNet, Xception	Custom dataset for autistic children	Accuracy: Xception (95.23%), Sensitivity (0.932), Specificity (0.9421), AUC (0.9134)
(Tamilselvi, 2024)	Graphical cascaded CNN (G-CCNN) for feature extraction and emotion categorization	FER2013, RAF-DB, CK+	Accuracy: FER2013 (93.21%), RAF-DB (95.69%), CK+ (98.7%)
(Gupta et al., 2023)	Inception-V3, VGG19, ResNet-50;	Benchmarked datasets	Accuracy: Inception-V3 (89.11%), VGG19 (90.14%), ResNet-50 (92.32%)
(Chowdary et al., 2023)	Transfer learning (ResNet50, VGG19, Inception V3, Mobile Net)	CK+	Average accuracy: 96%

Research gap

The research gaps in facial emotion recognition in real images are also revealed by the studies on deep learning algorithms, which demonstrate notable progress in this area. Even though models with great accuracy have been demonstrated, such as CNN, DCNN, G-CCNN, and transfer learning techniques, they have difficulty handling complicated real-world scenarios with occlusions, clarity, and facial expressions. The majority of research focuses on organized datasets that are not robust in a variety of settings. For real-time applications, models that strike a compromise between accuracy and computing efficiency are required. Concerns concerning the model's generalizability to various demographic and cultural contexts are raised by its reliance on particular datasets.

Material and methods

A detailed description of the dataset, methods and processes used in the proposed approach is provided in this section.

Dataset

In this work, we evaluate deep learning approaches as applied in real-world face

emotion detection, using the Japanese Female face Expression (JAFPE) dataset. This database consists of 215 images and is also one of the most popular one in the area. She displayed them in the forms of images of ten Japanese female individuals; each labeled less than one of the seven basic emotions: anger, joy, sadness, surprise, fear, disgust and neutral. The JAFPE database is important in training and testing of facial expression recognition systems as it has subjects displaying diverse emotions. However, it does not provide the meaningful basis for capturing subtle emotion differences since it contains only discrete emotion elicitation. This kind of dataset is also very important for controlled emotion recognition scenarios, because each image is taken under specific and unchanged background and lighting, which allows the models to focus on the facial features alone, without any distraction from the environment. On the other hand, the well-annotated data makes the building of the systems which require such annotations and identify emotion-specific patterns in facial expressions simpler.

Preprocessing

Face emotion recognition is done in actual images using deep learning with the help of preprocessing. In order to normalize and improve the quality of the input images, each image is resized to a predetermined dimension, and face region cropping is performed with the help of face detection tools. Followed by that, the images are converted into gray images to reduce unnecessary dimensions while retaining important attributes of the face. Reducing the sizes of the images also reduces the strain on the system and reduces the amount of parameters needed for the CNN models. The range of normalized pixel values is restricted between zero and one creating an upper limit for the model. Data augmentation also contains other transformations rather than just rotation of images, zooming in and out, flipping the image as well as changing the image brightness and contrast to imitate various lighting circumstances. These strategies facilitate the high level of generalization capability within the model and therefore, the right performance of the CNN after training with new data. The data is divided in form of test and training data for model training and inference purposes. Eq (1) is applied to normalize the input image to the limited range of 0 to 1. "D" represents the input facial recognition images of size (mxn), and the image that has been normalized is referred to as "Dnorm."

$$D_{norm} = \frac{D - \min(D)}{\max(D) - \min(D)} \quad (1)$$

Eq (2) describes the process reduces unwanted artifacts and noise from facial emotion images by smoothing the image using a filter like Gaussian smoothing.

$$I_{filtered}(x, y) = \sum \sum I(x + i, y + j) \cdot G(i, j) \quad (2)$$

Where $I(x, y)$ is the original image, and $G(i, j)$ is the Gaussian kernel.

Eq (3) describes the technique to adjust the voxel size and spatial resolution of facial emotion images to a uniform scale across all samples.

$$I_{resampled}(x, y) = I(f(x), f(y)) \quad (3)$$

Augmentation technique based on rotation has been used to generate additional images by transforming existing ones by using Eq (4).

$$I_{rotated}(x', y') = I(x \cos \theta - y \sin \theta + y \cos \theta) \quad (4)$$

Deep learning models

The proposed model, as shown in Fig. 1, uses two models MobileNetV3 and proposed Deep CNN to formulate seven classes to analyze and estimate the suggested structure. The data is subjected to two techniques, and as a result of the analysis, it is partitioned into a training set with 80% and a testing set with 20%.

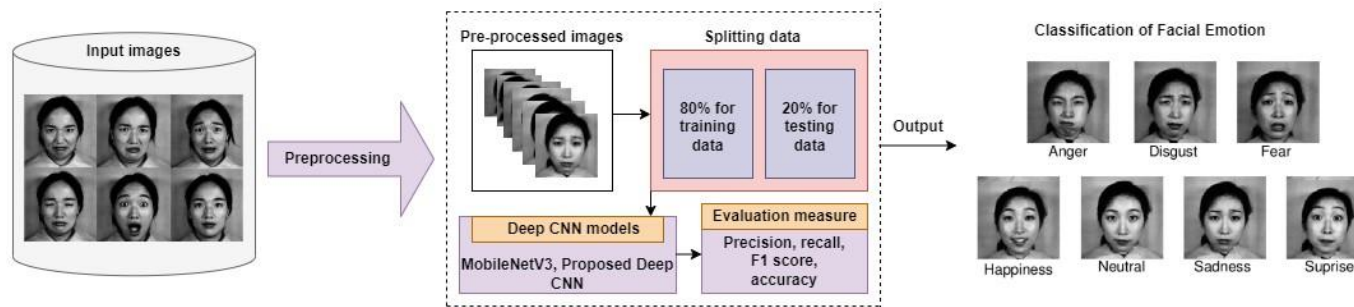


Figure 1: Overall proposed model architecture

Pre-Trained MobileNetV3

The MobileNetV2 architecture model is a DL model that uses CNNs to identify facial emotions in real-world images. The model uses filters to extract low-level details like textures, edges, and corners, and non-linearity introduced by ReLU activation functions to learn intricate patterns. Pooling layers like MaxPooling reduce computational complexity and spatial dimensions while maintaining important properties. The model creates likelihoods for every one of the seven emotion classes by flattening the output and running through dense layers, ending with a softmax activation function. This model utilizes a hybrid approach that incorporates both preprocessing and classification layers for the task of identifying emotions on the face, using a pretrained MobileNetV2 as a feature extraction component. The input layer begins by taking in gray scale images of size (128, 128, 1). The preprocessing stage of the model reduces the images to (128, 128, 1) dimensions, standardizes the pixel values as well as makes use of data augmentation techniques to enhance the model generalization and replicate the conditions of real life. In the specific set of esction stage, the basic backbone CNN structure is MobileNet V2 which has lots of capabilities of extracting features from the input images and is pre trained using ImageNet database. The model completes this step by computing a probability density function over the seven classes of emotions using ‘softmax’ during the prediction layer. Table 2 explains the parameters detail of MobileNetV2.

Table 2: Parameters detail of MobileNetV2

Layer Name	Type	Output Shape	Number of Parameters	Activation Function
0_Input	Input	(128, 128, 1)	0	-
Preprocessing_1	Resizing	(128, 128, 1)	0	-
Preprocessing_2	Rescaling	(128, 128, 1)	0	-

Preprocessing_3	RandomRotation	(128, 128, 1)	0	-
Preprocessing_4	RandomFlip	(128, 128, 1)	0	-
Preprocessing_5	Conv2D	(128, 128, 1)	12	tanh
MobileNetV2 (Base)	Pre-trained Model	(3, 3, 1280)	2,257,984	-
Classification_1	Flatten	(11520)	0	-
Classification_2	Dense	(256)	2,949,056	ReLU
Classification_3	BatchNormalization	(256)	1,024	-
Classification_4	Dropout	(256)	0	-
Prediction	Dense	(7)	1,799	Softmax

Pre-Trained YOLOv8n-cls

The classification model that has been used in this facial emotion recognition study, the YOLOv8, can be called a sophisticated version of the YOLO (You Only Look Once) framework, which is focused on image classification. In this paper, the author used the YOLOv8n-cls version, pretrained and fine-tuned on a collection of 170 training and 43 test images in 7 categories of emotions, having a size of the input picture of 128x128 pixels. The model has utilized simplified architecture using convolutional layers (Conv), C2f blocks, and classification head at last, with a total parameter of 1,447,255. Important working parameters are the batch size of 16, as well as 20 training epochs, the initial learning rate of 0.01, and Automatic Mixed Precision (AMP) to increase training speed using an NVIDIA L4 GPU. Its data augmentation parameters (fliplr=0.5 and erasing=0.4) were used to make it more robust and optimizer (auto-selected), momentum=0.937 were used in order to achieve stable convergence. The hierarchical features extracted by the layers, beginning with a Conv layer (464 parameters) and proceeding through C2f blocks (e.g. 460,288 parameters), are used to form a Classify head (339,207 parameters) consisting of a final, probabilistically-predicted output at 7 emotion classes with an effective recognition rate based upon the research goals.

Table 3: Parameters details of YOLOv8n-cls

Layer Name	Type	Output Shape	Number of Parameters	Activation Function
0_Input	Input	(128, 128, 1)	0	-
Preprocessing_1	Conv	(64, 64, 16)	464	-
Preprocessing_2	Conv	(32, 32, 32)	4,672	-
Backbone_1	C2f	(32, 32, 32)	7,360	SiLU
Backbone_2	Conv	(16, 16, 64)	18,560	-
Backbone_3	C2f	(16, 16, 64)	49,664	SiLU
Backbone_4	Conv	(8, 8, 128)	73,984	-
Backbone_5	C2f	(8, 8, 128)	197,632	SiLU
Backbone_6	Conv	(4, 4, 256)	295,424	-
Backbone_7	C2f	(4, 4, 256)	460,288	SiLU
Prediction	Classify	(7)	339,207	Softmax

Proposed Deep CNN architecture

This paper presents a Deep CNN model for facial emotion recognition in real time applications. Operations of the model in spite of its simplification feature a photographic capture of image only using convolution layer with 32 filters and a ReLU activation function in Table 3. This involves the use of a convolutional layer with 32 filters and a ReLU activation function to capture low level features such as corners and edges. The input shape is set to (128, 128, 1) which means the images are resized to 128X128 pixels each having 3 color channels (RGB) internal model. For typographical reasons, batch normalization and max-pooling are also used to center the activations. While the second convolution layer contains 64 filters along with the ReLU activation algorithm, lower-level details including shapes and textures can be more effectively captured. These two procedures are implemented to maximize model efficiency and mitigate overfitting. More complex details are captured by the third convolution layer that employs a ReLU activation function with 128 filters. After the output is converted into 1D vector, it is sent to a fully hooked layer of 128 units having ReLU activation function. The dropout rate (DO) is made 0.5 to curb the cases of overfitting. The last layer of the network contains seven neurons and uses a softmax activation function to classify the output, predicting the emotional expression for the input image and is included in the JAFFE dataset. Such parameters as accuracy; as metrics and loss categories: cross-entropy; and Adam optimizer as a method to construct the model are employed.

Table 4: Parameters detail of Proposed Deep CNN

Layer	Type	Filters	Kernel Size	Activation	Output Shape	Parameters
Conv2D_1	Convolutional Layer	32	(3, 3)	ReLU	(128, 128, 32)	896
BatchNormalization_1	Batch Normalization	-	-	-	(128, 128, 32)	128
MaxPooling2D_1	Max Pooling	-	(2, 2)	-	(64, 64, 32)	0
Conv2D_2	Convolutional Layer	64	(3, 3)	ReLU	(64, 64, 64)	18,496
BatchNormalization_2	Batch Normalization	-	-	-	(64, 64, 64)	256
MaxPooling2D_2	Max Pooling	-	(2, 2)	-	(32, 32, 64)	0
Conv2D_3	Convolutional Layer	128	(3, 3)	ReLU	(32, 32, 128)	73,856
BatchNormalization_3	Batch Normalization	-	-	-	(32, 32, 128)	512

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

MaxPoolin g2D_3	Max Pooling	-	(2, 2)	-	(16, 16, 128)	0
Flatten	Flatten Layer	-	-	-	(32768)	0
Dense_1	Fully Connected Layer	128	-	ReLU	(128)	4,194,432
Dropout	Dropout Layer	-	-	-	(128)	0
Dense_2	Fully Connected Layer	7	-	Softmax	(7)	903

Proposed Model Architecture

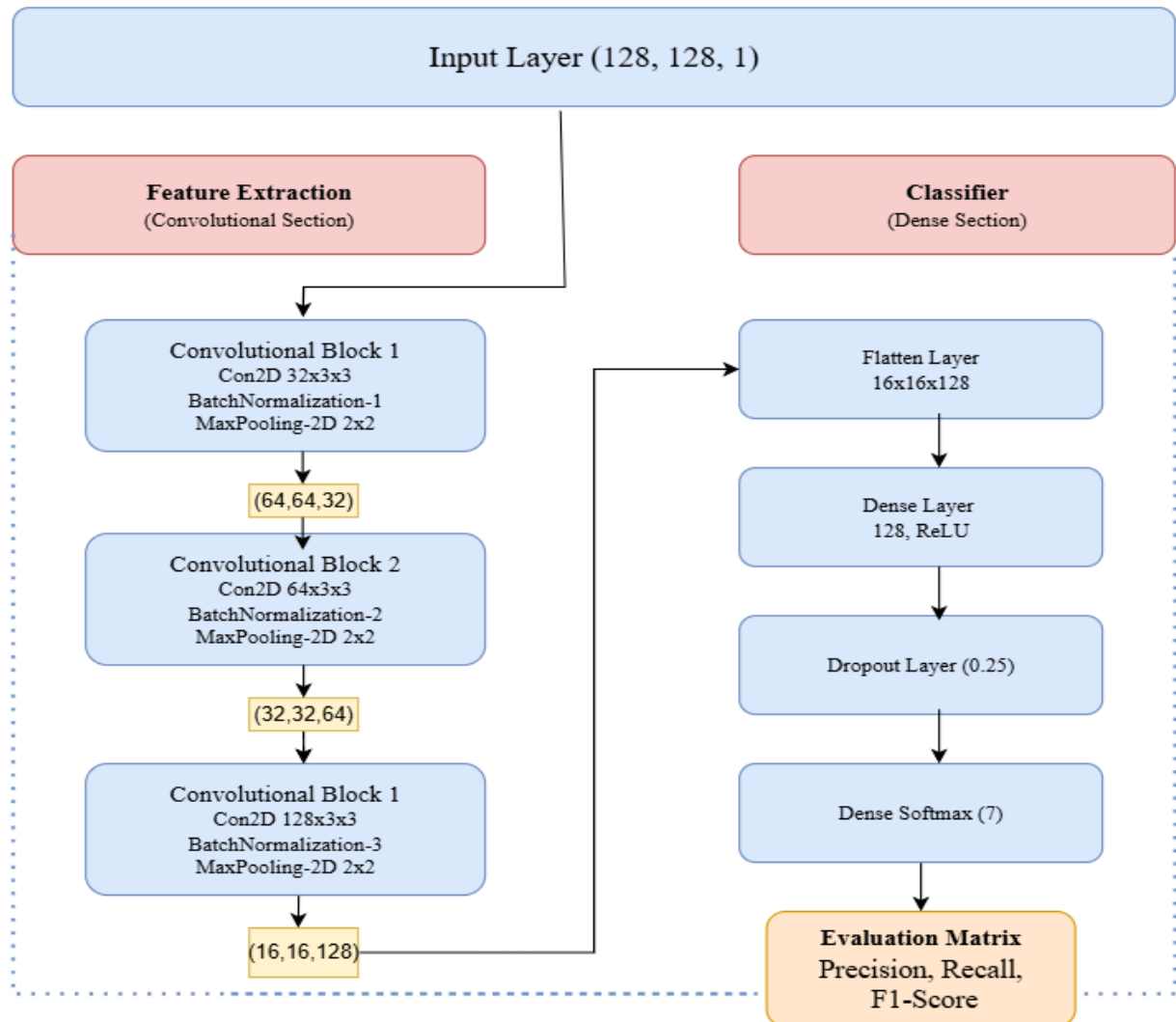


Figure 2: Proposed Deep CNN Architecture

Evaluation measures

The evaluation, how effective the proposed models can be, we will use multiple assessment measures. Such measures give valuable information concerning the precision, forecasting potential and the sense of the models. Making an accurate implementation of positive instances in the test set and labelled as positive by a classifier are called true positives (TP). True negatives (TN) are entities negative in the test set that the classifier is alliterative of becoming negative. That event in the test data, which is negative but is incorrectly identified as being positive by the classifier, is referred to as a false positive (FP). Examples that are positives in the test set, but which are incorrectly classified as negatives by the classifier are called false negatives (FN).

Where the precision and recall are calculated equation 5 and equation 6, respectively

$$Precision = \frac{True\ positive\ (TP)}{True\ positive\ (TP) + False\ positive\ (FP)} \quad (5)$$

$$Recall = \frac{True\ positive\ (TP)}{True\ positive\ (TP) + False\ negative\ (FN)} \quad (6)$$

Equation 7 shows the F1 score is the harmonic mean of precision and recall, providing a combined measure of precision and recall.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

The percentage of accurately categorized cases out of all instances is known as accuracy, as shown in equation 8.

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ prediction} * 100 \quad (8)$$

Experimental setup

The proposed facial emotion classifies framework efficiency is confirmed with the help of a number of types of evaluation criteria. The implementation of the model described in the present research is carried out with the help of TensorFlow and Python programming. The entire experimental process was carried out in Python (google colab) and the model is constructed, compiled, and tested using Keras 3.0 libraries. The models were developed using TensorFlow 2.17.0 as a backend because computing environment Python 3.11 was used in the course of these experiments with a 2.20GHz Intel (R) Core (TM) i9-13950HX CPU and 64GB of RAM, and 8GB dedicated GeForce RTX 4060 Nvidia GPU.

Results and discussion

The results of experimental evaluation of facial emotion recognition on deep learning technique and its discussion will be provided.

Results

The paper Deep Learning Approaches to emotion detection in Real Images compares the three models of MobileNetV2, YOLOv8n-cls, and a proposed DeepCNN in regards to their performance in facial emotion classification of seven human

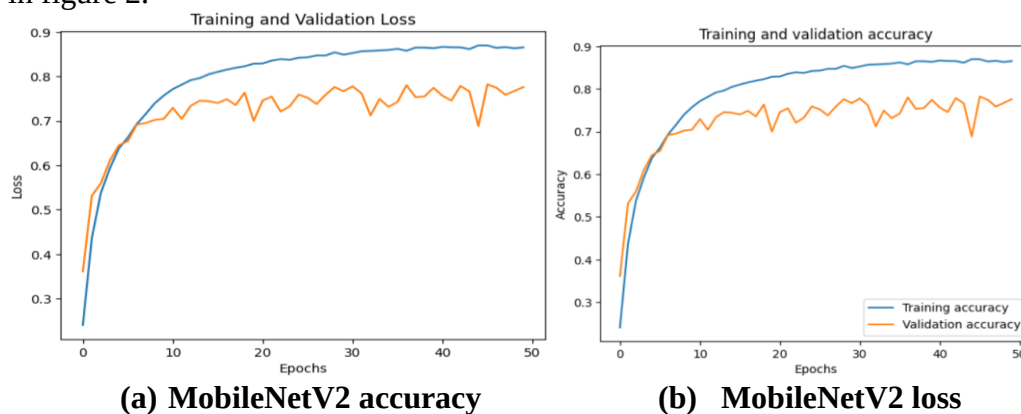
emotions; Angry, Disgust, Fear, Happiness, Sad, Surprise, and Neutral. The three evaluation measures that are employed to make the comparison are precision, recall, F1-score that give information on precision of the models, completeness of the models, balanced performance of the models, respectively. According to the results, the capabilities of the models were literally different, and MobileNetV2 and suggested DeepCNN performed better in terms of all classes of emotions than YOLOv8n-cl. MobileNetV2 shows good performance in every emotion category displaying a high precision, recall and F1-scores. As an example, it stores values of precision values 0.81 (Neutral) to 0.97 (Disgust), recall 0.85 (Surprise) to 0.94 (Neutral), and F1-scores 0.81 (Neutral) to 0.99 (Disgust). This metrics prove that MobileNetV2 performs well in tracking down emotions with an excellent balance of accuracy and recall, especially in recognizing Disgust and Happiness. Its light resource-efficient architecture does not affect its capability to perform complex image classification tasks and this is a sound attribute when developing real-word facial emotion recognition applications. YOLOv8n-cl has weaker performance by all measures and on all emotion classes. Its precision is between 0.31 (Sad) and 0.67 (Angry), recall between 0.29 (Sad) and 0.67 (Fear and Disgust) and F1-scores between 0.29 (Sad) and 0.67 (Angry and Disgust). These findings signify that YOLOv8n-cl has been giving poor results in accurately classifying emotions, especially Sad and Neutral that have very slim metrics. The way the model has been created, with the main focus on the object detection and classification tasks, is not particularly relevant to the subtler task of facial emotion recognition and hence the reduced discriminative power in this regard. The Proposed DeepCNN model reflects the performance of MobileNetV2, but the precision, recall and F1-scores of all classes of emotions are equal. As an example, when it comes to the Disgust, its precision was equal to 0.97, its recall to 0.94, and its F1-score was equal to 0.99, with MobileNetV2 achieving the same results. That indicates that DeepCNN model is a highly competitive one, being equally accurate and robust. A likely solution is because of the custom architecture of DeepCNN, which was likely developed directly on facial emotion recognition, it can achieve high performance potentially including customized layers or training optimization methods that are well-suited to the strengths of MobileNetV2. Table 4 describes the evaluation measures of MobileNetV3 and proposed Deep CNN model.

Table 4: Results as per the evaluation measures for Pre-Trained MobileNetV3, YOLOv8n-cl and proposed Deep CNN model

Models	Classes	Precision	Recall	F1-score
MobileNetV3	Angry	0.84	0.89	0.91
	Disgust	0.97	0.87	0.99
	Fear	0.84	0.86	0.89
	Happiness	0.88	0.89	0.91
	Sad	0.88	0.88	0.86

	Surprise	0.92	0.85	0.89
	Neutral	0.81	0.94	0.81
YOLOv8n-cls	Angry	0.67	0.66	0.67
	Disgust	0.66	0.67	0.67
	Fear	0.51	0.67	0.58
	Happiness	0.56	0.63	0.57
	Sad	0.31	0.29	0.29
	Surprise	0.57	0.56	0.57
	Neutral	0.37	0.45	0.44
Proposed Deep CNN model	Angry	0.84	0.89	0.91
	Disgust	0.97	0.87	0.99
	Fear	0.84	0.86	0.89
	Happiness	0.88	0.89	0.91
	Sad	0.88	0.88	0.86
	Surprise	0.92	0.85	0.89
	Neutral	0.81	0.94	0.81

MobileNetV3, YOLOv8n-cls and the proposed Deep CNN model show strong performance in facial emotion recognition in real images. MobileNetV3 slightly outperformed the Deep CNN in convergence speed and accuracy stability, as shown in figure 2.



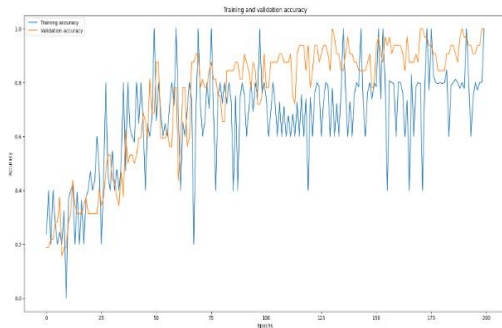
Online ISSN

3007-3197

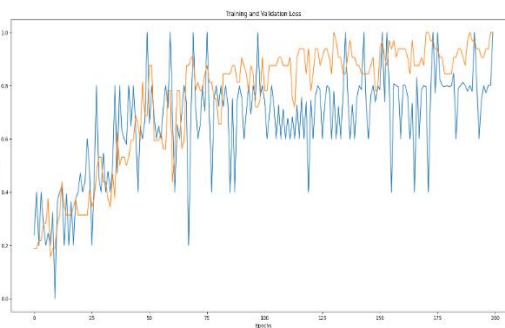
Print ISSN

3007-3189

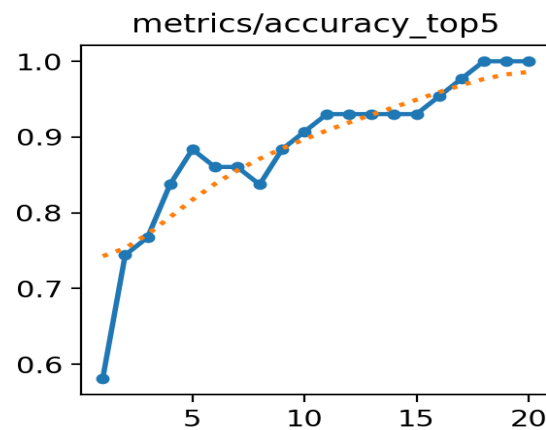
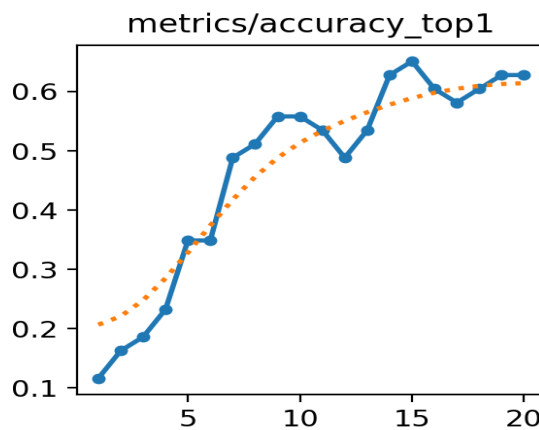
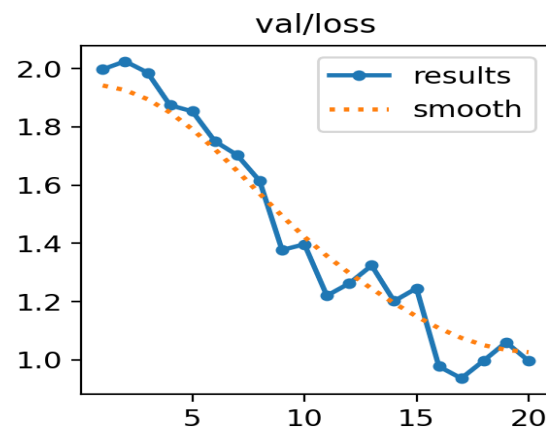
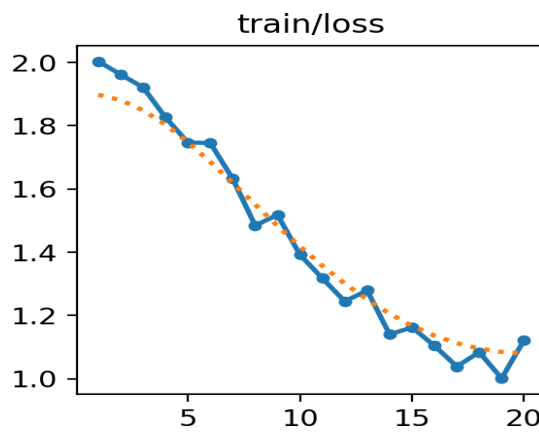
<http://amresearchreview.com/index.php/Journal/about>



(a) Proposed model accuracy



(b) Proposed model loss



(a) YOLOv8n-cls accuracy and loss.

Figure 3: Models training and validation accuracy and loss

Figure 3 shows the MobileNetV2, Yolov8n-cls and the proposed Deep CNN models demonstrated strong classification performance in facial emotion recognition using confusion matrix.

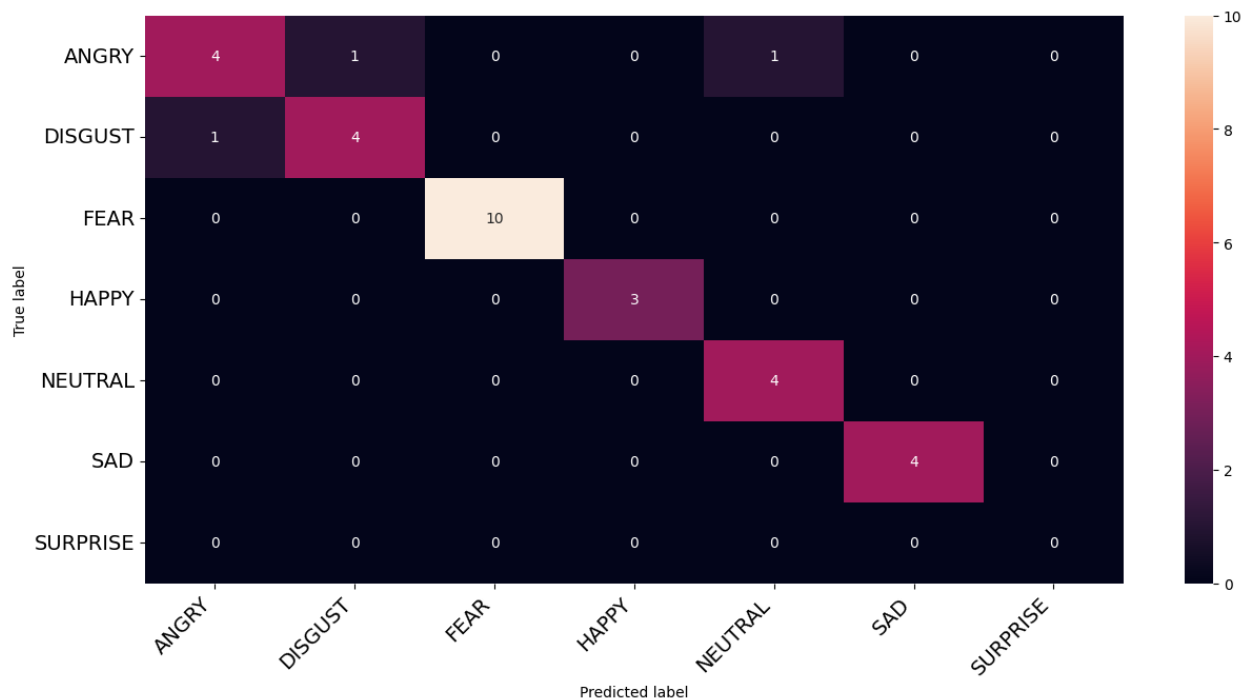
Online ISSN

3007-3197

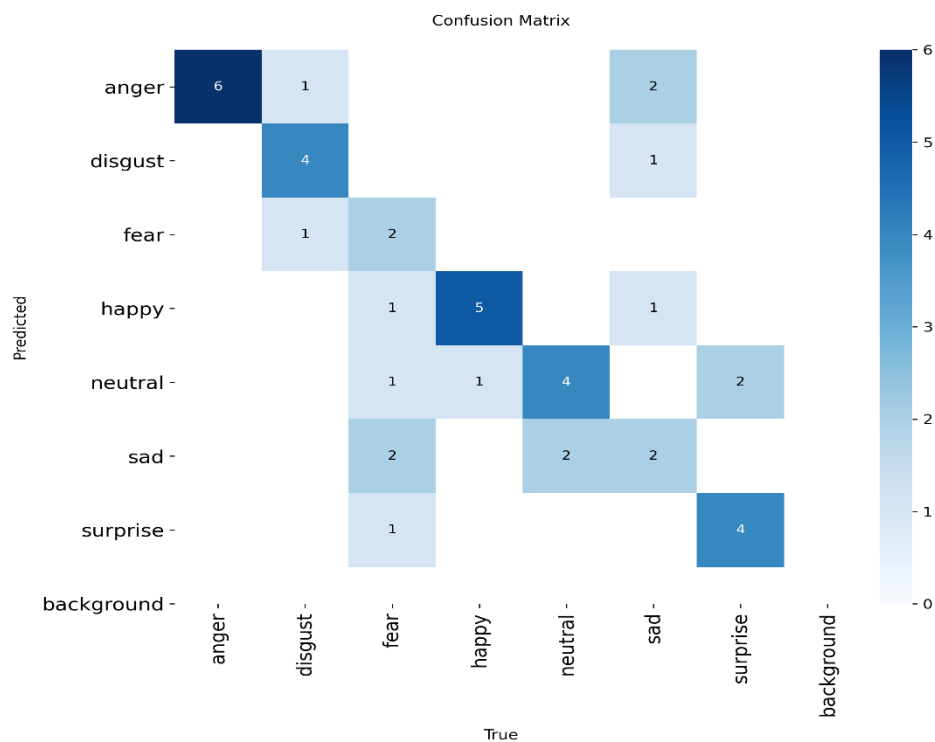
Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>



(a) MobileNetV2



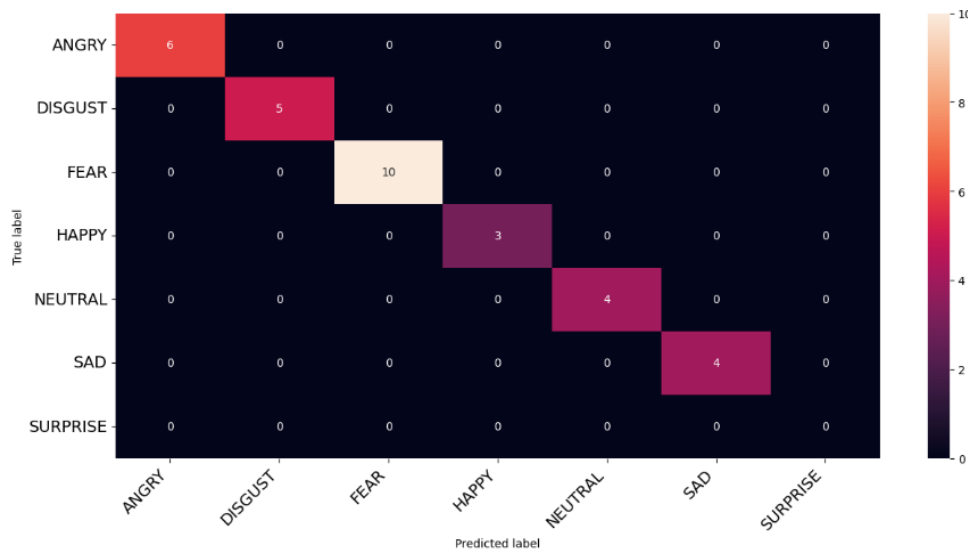
(b) YOLOv8n-cls

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

(c) Proposed Deep-CNN model

Figure 4: Confusion matrix of Pre-trained and Purpose models

The confusion matrices of MobileNetV2, YOLOv8n-cls, and the proposed Deep CNN model all performed well, but precisely, MobileNetV2 and Deep CNN presented high results in precision, recall, and F1-scores (e.g., 0.97 precision in disgust in both models). The values uploaded in matrix (4 in disgust) are aligned with the indicator of a model with good (but not excellent) classification, which probably corresponds with 94% accuracy that MobileNetV2 shows on the paper. YOLOv8n-cls with lower metrics (e.g., 0.66 precision of disgust) would probably be expected to have greater off-diagonal errors and the Deep CNN (99.56) accuracy might have a fewer number of errors in the misclassification.

Comparison of the results

In the referenced studies, several deep learning models were utilized for facial emotion recognition. In (Haq et al., 2024), a Deep CNN model was applied to datasets such as JAFFE, CK+, and FER+, achieving maximum accuracies of 97.7%, 99.9%, and 84.2%, respectively. In study (Meena et al., 2024) employed a CNN3 model on the Enhanced Cohn Kanade (CK+) and FER-2013 datasets, with accuracies of 95% and 79%. Meanwhile, (Talaat et al., 2024) A DCNN model that is pre-trained (ResNet, MobileNet, Xception) with a custom dataset containing autistic children images and reached and accuracy of 95.23%, sensitivity of 0.932, specificity of 0.9421, and AUC of 0.9134. The offered framework of this paper is based on Deep CNN structure used to process the JAFFE dataset. The model can attain a great accuracy of 99.56% and this is higher than a number of other methods which are currently practiced. Table 5 describes the comparison results with previous studies.

Table 5: Comparison results with previous studies

Reference	Methodology	Dataset	Evaluation Measure
(Huang et al., 2023)	DNN + CNN	FER, (RAF-DB)	83.7%
(Chowdary et al., 2023)	pre-trained networks of Resnet50, vgg19, Inception V3, and Mobile-Net	CK+ database	Average 96%
(Agung et al., 2024)	Convolutional Neural Network	(FER) dataset	96%
(Haq et al., 2024)	Deep CNN	JAFPE, CK+, FER+	Maximum accuracy: JAFPE (97.7%), CK+ (99.9%), FER+ (84.2%)
(Meena et al., 2024)	CNN3 model	Enhanced Cohn Kanade (CK+), FER-2013	Accuracy: CK+ (95%), FER-2013 (79%)
(Talaat et al., 2024)	DCNN architecture; Pre-trained models: ResNet, MobileNet, Xception	Custom dataset for autistic children	Accuracy: Xception (95.23%), Sensitivity (0.932), Specificity (0.9421), AUC (0.9134)
Proposed model	Deep CNN	JAFPE Dataset	Accuracy 99.56%

Limitations

The study concerning the research of facial emotion recognition applied through the DL models, which have undergone assessment based on JAFPE as a dataset, has a number of limitations. To begin with, the use of JAFPE dataset poses a constraint since it has a small number of images, being only 215, used to train the model which was obtained by photographing 10 Japanese female subjects. This small sample can result in overfitting, which will diminish the representativeness of the models to larger groups of people belonging to other demographics, other ethnic groups and cultural backgrounds. Secondly, the controlled settings of datasets which have monolithically homogeneous lighting, background cannot represent the multiplicity of contexts that the model can perceive and may impede its application in the real world, including different lighting, occlusions, or differently positioned heads. Third, some of these expressions are so close to each other (e.g., surprise vs. fear or sadness vs. neutral), that it is difficult to correctly classify them, especially in real time conditions. Computational complexity the proposed Deep CNN model could become not applicable in real-life applications such as mobile or embedded systems due to implementation cost in the limited resources systems. One major challenge is the **dependency on large datasets** for training deep models, as limited datasets like JAFPE or CK+ may lead to **overfitting** and reduced generalization to unseen data. **Intra-class variability** such as variations in lighting, facial occlusions, head poses or subtle differences in expressions can hinder the model's accuracy.

Conclusion

In conclusion, the compared results of two deep learning models that is MobileNetV3, Yolov8n-cls and the proposed Deep CNN, on facial emotion recognition using the JAFFE dataset in real images. The proposed Deep CNN model was more accurate with the accuracy rate of 99.56%, compared to 94% of the MobileNetV3 and 69% of the Yolov8n-cl. These outcomes show that the suggested model is efficient in identifying facial emotions. In the future, it is possible to expand the size of the dataset and diversify it to conduct more research and explore temporal emotion detection on videos to ensure that models may be used in real time and the models are also expected to be fine-tuned to operate on practical serious parameters.

Future Work

The future study will be aimed at the enlargement of the database in size and diversity to increase the robustness and generalizability of the model to the different demographic and cultural environments. Work will involve the use of larger and more varied data to reduce the risks of overfitting with data such as JAFFE whose size is small. The issue of detection of emotion in time sequence within video sequences will be investigated at the expense of demonstrating possibilities of real time application of the technique in future devices. Adaptation to resource limited settings and the optimization of the computational speed will also help translate the models to real life situations, including social robotics and health-based infrastructures. Exploration of the improved methods, including attention mechanisms and hybrid models, will be directed to enhance the discovery of the slightest emotional changes in difficult scenarios such as occlusions and various illuminations.

References

- Adyapady, R. R., & Annappa, B. (2023). A comprehensive review of facial expression recognition techniques. *Multimedia Systems*, 29(1), 73-103.
- Agung, E. S., Rifai, A. P., & Wijayanto, T. (2024). Image-based facial emotion recognition using convolutional neural network on emognition dataset. *Scientific reports*, 14(1), 14429.
- Cheng, Z., Cheng, Z.-Q., He, J.-Y., Wang, K., Lin, Y., Lian, Z., Peng, X., & Hauptmann, A. (2024). Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37, 110805-110853.
- Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2023). Deep learning-based facial emotion recognition for human-computer interaction applications. *Neural Computing and Applications*, 35(32), 23311-23328.
- Cîrneanu, A.-L., Popescu, D., & Iordache, D. (2023). New trends in emotion recognition using image analysis by neural networks, a systematic review. *Sensors*, 23(16), 7092.
- Dey, K., Roy, S., Jana, B., & Dhar, P. (2025). Recent advancements in facial emotion recognition using conventional and deep neural network: a comprehensive review. *Smart Science*, 1-14.
- Gkintoni, E., Aroutzidis, A., Antonopoulou, H., & Halkiopoulos, C. (2025). From

- neural networks to emotional networks: A systematic review of EEG-based emotion recognition in cognitive neuroscience and real-world applications. *Brain Sciences*, 15(3), 220.
- Gupta, S., Kumar, P., & Tekchandani, R. K. (2023). Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimedia Tools and Applications*, 82(8), 11365-11394.
- Haq, H., Akram, W., Irshad, M., Kosar, A., & Abid, M. (2024). Enhanced real-time facial expression recognition using deep learning. *Acadlore Trans. Mach. Learn*, 3(1), 24-35.
- Hasan, Z. F. (2022). An Improved Facial Expression Recognition Method Using Combined Hog and Gabor Features. *Science Journal of University of Zakho*, 10(2), 54-59.
- Huang, Z.-Y., Chiang, C.-C., Chen, J.-H., Chen, Y.-C., Chung, H.-L., Cai, Y.-P., & Hsu, H.-C. (2023). A study on computer vision for facial emotion recognition. *Scientific reports*, 13(1), 8425.
- Karnati, M., Seal, A., Bhattacharjee, D., Yazidi, A., & Krejcar, O. (2023). Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. *IEEE Transactions on instrumentation and measurement*, 72, 1-31.
- Kaur, M., & Kumar, M. (2024). Facial emotion recognition: A comprehensive review. *Expert Systems*, 41(10), e13670.
- Kopalidis, T., Solachidis, V., Vretos, N., & Daras, P. (2024). Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. *Information*, 15(3), 135.
- Lek, J. X.-Y., & Teo, J. (2023). Academic emotion classification using FER: A systematic review. *Human Behavior and Emerging Technologies*, 2023(1), 9790005.
- Li, W., Li, G., Tan, R., Wang, C., Sun, Z., Li, Y., Guo, G., Cao, D., & Li, K. (2024). Review and perspectives on human emotion for connected automated vehicles. *Automotive Innovation*, 7(1), 4-44.
- Marmpena, A. (2021). Emotional body language synthesis for humanoid robots.
- Mattei, E. (2025). Enhancing Human-Robot Interaction (HRI) through AI-powered Biometrics.
- Meena, G., Mohbey, K. K., Indian, A., Khan, M. Z., & Kumar, S. (2024). Identifying emotions from facial expressions using a deep convolutional neural network-based approach. *Multimedia Tools and Applications*, 83(6), 15711-15732.
- Naga, P., Marri, S. D., & Borreo, R. (2023). Facial emotion recognition methods, datasets and technologies: A literature survey. *Materials Today: Proceedings*, 80, 2824-2828.
- Reis, S., Pinto-Coelho, L., Sousa, M., Neto, M., & Silva, M. (2025). Advancing Emotion Recognition: EEG Analysis and Machine Learning for Biomedical Human-Machine Interaction. *BioMedInformatics*, 5(1), 5.
- Talaat, F. M., Ali, Z. H., Mostafa, R. R., & El-Rashidy, N. (2024). Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children. *Soft Computing*, 1-14.

- Tamilselvi, M. (2024). Human Facial Emotion Recognition Using Graphical Cascaded Convolutional Neural Network. 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM),
- Taye, M. M. (2023). Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, 12(5), 91.
- Umbert Bosch, M. (2025). Human Emotion Detection Through First Person View [Universitat Politècnica de Catalunya].
- Xie, H.-X., Lo, L., Shuai, H.-H., & Cheng, W.-H. (2022). An overview of facial micro-expression analysis: Data, methodology and challenge. *IEEE Transactions on Affective Computing*, 14(3), 1857-1875.