OPEN ACCESS

International Journal of Innovations in Science & Technology

RESEARCH & INNOVATION DIVISION

IJIST

Modified Convolutional Neural Networks for Facial Emotion Classification

Sobia Yousaf[1], Saiqa Anjum[1], Nimra Ibrar[2], Ruqia Bibi[*3], Muhammad Adeel Asghar[4]

[1]Department of Software Engineering, National University of Modern Languages, Rawalpindi, Pakistan.

[2]University of Roehampton, United Kingdom

[3]University Institute of Information Technology, PMAS Arid Agriculture University, Rawalpindi, Pakistan

[4]Department of Computer Science, National University of Modern Languages, Rawalpindi, Pakistan.

[*]Correspondence: ruqia.bibi@uaar.edu.pk.

Citation |Yousaf S, Anjum S, Ibrar N, Bibi R, Asghar A M "Modified Convolutional Neural Networks for Facial Emotion Classification", IJIST, Vol 6, Issue 4. pp 1897-1912, Nov 2024

Received | Oct 21, 2024; Revised | Nov 18, 2024 Accepted | Nov 19, 2024; Published | Nov 20, 2024.

Facial expression analysis is a fascinating yet challenging problem in the realm of artificial intelligence. The vast variability in human expressions poses a significant hurdle for machine learning methods to detect them accurately. Recently, machine learning and deep learning approaches have made notable strides in this area, leveraging Deep Neural Networks (DNNs) to identify human emotions. Convolutional Neural Networks (CNNs), in particular, have proven effective in resolving the complexities involved in human facial expressions, making them a preferred choice for these tasks. In this study, we proposed a modified CNN architecture by introducing a new layer to enhance accuracy. The CNN network is trained on both frontal face images and images with varying poses. We utilized three distinct datasets FER 2013, CK+ and our own dataset to achieve the desired results. The evaluation results obtained using the proposed network surpass those achieved by conventional CNN networks. Notably, our proposed network achieves an average accuracy of 97.5% on our collected dataset.

Keywords: Facial Expression Recognition, Facial Action Coding System, Machine Learning, Convolutional Neural Networks, Artificial Intelligence, Deep Learning, Extended Cohn-Kanade (CK+).

Introduction:

Since the invention of advanced computers, it has become the desire of scientists to develop an intelligent system that can meet human capabilities both physically and mentally. In recent decades, computational power has continuously increased and is facilitating the development of fast-learning machines by providing a huge amount of data for training. A lot of machine learning techniques have been developed so far but they couldn't provide the promising results. Among them, deep learning techniques are considered powerful for developing AI systems/machines that approach human-level intelligence [1]. To accomplish this task, machines must be designed in such a way that they must be equipped with the capacity to understand, perceive, and respond accordingly to their environments. It also includes advanced interaction capacities due to their design involving some degree of autonomy [2]. Human emotion recognition is one of the critical tasks for machines. Expressions and emotions can serve as indicators of someone's emotional state. More so than words, a person's feelings may be expressed via body language and facial expressions. Academics have recently become interested in the application of facial expression detection in industries including virtual reality, healthcare, and intelligent tutoring systems. The main focus of the Facial Expression Recognition system is to find out emotional states like anger, surprise, fear, sadness, disgust, happiness, etc. based on given facial images [3]. Eight basic facial expressions are shown in Figure 1.

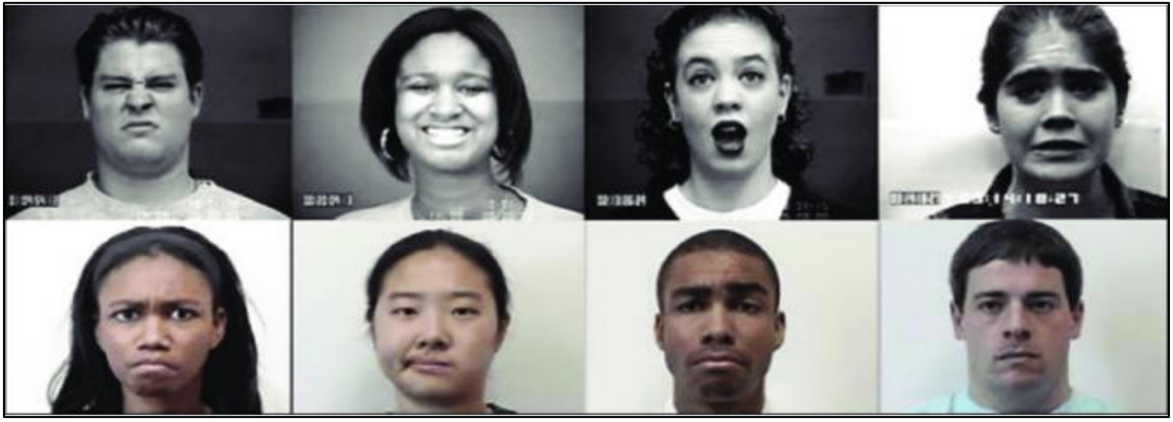


Figure 1. Expressions images from CK+ Dataset [4]

Various machine learning as well as deep learning techniques have been applied to recognize emotions. Among them, Convolutional Neural Networks (CNN) have been made an improvement to accomplish this task. A few challenges are still present including a low recognition rate and a longer time for training. CNN extracts the features automatically and classifies them based on given classes. Most of the models based on CNN developed in literature achieve reasonable accuracy but still face challenges due to variability in pose, lighting condition, occlusion, and/or dataset limitations [5] [6]. To address these issues, we have collected our dataset, in which images are extracted from video sequences available on social media.

Deep Learning models have shown significant success in this area in recent years. As a result, combining this data with appropriate algorithms allows for the generation of machine vision [7]. According to affective computing, emotion detection is required for machines/robots to effectively fulfill their purpose [8]. Firstly, determine what feelings the human being experiences or genuinely feels, what technology is most suited to capturing them, and which theories and techniques are most successful. For example, the employment of robots in aged people care or as hospital, attendants necessitates a thorough awareness of the surroundings [9]. This problem has not yet been adequately solved because of all of these factors. Facial expressions give information about the inner condition of a person [6]. If a system can get a sequence of face images, deep learning models can assist machines in understanding their interlocutor's emotions. Deep learning models play a vital role in human-computer interaction tasks as they provide self-awareness to machines and help to enhance communication in an intelligent way [10], [11], [12].

Using various datasets, few researchers evaluate some cutting-edge deep learning and machine learning models for expression identification. The fundamentals of affect expression through body language have been discussed previously and different accuracies are achieved on different standard datasets. In this study, we proposed CNN based facial expression recognition system. The contribution of this work is as follows;

- A modified CNN architecture is produced by introducing additional layers.
- Different pose images from video sequences for seven different expressions are included.
- Data balancing and augmentation are performed to improve accuracy.

Literature Review:

Artificial Intelligence (AI) and deep face recognition is the emerging topic which has a much bigger impact on the enhancement of user experience and recognition of human emotions. The same emotions can be expressed by different human beings in different ways. Due to this, accurate detection of emotions is difficult. In light of this, for accurate detection of emotions, AI technology is used. Understanding of human emotions is a very difficult task. Human-machine technology is used to interpret human facial expressions and gestures. Voice, text, facial expressions, and language can be used to show emotions. The best way to detect the emotions of a person is their facial expressions.

The paper reports a study utilizing Deep Face and artificial intelligence for human emotional recognition[13] with a particular emphasis on detecting emotions such as happiness, sadness, anger, surprise, and neutrality in near real-time. The system makes use of deep learning and convolutional neural networks to evaluate facial characteristics and the model has an overall accuracy of 94%. The researchers trained and validated their model using the FER 2016 database since it gives a wide range of facial motions concerning emotions. Their system is quite effective even with irregular illumination and head tilts of up to 25 degrees.

In [14] researchers focus on user-generated collections of videos which have expanded in recent years and have led to a high demand for analyzing these expressions automatically. On the other hand, semantic analysis such as "skiing," and "birthday party" are used for translating emotions like "sadness" and "joy". In this paper author collected a dataset from different well-known video-sharing platforms and this dataset is used for benchmarking. Low-level semantics to high-level semantics features are extracted and expressions are recognized. In [15] a Google Net-based CNN model using CWT has been proposed for the detection of human emotions. The proposed technique indulges EEG images limiting real-time measurement and calculation of human emotions in the fastest possible time-frame besides incorporating only limited dataset techniques namely the GMEEMO dataset has been used with k-NN and SVM classifiers attaining 98% efficiency.

In [16], the author reported a review of the textual emotions detection (TED) in the area of medical sciences focusing on the COVID-19 pandemic, its impact and subsequently detecting the health issues based on people’s emotions. In the author's views, deep learning models using NLP techniques were mostly preferred over others having better results. In [17], VADER’s emotion semantic recognition library for audio and Kaggle datasets for speech recognition have been experimented with, achieving 70-72% accuracy. The proposed technique accurately predicts the wavelength of the target sample’s speech. However, speech emotion identification, audio, and facial expression recognition are scored and calculated separately, resulting in a lack of correlation for net output weights predicting total results, thus, lowering model accuracy.

A multi-modal fusion and feature analysis model [18] has been developed for processing multimodal signals. Authors designed a new method of processing multimodal signals taking into account the hysteresis and delays characteristic of said data with the ultimate objective of representing the emotions using a multimodal feature fusion approach. The dataset was evaluated against benchmark CMU Multimode Opinion Sentiment and emotions Intensity Corpus with significantly improved results. An effective method for fine-tuning multi-modal along with the fusion technique was utilized, formulating a new sentiment detection model.

In daily life, Facial expressions are very vital for understanding emotions. According to a survey in this paper, verbal part that is linguistic language contributed 7% of communication between humans, the vocal part paralanguage is 38% and facial expressions are 55%. In face-to-face communication, facial expressions play an important role. AI plays a vital role in various aspects of machine and human interactions. It also reduces the gap between technology and human beings [19]. A big part of algorithms of AI techniques are used to recognize human emotions. In this paper, author focuses on numerous studies of learning techniques to learn the computer to distinguish particular emotions.

In [20] author uses the concept of action units for analysing facial expressions. The action unit consists of the emotions that are produced by a person by the movement of muscles. The Facial Action Coding System (FACS) is a superset of AUs described by Ekman and Friesen in 1971 [12] to track the motions of a single muscle or a group of muscles during facial emotion activation. Another way to analyse facial expressions is to employ action units. An Action Unit (AU) describes the activity of muscles when a person produces facial expressions. FACS is a superset of AUs described by Ekman and Friesen in 1971 to track the motions of a single muscle or a group of muscles during facial emotion activation. It is observed that the upper portion of facial activities has minor interactions with the facial emotion of the lower part.

Facial activities are divided into two groups upper face and lower face. Eyes, brows, head, and cheeks are considered as the upper portion. The lower part consists of lips, nose, root areas, and chin. More than 7000 action unit combinations have been reported in this paper which contains a description of the specifics of facial expression [20], [21]. Facial expression recognition tasks in computer vision are not just a technological pursuit but also a framework for human emotion dynamics and cognition [22],[23]. Deep learning models produced promising results on FER system by offering robust schemes for the extraction of features and classification [24][25]. Models like CNNs and RNNs have a high capability to handle complex facial data. More advanced models like spatiotemporal, inception models, and attention mechanisms produce enhanced accuracy. Training and testing have been performed on different datasets having different challenges like illumination change, facial pose, and image size [26].

Lately, researchers have focused on facial expression recognition in the wild. The images included in the wild dataset have many factors like face blurring and occlusion due to which many challenges occurred. To address these challenges, [27] proposed a dual-attention network. It included three parts, 1st is cross fusion dual attention scheme to obtain the local features and global information, 2nd is the activation function and 3rd is a module to introduce generalization ability in the model. FER Plus, RAF-DB, and Affect Net datasets were used and obtained 92.02%, 92.78%, and 63.58%, respectively. Experimental results show that this model provides a better solution to FER problem. The researcher [28] proposed a hierarchical network with an attention mechanism along the feature fusion technique. Local and global context features are extracted from diverse feature extraction modules. The gradient features are also extracted because they are robust to illumination changes. These features are further fused using the hierarchical module to skip the irrelevant features. The model achieves better accuracy among the existing FER methods.

The researcher [29] proposed a lightweight fully convolutional network including a multi-scale fusion mechanism. It also includes the two blocks called Mass attention that generate the spatial attention map and point-wise feature selection selects the important and suppresses the irrelevant features. They achieved an accuracy of 90.77% on FDEF, 70.4% on FER, and 86% on FER+ datasets. Table 1 shows the comparison of a few deep learning methods developed for facial expression recognition.

Table 1. Deep learning-based facial expression recognition systems

Author	Method used	Dataset	Accuracy
O. Khajuria et al [30]	CNN and VGG-16 with transfer learning	FER dataset	91%
Erlangga et al [31]	CNN with transfer learning (fine-tuned) with pre-trained models of Inception-V3 and MobileNet-V2	Video dataset	96%
Fatima et al [32]	CNN with hyperparameter tuning	FER13 dataset	89%
Burhan et al [33]	MobileNet-V1	Real time dataset	97.9%
Hadhami et al [34]	Hybrid CNN-SVM	Ryerson Multimedia Laboratory (RML) dataset	97.6%
Brijesh et al [35]	Convolution Neural Network (CNN) based deep learning approach	OAHEGA and FER-2013	73%
Mehrotra et al [36]	CNN	FER13 dataset	71%
Chowdary et al [37]	pre-trained networks of Resnet50, vgg19, Inception V3, and Mobile Net	CK+	96%
Deepak et al [26]	Retina Net model, Neural Architectural Search (NAS Net) Large feature extractor with a gated recurrent unit (GRU) model	KDEF and KMU-FED datasets	99.6%
Pan et al [38]	Improved GhosNet	CK+	98.2%
Rajesh et al [39]	3D-CNN and Conv LSTM	SAVEE, CK +, and AFEW	95.1%
Hangaragi et al [40]	Deep learning model	Labeled wild face (LWF) dataset	94.23%

Objective of the Research:

The objective of this study is to produce a robust algorithm for facial emotion classification that can accurately recognize facial expressions in images and videos. It also enhances human emotional understanding through an automated system that significantly contributes to the field of psychology, autism patients, enhancing classroom teaching environment, AI and robotics systems, etc. In this study two novelties are introduced;

CNN is modified by incorporating the additional layers in it to improve the performance of the network as compared to standard CNN. The modification involves the addition of a dropout layer and max pooling layer after every convolutional layer. Max pooling reduces the feature map size which helps to decrease the computational cost and no of parameters. The dropout layer prevents the model from being over-fit. The novel dataset is created by using videos on social media I.e. FaceBook, Instagram, YouTube, etc. The dataset includes images of seven different expressions with poses.

Methodology:

CNN is known as one of the popular methods to analyze images. It includes the hidden layers called the convolutional layer which is why different from Multi-Layer Perceptron (MLP). We modified conventional CNN to use it for our purpose. The proposed system (as shown in Figure 2) in this research comprises three main steps; (i) Pre-processing (ii) Emotion recognition using modified CNN (iii) Post-processing and emotion labeling.

In the first step, we acquired the image from video frames. Video holds a lot of information. Our main focus is to detect and extract the human face from the video frame to avoid unnecessary information processing. The acquired face in this step is pre-processed and passed to the proposed CNN model. The network is trained to classify the seven different expressions. Finally, the recognized emotions are labeled.

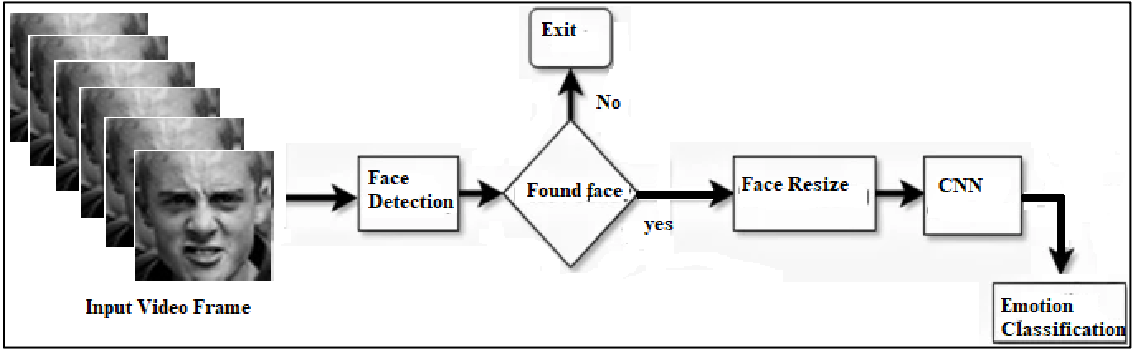


Figure 2. Proposed AI-based FER System dataset

Datasets:

Generally, the neural network's performance depends upon the large enough training data to get the better performance of networks. By keeping this concept in mind, we used three datasets Extended Cohn-Kanade (CK+) [8], Facial Expression dataset (FER-2013) [41], and our own images dataset prepared from videos.

Extended Cohn-Kanade:

The Extended Cohn-Kanade (CK+) database is an extended version of the Cohn-Kanade (CK) database which was released in the early years of 2000. In this database, 27% no. Of persons, the sequence numbers are increased by 22%. The CK+ database is made up of 593 sequences from 123 subjects. Frontal view images are captured using AG-7500 cameras and these are also transformed into 640x480 pixels with 8 or 16-bit color values. The sequence contains the neutral expression that leads to the peak expression of a particular form. Figure 3 shows a particular image sequence from the CK and CK+ database. The database comprises seven different expressions i.e. Happiness, Sadness, Surprise, Contempt, Disgust, Anger, and Fear.

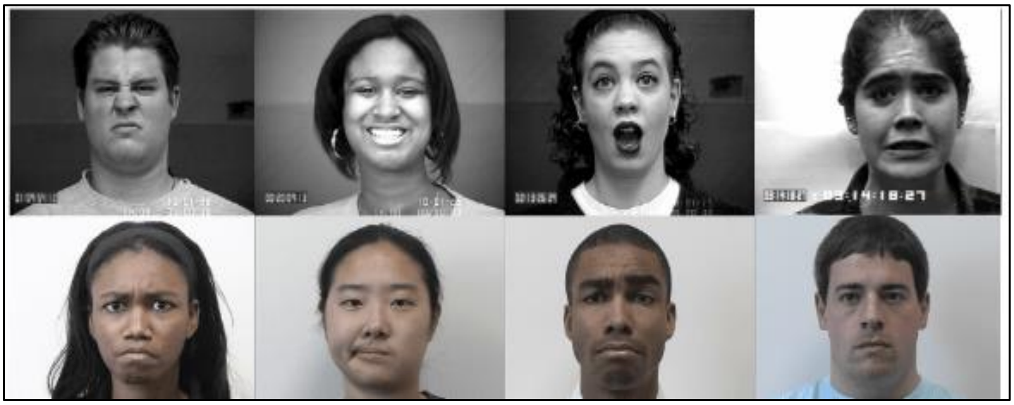


Figure 3. Different emotions – Row1 represents the emotions from CK while Row2 represents the emotions from CK+ dataset [4]

FER Dataset:

In ICML 2013, the FER-2013 database was first introduced. It consists of a total of 35,887 images of seven different expressions (e.g. happy, neutral, disgust, surprise, sad, and fear). Each image is taken from the front view and is of grayscale. The highest number of images are included from happy expressions while disgust images are very low in count as compared to other expressions. Figure 4 shows the sample images of the FER+ database.

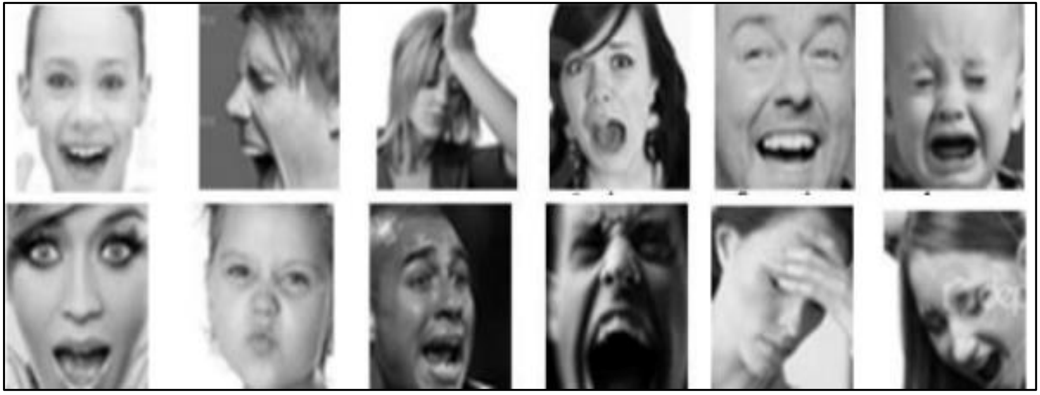


Figure 4. Images of FER+ Dataset [41]

Own Dataset:

Apart from the publicly available dataset mentioned above, we have collected our images from different videos available on social media platforms like YouTube, Facebook, Instagram, Twitter, etc. Most of the images that we collected are from Bollywood, Hollywood actors of different ages. These actors are from drama and movie series. 10000 images of seven different expressions are collected and labeled according to the category of expression. Figure 5 shows the sample images of different expressions extracted from the video sequence;

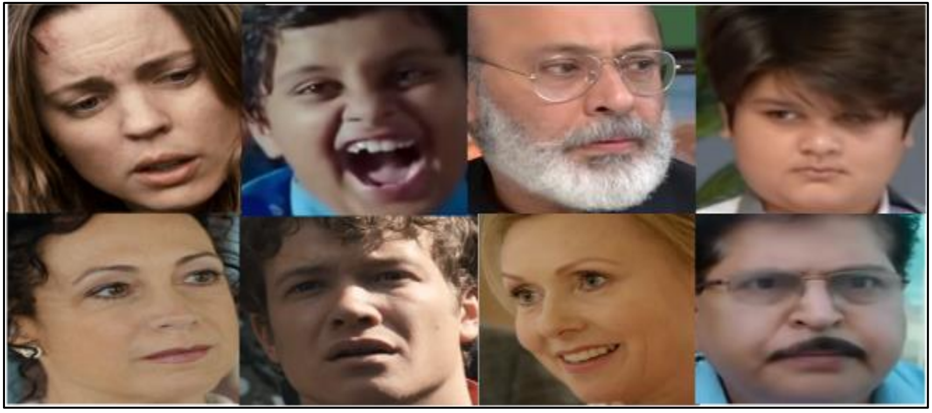


Figure 5. Collection of datasets having different expressions

Input Dataset Preparation:

Due to several difficulties with this dataset, we have to perform a few steps to prepare the dataset for input to CNN.

Frame Extraction:

To collect the images from videos and form a dataset, our first step is to extract the frame from videos. After starting the video, when the image frame that contains the human face is obtained or within a sec a face arrives, it is first detected and then cropped.

Data Imbalance:

When one class contains significantly more images than another class, there is an imbalance. As a result, the model favors one class over another. For instance, if there are 2000 images of happy faces and 500 images of fearful faces, the model will be biased in favor of the happy faces. To get around this issue, we used horizontal/vertical flipping and image rotation with a random angle from -45 to 45 (in degrees) to augment the data. Image re-scaling is also done because the standard input image size for all three datasets is 48 x 48.

Brightness Difference:

The image collection comprises both low-contrast and high-contrast types. The images contain visual information, and consequently, those with brighter images possess more details than those with lower contrast. The performance of Convolutional Neural Networks (CNNs) can be significantly impacted by fluctuations in image contrast. To address this issue, histogram equalization is employed to enhance the contrast of an image.

Face Detection:

Face detection has been a subject of extensive research, with the ultimate goal of designing a system that can effectively detect faces across various illuminations, skin tones, and backdrop sizes. This challenging task is crucial for our Facial Emotion Recognition (FER) method, which relies on the Haar Cascade Classifier for face detection. This well-established classifier is renowned for its computational efficiency and robust performance. The Haar Cascade Classifier is trained using a dataset comprising both positive and negative image samples. Feature extraction is then performed on these images, where feature values are calculated by subtracting the total number of pixels in the white rectangle from those in the black rectangle. This enables the classifier to recognize multiple faces in diverse settings. In constant time, integral images enable the calculation of the Haar-like feature of any size [2]. Figure 6 shows the resultant image after the detection of the face.

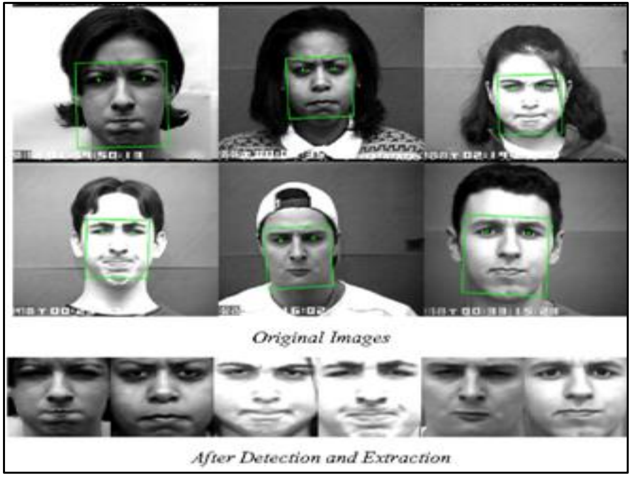


Figure 6. Face Detection and Extraction from Different Expressions [4]

Proposed Network Overview - CNN Architecture: The proposed Convolutional Neural Network (CNN) model, illustrated in Figure 7, consists of two fully connected layers and four convolutional layers. Unlike traditional CNNs, our model incorporates max pooling and dropout layers after each convolutional filter to reduce network complexity and enhance performance. The pooling layers utilize a constant function to transform activations, while the convolutional layers employ learnable weights. The entire network is introduced to non-linearity through the Rectified Linear Unit (ReLU), which does not affect the receptive fields of the convolutional layers. Looping mechanisms are used to optimize the convolutional layers' efficiency. Furthermore, the learning parameters, including weights and kernels, are updated during training, and the model's output is determined using a training dataset. To generate a significant output, it is beneficial to incorporate or remove specific layers, such as convolutional or pooling layers, during training cycles. The pooling layer contributes to reducing computations and parameters, thereby controlling overfitting. The features obtained after the pooling layer are remapped from a 2D structure to a 1D feature vector using two fully connected layers. Consequently, a flat pooling feature map is obtained. This feature vector is then forwarded to the Fully Connected (FC) layer for classification.

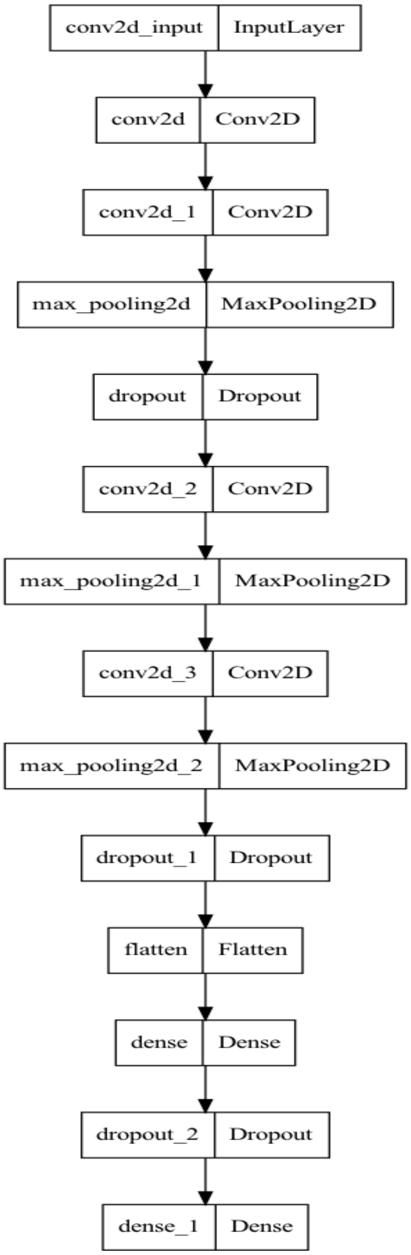


Figure 7. Proposed CNN Architecture Overview

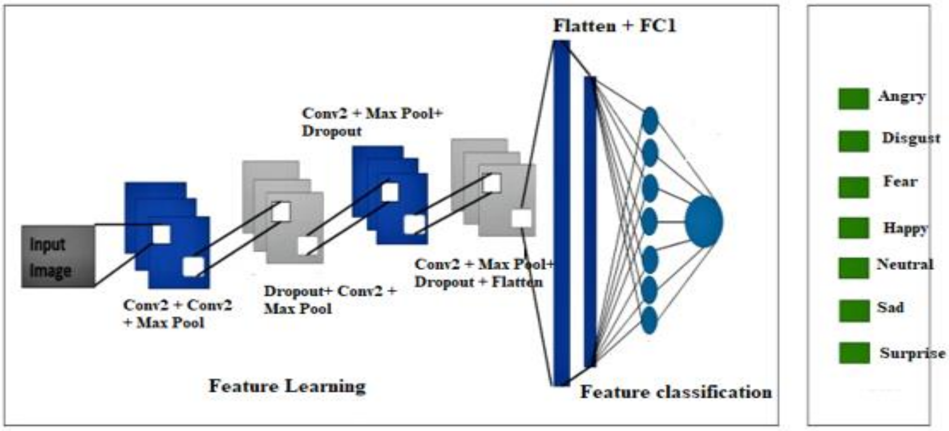


Figure 8. Facial Expression Classification System based on proposed CNN

Convolutional Neural Network (CNN) Tuning:

To optimize the CNN model, hyperparameter tuning is essential. This involves adjusting parameters such as padding, filtering, stride, batch size, and learning rate. These parameters regulate the output convolutional layer's size. Padding adds zeros to the input's border, while stride determines the distribution of width and height parameters. Smaller stride sizes result in larger output with extensively overlapping receptive fields, whereas larger strides reduce the overlapping ratio, yielding smaller output dimensions.

In our proposed model, comprising two fully connected layers and four convolutional layers, only the fully connected layers acting as a classifier and the high-level detailed feature block require modifications. Additionally, we reset the Softmax ranking to 7 grades from 1000 rankings, as our dataset consists of only seven emotions.

Data Splitting:

We employed three different datasets with varying splitting schemes. For the FER dataset, we split the data into training and testing sets with a 75:25 ratio. In contrast, the CK+ and our dataset were split into 80% training and 20% testing sets. We also evaluated our model's performance on videos obtained from YouTube. Our collected dataset was split into a 90:10 ratio.

Model Training:

The model was trained on all three datasets. The calculation of the model's weights depends on the learning parameter selection. A low learning rate yields more accurate results but requires longer computation time, whereas a larger learning rate leads to faster convergence. The dataset is run through the neural network both forward and backward in multiple epochs. To reduce processing time, the dataset is divided into batches, and the batch size refers to the number of training images in each batch.

Model Evaluation:

The model is evaluated on the test datasets. The corresponding weight values and parameters learned during the training phase enable the recognition of novel facial expressions. Given that the developed model already possesses weights, Facial Emotion Recognition (FER) can be performed quickly for real-time images.

Experimental Results:

Table 2 presents the classification results obtained by the proposed model, while Tables 3-5 show the confusion matrix of training results on all three datasets. The results demonstrate that most expressions are well-classified and recognized accurately. As the epoch increases, the loss decreases, and the accuracy rises, as shown in Figure 10. The training and validation precision is determined using the training and validation accuracy. A slight difference between both training and validation accuracy is observed as the epochs increase. The validation loss may still be lower, indicating that the model fits the training results accurately.

Table 2. Classification Results on FER, CK+, and our dataset

Datasets	Parameters/Expressions	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
FER	Precision	0.8	0.87	0.86	0.91	0.85	0.76	0.9
	Recall	0.81	0.84	0.79	0.96	0.86	0.82	0.9
	F1-Score	0.82	0.85	0.8	0.93	0.85	0.81	0.9

CK+	Precision	0.85	0.89	0.9	0.96	0.88	0.87	0.95
	Recall	0.84	0.9	0.9	0.97	0.87	0.87	0.92
	F1-Score	0.85	0.89	0.91	0.95	0.86	0.87	0.92
Our Dataset	Precision	0.96	0.96	0.96	0.99	0.97	0.96	0.97
	Recall	0.98	0.97	0.97	0.99	0.97	0.97	0.97
	F1-Score	0.97	0.97	0.96	0.98	0.98	0.96	0.97

Table 3. Confusion Matrix of Training results on FER dataset [41]

Expression	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.81	0.0	0.02	0.03	0.05	0.08	0.01
Disgust	0.12	0.89	0.01	0.01	0.01	0.07	0.0
Fear	0.07	0.0	0.86	0.07	0.0	0.0	0.0
Happy	0.02	0.0	0.01	0.96	0.0	0.00	0.01
Neutral	0.03	0.0	0.05	0.02	0.83	0.06	0.01
Sad	0.16	0.0	0.02	0.0	0.0	0.82	0.0
Surprise	0.0	0.0	0.02	0.08	0.0	0.0	0.90

The accuracy of face emotion recognition systems can be compromised by several factors, including incorrectly and poorly cropped images. Incorrectly cropped images can significantly decrease the efficiency of face emotion recognition, emphasizing the need for an improved face detector to minimize or avoid this impact. However, a more advanced face detector may involve more complex processes, potentially increasing the running time. Poor-quality cropped images can result from various factors, including user movement and location. When the user is far from the camera, the cropped image may be smaller than required, leading to a blurred image after resizing. These factors highlight the importance of optimizing image cropping and face detection processes to ensure accurate face emotion recognition.

Table 4. Confusion Matrix of Training results on CK+ dataset [4]

Expression	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.85	0.05	0.02	0.03	0.0	0.0	0.05
Disgust	0.0	0.90	0.0	0.0	0.0	0.10	0.0
Fear	0.03	0.0	0.91	0.0	0.0	0.0	0.06
Happy	0.02	0.0	0.01	0.95	0.0	0.00	0.02
Neutral	0.0	0.06	0.0	0.0	0.88	0.06	0.0
Sad	0.10	0.03	0.0	0.0	0.0	0.87	0.0
Surprise	0.0	0.0	0.0	0.06	0.0	0.0	0.94

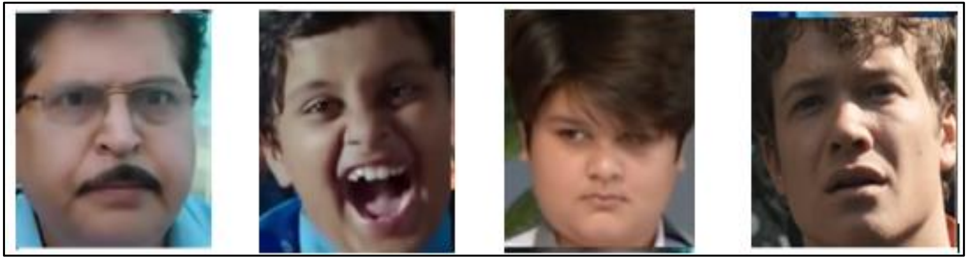


Figure 9. Miss-classified Expressions

Figure 9 illustrates examples of misclassified images by the facial expression recognition system. Specifically, an angry expression is incorrectly identified as fear, while a surprised expression is misclassified as happy. Additionally, a neutral expression is mistakenly labeled as depressed, and a sad expression is incorrectly classified as surprised. These examples highlight the challenges faced by the system in accurately recognizing facial expressions with nuanced or subtle differences.

Table 5. Confusion Matrix of own collected dataset (Test)

Expression	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.98	0.10	0	0	0	0.02	0
Disgust	0.30	0.97	0	0	0	0	0
Fear	0	0	0.97	0	0.10	0	0.15
Happy	0	0	0	0.99	0	0	0.01
Neutral	0	0	0	0	0.98	0.02	0
Sad	0	0	0	0	0.03	0.97	0
Surprise	0	0	0	0.03	0	0	0.97

Discussion:

The proposed model was evaluated using accuracy and confusion metrics on three distinct datasets: CK+, FER+, and our own dataset. The CK+ database comprises 593 sequences from 123 subjects, while FER+ consists of 35,887 images of seven different expressions. Both datasets primarily feature frontal-view images with minimal pose variations. In contrast, our dataset focuses on capturing pose variations, consisting of 250 expression images of individuals with diverse ages and poses.

For evaluation purposes, we selected 350 sequences from the CK+ dataset, 200 from our own dataset, and 25,000 images from the FER dataset. We also utilized images from these datasets for testing. All images were resized and fed into the Convolutional Neural Network (CNN) for recognition purposes.

The Adam optimizer was employed to train the model, with a learning rate (LR) of 0.01, a training ratio of 0.8, a batch size of 16, and 100 epochs. After training the model for 100 epochs, we observed an accuracy range of 68% to 72% across two attempts. Notably, increasing the number of epochs led to higher accuracy, eventually stabilizing at 86% after five attempts for the FER 2013 dataset.

Our results, summarized in Table 1, demonstrate that the modified CNN outperforms existing models in classifying expressions such as sad, surprise, and happy, which involve intense changes in facial expressions. However, the model struggles to distinguish between disgust and angry expressions, likely due to their similar characteristics.

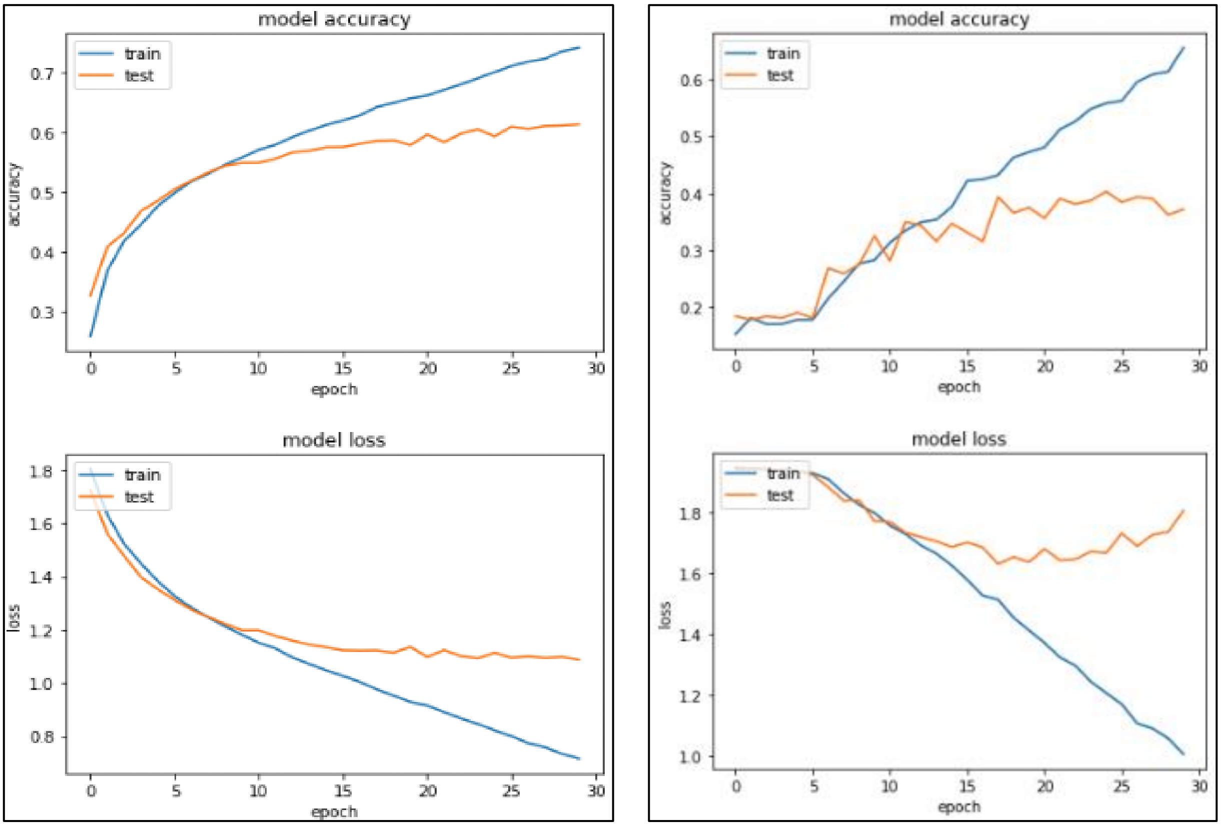
In the literature, video sequences rather than still images were taken into account. Action units were utilized in place of face landmarks in [22], [23]. The two sides of the face (top and bottom) and upper and lower lip parts are represented by using these action units. Seven expressions happy, fear, sad, angry, surprise, disgust, and neutral were taken into account for categorization in [5]. The anger expression accuracy was 53%, disgust was 70.0, fear was 46.0%, happiness was 80.5%, sadness was 63.0, surprise was 62.5, and neutral was 51.5. Comparison of these findings with our system shows that proposed CNN models outperform previous techniques in the literature in terms of accuracy and computing time.

Table 6: Illustrates the comparative analysis of the proposed method with previous methods on the CK+ dataset, FER, and our dataset.

Table 6. Performance comparison of the proposed Emotion Recognition System

Description of Model	Database	Accuracy (%)
Facial Expression Recognition Using Convolutional Neural Networks and SVM [42]	CK+ BU4D	83.4% 94.69
Facial Expression Recognition with Convolutional Neural Networks [43]	FER 2013	75.2%
Deep learning-based facial emotion recognition for human–computer interaction applications [37]	CK+	97.7%

Modified Convolutional Neural Networks for Facial Emotion classification **(Proposed)**	FER 2013 CK+ our dataset	85.4% 89.4% **97.4%**



Training data Model Accuracy vs Model Loss Test data model accuracy vs loss

Figure 10. Model Accuracy vs Model Loss

Referring to Table 6, the proposed model showed better performance on all three datasets as compared to the conventional CNN [42][37] applied to the same dataset. The proposed method achieved an average recognition accuracy rate of 97.4% on our self-collected dataset, outperforming existing studies, which reported average accuracy rates ranging from 85% to 90%. These results demonstrate that our model is particularly effective for facial emotion recognition (FER) in challenging expression recognition scenarios and exhibits excellent performance on our own dataset, where most expressions are accurately classified.

Figure 10 illustrates the model's accuracy on training and test data. As the training period progresses, the validation loss increases, while the training loss decreases. This trend suggests that adjusting the weights leads to improved performance on the validation data. Furthermore, as the epoch number increases, the validation loss rate is expected to be lower than the training loss rate, which is consistent with our observations in the later stages of training.

Conclusion:

This study aimed to develop a system for categorizing facial emotions from static facial images extracted from videos using deep learning techniques. Although several alternative approaches have been employed to address this challenging problem, this research introduces a modified Convolutional Neural Network (CNN) model. While the outcomes were not groundbreaking, they demonstrated a slight improvement over previous models. This suggests that, given sufficient labeled samples, deep learning approaches can ultimately resolve this issue. Larger datasets may enable the use of networks with greater feature learning capacity.

Our proposed model performed well on all three datasets. Future research will focus on improving the consistency of the CNN model and exploring potential feature enhancements or fusion methods. Additionally, we will investigate various human factors, including personality traits, gender, and age, that influence emotional recognition abilities.

Acknowledgment:

I would like to express my sincere gratitude to the Department of Software Engineering and Computer Science, NUML Rawalpindi Campus, for providing access to their high-end computing labs. These state-of-the-art facilities significantly enhanced the quality and efficiency of my research.

Author's Contribution: This research is a teamwork and all authors contributed equally.

Conflict of Interest: The authors declare no conflict of interest regarding the publication of the paper.

Project Details: NIL

References:

[1] Y. Wang, "Research on the Construction of Human-Computer Interaction System Based on a Machine Learning Algorithm," *J. Sensors*, 2022, doi: https://doi.org/10.1155/2022/3817226.

[2] Y. G. and Y. S. Y. -J. Liu, M. Yu, G. Zhao, J. Song, "Real-Time Movie-Induced Discrete Emotion Recognition from EEG Signals," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 550–562, 2018, doi: 10.1109/TAFFC.2017.2660485.

[3] S. L. Dingus, Thomas A, Feng Guo, "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proc. Natl. Acade*, vol. 13, no. 10, pp. 2636–2641, 2016, doi: 10.1073/pnas.1513271113.

[4] S. M. H. Garcia, "CK+," *DataCite Commons*, 2024, [Online]. Available: https://commons.datacite.org/doi.org/10.5281/zenodo.11221350

[5] G. H. et Al, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012, doi: 10.1109/MSP.2012.2205597.

[6] M. J. M. P S Bellet, "The Importance of Empathy as an Interviewing Skill in Medicine," *JAMA*, vol. 266, no. 13, pp. 1831–2, 1991, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/1909761/

[7] and I. T. K. Han, D. Yu, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," *Interspeech*, 2014, doi: 0.21437/Interspeech.2014-57.

[8] H. L. and E. M. P. Y. Kim, ""Deep learning for robust feature generation in audiovisual emotion recognition," *IEEE Int. Conf. Acoust. Speech Signal Process. Vancouver, BC, Canada*, pp. 3687–3691, 2013, doi: 10.1109/ICASSP.2013.6638346.

[9] B. C. Ko, "A Brief Review of Facial Emotion Recognition Based on Visual Information," *Sensors (Basel)*, vol. 18, no. 2, p. 401, 2018, doi: 10.3390/s18020401.

[10] and L. C. H. Li, J. Sun, Z. Xu, "Multimodal 2D+3D Facial Expression Recognition With Deep Fusion Convolutional Neural Network," *IEEE Trans. Multimed.*, vol. 19, no. 12, pp. 2816–2831, 2017, doi: 10.1109/TMM.2017.2713408.

[11] and W. B. H. Bejaoui, H. Ghazouani, "Fully Automated Facial Expression Recognition Using 3D Morphable Model and Mesh-Local Binary Pattern," *Adv. Concepts Intell. Vis. Syst.*, pp. 39–50, 2017, doi: 10.1007/978-3-319-70353-4_4.

[12] P. E. and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personal. Soc. Ps*, vol. 17, no. 2, pp. 124–9, 1971, doi: 10.1037/h0030377.

[13] and T. J. J. R. Venkatesan, S. Shirly, M. Selvarathi, "Human Emotion Detection Using DeepFace and Artificial Intelligence," *Eng. Proc.*, 2023, doi: 10.3390/engproc2023059037.

[14] X. X. Yu-Gang Jiang, Baohan Xu, "Predicting emotions in user-generated videos", [Online]. Available: https://cdn.aaai.org/ojs/8724/8724-13-12252-1-2-20201228.pdf

[15] M. Aslan, "CNN based efficient approach for emotion recognition," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7335–7346, 2022, doi: https://doi.org/10.1016/j.jksuci.2021.08.021.

[16] B. O. Alieh Hajizadeh Saffar , Tiffany Katharine Mann, "Textual emotion detection in health: Advances and applications," *J. Biomed. Inform.*, 2023, doi: 10.1016/j.jbi.2022.104258.

[17] J. H. Yuwei Chen, "Deep Learning-Based Emotion Detection," *J. Comput. Commun.*, vol. 10, no. 2, 2022, doi: 10.4236/jcc.2022.102005.

[18] and R. Z. Q. Qi, L. Lin, "Feature Extraction Network with Attention Mechanism for Data Enhancement and Recombination Fusion for Multimodal Sentiment Analysis," *Information*, vol. 12, no. 9, p. 342, 2021, doi: https://doi.org/10.3390/info12090342.

[19] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 15, pp. E1454–E1462, Apr. 2014, doi: 10.1073/PNAS.1322355111/ASSET/C2BBF816-8771-46B5-9871-0AA84A513697/ASSETS/GRAPHIC/PNAS.1322355111I91.GIF.

[20] T. K. and J. F. C. Y. . -I. Tian, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, 2001, doi: 10.1109/34.908962.

[21] W. V Ekman, P., & Friesen, "Facial Action Coding System," *APA PsycNet Direct*, 1978, doi: https://doi.org/10.1037/t27734-000.

[22] P. P. R. and D.-M. J. J. -H. Kim, B. -G. Kim, "Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure," *IEEE Access*, vol. 7, pp. 41273–41285, 2019, doi: 10.1109/ACCESS.2019.2907327.

[23] E. C. D. and R. I. B. Ramdhani, "Convolutional Neural Networks Models for Facial Expression Recognition," *Int. Symp. Adv. Intell. Informatics (SAIN), Yogyakarta, Indones.*, pp. 96–101, 2018, doi: 10.1109/SAIN.2018.8673352.

[24] and Z. T. Zhang, Tao, "Survey of deep emotion recognition in dynamic data using facial, speech and textual cues," *Multimed. Tools Appl.*, vol. 83, pp. 66223–66262, 2024, [Online]. Available: https://link.springer.com/article/10.1007/s11042-023-17944-9

[25] G. Muhammad Sajjad, Fath U Min Ullah, Christodoulou and J. J. P. C. R. Ullah, Mohib, Faouzi Alaya Cheikh, Mohammad Hijji, Khan Muhammad, "A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines," *Alexandria Eng. J.*, vol. 68, pp. 817–840, 2023, doi: https://doi.org/10.1016/j.aej.2023.01.017.

[26] S. A. Jain, Deepak Kumar, Ashit Kumar Dutta, Elena Verdú and A. R. W. Sait, "An automated hyperparameter tuned deep learning model enabled facial emotion recognition for autonomous vehicle drivers," *Image Vis. Comput.*, vol. 133, p. 104659, 2023, doi: https://doi.org/10.1016/j.imavis.2023.104659.

[27] H. W. & C. Z. Fan Zhang, Gongguan Chen, "CF-DAN: Facial-expression recognition based on cross-fusion dual-attention network," *Comput. Vis. Media*, vol. 10, pp. 593–608, 2024, doi: https://doi.org/10.1007/s41095-023-0369-x.

[28] and Q. D. Tao, Huanjie, "Hierarchical attention network with progressive feature fusion for facial expression recognition," *Neural Networks*, vol. 170, pp. 337–348, 2024, doi: https://doi.org/10.1016/j.neunet.2023.11.033.

[29] A. A. Ali Ezati, Mohammadreza Dezyani, Rajib Rana, Roozbeh Rajabi, "A Lightweight Attention-based Deep Network via Multi-Scale Feature Fusion for Multi-View Facial Expression Recognition," *Comput. Vis. Pattern Recognit.*, 2024, doi: https://doi.org/10.48550/arXiv.2403.14318.

[30] R. K. and M. G. O. Khajuria, "Facial Emotion Recognition using CNN and VGG-16," *Int. Conf. Inven. Comput. Technol. (ICICT), Lalitpur, Nepal*, pp. 472–477, 2023, doi:

10.1109/ICICT57646.2023.10133972.

[31] T. W. Erlangga Satrio Agung , Achmad Pratama Rifai, "Image-based facial emotion recognition using convolutional neural network on emognition dataset," *Sci. Rep.*, vol. 13, no. 1, p. 14429, 2024, doi: 10.1038/s41598-024-65276-x.

[32] S. S. Fatimatuzzahra, Lindawati, "Development of Convolutional Neural Network Models to Improve Facial Expression Recognition Accuracy," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 10, no. 2, pp. 279–289, 2024, doi: 10.26555/jiteki.v10i2.28863.

[33] amna kosar hafiz burhan ul haq, waseem akram, muhammad nauman irshad and M. Abid, "Enhanced real-time facial expression recognition using deep learning," *Acadlore Trans. AI Mach. Learn.*, vol. 3, no. 1, pp. 24–35, 2024, doi: https://doi.org/10.56578/ataiml030103.

[34] H. A. & Y. Ben Ayed, "Deep facial expression detection using Viola-Jones algorithm, CNN-MLP and CNN-SVM," *Soc. Netw. Anal. Min.*, vol. 14, p. 65, 2024, doi: https://doi.org/10.1007/s13278-024-01231-y.

[35] R. R. & K. K. M. Brijesh Bakariya, Arshdeep Singh, Harmanpreet Singh, Pankaj Raju, "Facial emotion recognition and music recommendation system using CNN-based deep learning techniques," *Evol. Syst.*, vol. 15, pp. 641–658, 2024, doi: https://doi.org/10.1007/s12530-023-09506-z.

[36] K. P. S. and Y. B. S. M. Mehrotra, "Facial Emotion Recognition and Detection Using Convolutional Neural Networks with Low Computation Cost," *2nd Int. Conf. Disruptive Technol. (ICDT), Gt. Noida, India*, pp. 1349–1354, 2024, doi: 10.1109/ICDT61202.2024.10489678.

[37] and D. J. H. M. K. Chowdary, T. N. Nguyen, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput. Appl.*, vol. 35, no. 4, pp. 1–18, 2021, doi: 10.1007/s00521-021-06012-8.

[38] Z. Z. and S. W. J. Pan, W. Fang, Z. Zhang, B. Chen, "Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG," *IEEE Open J. Eng. Med. Biol.*, vol. 5, pp. 396–403, 2024, doi: 10.1109/OJEMB.2023.3240280.

[39] A. V. & S. S. Rajesh Singh, Sumeet Saurav, Tarun Kumar, Ravi Saini, "Facial expression recognition in videos using hybrid CNN & ConvLSTM," *Int. J. Inf. Technol.*, vol. 15, pp. 1819–1830, 2023, doi: https://doi.org/10.1007/s41870-023-01183-0.

[40] N. Shivalila Hangaragi, Tripty Singh, N, "Face Detection and Recognition Using Face Mesh and Deep Neural Network," *Procedia Comput. Sci.*, vol. 218, pp. 741–749, 2023, doi: https://doi.org/10.1016/j.procs.2023.01.054.

[41] "Challenges in Representation Learning: Facial Expression Recognition Challenge," *Res. Predict. Compet.*, [Online]. Available: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data

[42] and C.-S. L. J.-C. Kim, M.-H. Kim, H.-E. Suh, M. T. Naseem, "Hybrid Approach for Facial Expression Recognition Using Convolutional Neural Networks and SVM," *Appl. Sci.*, vol. 12, no. 11, p. 5493, 2022, doi: 10.3390/app12115493.

[43] K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. Santosh Kumar, R. Manjunath, and S. Jaiswal, "Facial emotion recognition using convolutional neural networks," *Mater. Today Proc.*, vol. 80, pp. 3560–3564, Jan. 2023, doi: 10.1016/J.MATPR.2021.07.297.