



TO IDENTIFY FACIAL EMOTIONS USING AN ENHANCED CNN-TCN MODEL

B. Maidhili^{*1}, Dr.R. Thirumalai Selvi²

¹Research Scholar, PG & Research, Department of Computer Science, Government Arts College, Nandanam, Chennai. Affiliated to University of Madras, India.

²Associate Professor, PG & Research, Department of Computer Science, Government Arts College, Nandanam, Chennai. Affiliated to University of Madras, India.

¹<https://orcid.org/0009-0006-5369-5784>, ²<https://orcid.org/0009-0001-7864-1196>

Email: *maidhili.b@gmail.com, sarassselvi@gmail.com

ARTICLE INFO

Article History

Received: January 21, 2026

Reviewed: March 22, 2026

Accepted: April 22, 2026

Published: May 29, 2026

Keywords:

Facial Emotion Recognition (FER), Convolutional Neural Network (CNN), Temporal Convolutional Network, Temporal dimension, Spatial Dimension.

ABSTRACT

A fascinating area of research, facial expression-based automatic emotion recognition has been introduced and applied in a variety of fields, such as safety, health, and human-machine interactions. A computer technique for recognizing different emotions is called Facial Emotion Recognition (FER), and it is crucial for artificial intelligence (AI) research. For FER classification using video frames, this study employs convolutional neural networks (CNNs) and temporal convolutional networks (TCNs) using the Nadam optimizer in conjunction with preprocessing techniques. The preprocessing procedures include shuffling, one-hot encoding, and video pixel normalization. These are first used in the video frames to identify emotions. In order to manage the temporal and spatial dimensions of the video input, the classifier first incorporates 3D Convolution layers. The result should then be reformatted to be fed into the 1D Convolution layers of the TCN section. This work illustrates the architecture, the databases used, and the contributions addressed by comparing the results with the proposed methods. The performance of the proposed model is compared with other pre-trained deep learning models, including GoogleNet, ResNet, and Inception v3. The suggested enhanced CNN-TCN model with Nadam optimizer had 98% accuracy, precision, recall, and F1 score.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Facial expressions are essential for interpreting and identifying emotions. The importance of the face of communication is even indicated by the name "interface." between two things. Recognizing emotions is crucial for effective communication, as emotions account for up to 93% of everyday communication cues. Computer-human interfaces (HCI) would like technology to be able to understand human emotions [1]. Emotion detection has garnered a lot of attention lately because of its important uses in human-computer interaction. A variety of verbal and nonverbal languages, including text, audio, and visual, are used to convey human emotions [2]. Humans are able to identify the emotions of others based on the situation. For instance, depending on the situation, a grin might be seen as a message of greeting, irony, delight, or politeness. Humans use a variety of methods to communicate their feelings, including voice, gestures, facial expressions, and text emoticons. Emotions may be more clearly defined with the help of facial expression recognition. Many years ago, studies of facial expressions were initiated [3]. Enhancing facial classification is a complex and nuanced task, especially when it involves transferring features between different faces. This area presents numerous challenges, including variations in head orientation, lighting conditions, age, gender, and background. Occlusion caused by factors like sunglasses, scarves, or skin conditions further complicates the process. Conventional approaches for extracting geometric and textural features from faces utilize techniques like LBP, FAC, and LDA [4]. In contrast, recent advancements in deep learning, particularly CNNs and RNNs, have shown strong potential in automating feature extraction and classification processes. These methods have gained significant popularity in facial emotion recognition, encouraging researchers to experiment with various deep neural network models that have demonstrated promising outcomes. This work provides an insightful and useful learning paradigm. Pattern recognition is a well-established paradigm in deep learning. The Convolution Neural Network (CNN) method is used, along with a sophisticated optimizer technique. CNN is a unique type of artificial neural network that uses a machine learning tool. CNN has to do with things like object detection, facial identification, and image processing.

Layers of neural networks make up a deep neural network (DCNN) [5]. The paper is organized as follows: Section 2 reviews the work related to emotion recognition through facial expressions; Section 3 discusses the methodology and algorithms proposed for emotion recognition; Section 4 presents the experimental analysis, including results and discussion; and finally, Section 5 provides the conclusion of the proposed system.

II. LITERATURE SURVEY

The author of this study [6] developed a model utilizing Deep learning methods that offer advantages over traditional machine learning, particularly for analyzing unstructured data such as videos and multimedia. The model implemented in this work is capable of real-time face recognition, emotion detection, and even music recommendations. Experimental analysis was conducted using the FER-2013 and OAHEGA datasets. The cutting-edge FER, POSTER, effectively merges facial landmark and image feature data. Nonetheless, POSTER's construction is unquestionably intricate. This study presents POSTER++ to tackle the computational difficulties related to POSTER [7]. POSTER++ replaces POSTER's pyramidal architecture with a lightweight multi-scale interaction (MSI) network, simplifying the two-stream approach used in POSTER. Through attentional interaction, MSI directly incorporates face data at various scales into the token embeddings. This study [8] leverages deep learning alongside saliency maps for analyzing facial images. The reprocessing of data includes applying a bilateral filter, generating a saliency map, and utilizing both a generative adversarial network (GAN) and CLAHE. For effective facial emotion classification, a deep (CNN) is then trained using the Nadam optimizer. Study [9] introduces a new Graph Neural Network (GNN) model that considers the systematic process of FER in human visual perception. First, an area division approach is to divide the face region into six pieces in order to standardize the selection of significant facial features. This leads to the proposal of a human visual cognition method that considers the connotative link between different components of facial expressions. Together with the human regional cooperative recognition process, the GNN model is built by extracting the connotative link (such as relative position and comparable structure) between graph nodes. In this paper, the author tackles the problem of compound facial expression recognition in uncontrolled settings [10].

The author creates a brand-new loss function based on center loss (CL) that is specifically designed for compound facial emotion detection: loss bi-center, in contrast to CL, which utilizes core emotion centers to guide deep neural networks in capturing complex emotion characteristics. Furthermore, the author presents a basic-center regularization term to manage differences between these core centers, enhancing the discriminative strength of the learned features. The study's author [11] specifically designed to identify emotions in autistic children, with a particular focus on recognizing distress and anger. The system follows three key steps: detecting faces, extracting facial features, and classifying emotions. With high accuracy in emotion classification, a DCNN architecture was utilized. Due to the extensive dataset, an autoencoder was employed to enhance feature extraction and selection. Additionally, models such as ResNet, MobileNet, and Xception were incorporated to boost system performance. This study [12] explores enhancing video-based emotion recognition by capturing spatio-temporal characteristics through enriched temporal encoding. The study tackles the difficulties of accurately recognizing emotions, aiming to enhance performance in these areas. To monitor overarching temporal patterns in emotional shifts and capture intricate spatio-temporal details, two specialized models are introduced for facial expression recognition: the SISTCM and LSTM model. In terms of body motion analysis, the method, along with an ACCM, focuses on extracting essential joint features and distinguishing attributes for each motion joint. The dynamic face emotion recognition accuracy, the author [13] an advanced spatial-temporal learning network (ESTLNet), which integrates a TTEM and a SFLM. The SFLM improves spatial feature representation through a dual-channel feature fusion learning strategy.

Subsequently, the TTEM employs an encoder that integrates a gated feed-forward network with a self-attention learning network, enabling the capture of a more comprehensive temporal context, building on the spatial data obtained previously. The author [14] presents a CNN model, a DL approach, to determine age and gender from facial expressions and to detect emotions from facial pictures. Experimental results showed that the proposed model outperformed the existing baseline models, achieving 0.95 accuracy in emotion detection, 98.5% accuracy in age prediction, and 0.99 accuracy in gender recognition. In order to improve each pre-trained block of a deep convolutional neural network (DCNN), the deep layer or layers were first trained using an advanced pipeline. FER system was evaluated on face image datasets utilizing eight well-known pre-trained DCNN architectures: VGG-16, VGG-19, ResNet-18, ResNet-34, ResNet-50, ResNet-152, Inception-v3, and DenseNet-161 [15]. A sophisticated facial expression recognition (FER) system is shown in the article [16], which uses a DCNN in conjunction with TL to achieve high accuracy. The technique entails adapting a pre-trained DCNN's dense layers to FER tasks. The pre-trained blocks are progressively refined after the thick layers have been taught in order to enhance recognition performance. Eight different pre-trained DCNN models (VGG-16, VGG-19, ResNet-18, ResNet-34, ResNet-50, ResNet-152, Inception-v3, and DenseNet-161) are tested using popular face expression datasets including KDEF and JAFFE to assess the system's efficacy.

A DL-based emotion, capable of achieving high accuracy across various use cases, such as feedback analysis and facial recognition for security purposes, can be developed [17]. DCNN model that can classify five distinct emotions based on facial expressions. The model is trained, tested, and validated using a custom-collected image dataset. The paper presents deep learning methods for accurate detection and identification in facial recognition [18]. Face recognition is used to identify and verify facial features. On the other hand, real-time facial feature acquisition is handled by haar cascade detection. The work's sequential procedure consists of three separate stages: the first is identifying a human face from the camera, and the second is using a database. The last step classifies emotions as neutral, pleased, angry, sad, disgusted, and startled. A human face is validated in the last step. Emotional analysis frequently identifies facial expressions that reflect the benefits of employing neural network-based schemes, which were the main focus of an analysis of the current methods for building emotion identification systems based on human facial expressions [19]. A neural network-based approach for building an emotion identification system that comprises figuring out the neural network's logical number of layers and an efficient technique for creating training and test samples. Real-time emotion recognition has been an ongoing research focus for several decades. In one study [20], emotional expressions of children with autism and individuals with physical disabilities (such as those who are deaf, mute, or bedridden) were classified using facial landmarks and electroencephalogram (EEG) signals. This classification was achieved through CNNs and LSTM classifiers.

Additionally, developed an optical flow algorithm that employs virtual markers, which perform well under various conditions such as uneven lighting, head rotations (up to 25°), different backgrounds, and diverse skin tones. The system successfully identifies six facial emotions, surprise, fear, contempt, anger, sadness, and happiness, utilizing ten virtual markers. In the publication, the author offers a comprehensive examination of FER [21]. The author's content is a compilation of a number of reliable studies published during the past decade. Different techniques from DL and conventional ML are used for the assessment. Evaluation measures are also used to assess and compare benchmark results with a number of publicly available FER datasets. In addition to discussing both traditional machine learning and deep learning approaches, this work offers a thorough analysis of FER, pointing out any knowledge gaps that may be filled by further study. Ultimately, this study functions as a guide, providing valuable insights for both experienced researchers looking for promising research directions and newcomers to FER, equipping them with a solid understanding of the latest techniques and foundational knowledge [22]. The dual-channel expression recognition system integrates philosophy and machine learning principles to address the challenge of subtle facial expression variations. The convolutional neural network (CNN) feature extraction technique tends to overlook these subtle changes. The method's main goal is to extract Gabor features from the ROI so that they can be used as input for the first path. The algorithm efficiently extracts emotion-relevant characteristics from the ROI by separating them. A more accurate analysis of the local area results from this method guarantees that the complex characteristics of the active facial expression area are completely utilized. Pre-trained models like ResNet50, VGG19, Inception V3, and MobileNet are used in the study [23]. These pre-trained Convolutional Neural Networks (ConvNets) have their completely connected layers deleted and replaced with new fully connected layers that are tailored to the number of instructions in the particular task. Only the weights of these recently added layers are changed during training. Using the CK+ dataset, the experiment's average accuracy for emotion recognition tasks was 96%. The network analyzes key facial features and uses low-light image enhancement techniques to identify facial emotions in mask-covered images.

The approach uses the Affect Net image collection, which includes 420,299 images and eight types of facial emotions [24]. This study [25] offers a comprehensive and well-structured comparative empirical analysis of face emotion detection using local binary pattern features, convolution neural networks, and frequency domain in deep learning. By combining the seven basic emotional categories neutral, happiness, anger, fear, sadness, disgust, and surprise - the author was able to conduct facial emotion recognition (FER). To highlight the differences and similarities, examine the potential technologies that may be employed in each component of the broad framework described for a standard FER approach. This work [26] extends the use of deep learning for facial emotion recognition (FER) using the Emognition dataset, which contains 10 target emotions: neutral, fear, awe, excitement, amusement, surprise, wrath, disgust, and astonishment. This study suggests two methods for developing Convolutional Neural Network (CNN) models: create from scratch using the Taguchi technique to identify a resilient combination of hyperparameters setting, and transfer learning (fine-tuned) with pre-trained models of Inception-V3 and MobileNet-V2. The goal of this study [27] is to create a hybrid Convolutional Neural Network – Recurrent Neural Network (CNN-RNN) model that can recognize human emotions from video data by analysing facial expressions. The Emotional Wearable Dataset 2020 served as the basis for the models' development. For the classification, three models are created: MobileNetV2-RNN, InceptionV3-RNN, and custom CNN-RNN. While the MobileNetV2-RNN and InceptionV3-RNN transfer learning models provide 59% and 66% accuracy rates, respectively, the bespoke CNN-RNN model obtained 63%. The Emognition dataset in this study [28] is devoted to evaluating techniques for emotion recognition (ER) using facial expressions and physiological reactions. 43 individuals who saw brief film snippets that evoked nine distinct emotions—amusement, amazement, enthusiasm, liking, surprise, wrath, disgust, fear, and sadness—were the subjects of our data collection. In conjunction with the upper-body videos, physiological data was recorded using three wearables: EEG, BVP (2x), HR, EDA, SKT, ACC (3x), and Gyro (2x).

III. CNN-TCN ARCHITECTURE

The CNN-TCN architecture of proposed work is shown in Figure 1, represents a structured approach to developing an emotion recognition system from video data, emphasizing key stages from data preparation to final prediction.

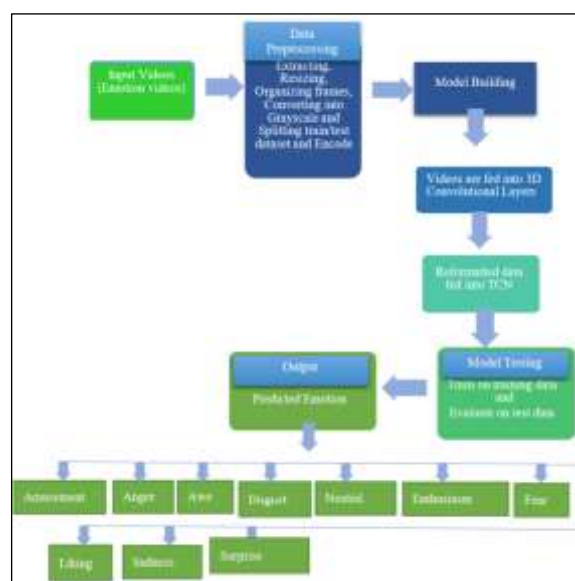


Figure 1: Overall Architecture of Enhanced CNN-TCN.

Source: Authors, (2026).

III.1 DATASET DESCRIPTION

The Proposed work begins with Input of Video Dataset, which consists of emotion-labelled videos that capture various expressions across time. These videos are processed through several steps to make them suitable for training a deep learning model that can accurately recognize emotions. The Emognition dataset includes upper body video recordings and physiological signals from 43 subjects who watch emotionally validated movie snippets that elicit ten distinct emotions. The Emognition dataset has a number of advantages over other datasets, such as highlighting the differences between positive emotions, allowing for a variety of analysis in emotion recognition (ER) from the physiology and facial expressions domain, and offering nine distinct emotions - amusement, awe, enthusiasm, liking, surprise, anger, disgust, fear, and sadness - as well as one neutral emotion [26]. This part presented the dataset used and the assessment of the suggested model. The copied videos and stock photos are no longer there. Figure 2 shows the number of videos available in the individual categories of the dataset.

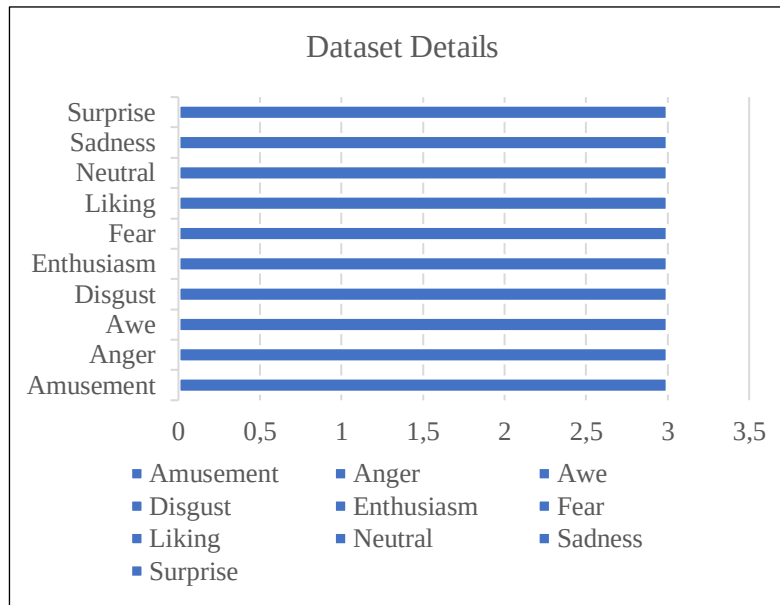


Figure 2: 10 Emotions in a Dataset for Facial Emotion Recognition.
Source: Authors, (2026).

III.2 DATA PREPROCESSING

The first step is Data Preprocessing, which plays a critical role in standardizing the raw video data. In this stage, frames are extracted from each video to represent moments in time, breaking down a continuous sequence into discrete images. These frames are resized to a standard dimension, ensuring that all input data has a consistent resolution, which simplifies further processing and model training. The frames are also converted to grayscale, reducing the data's complexity by focusing on intensity values rather than color information, which may be less relevant for emotion detection. This transformation helps lower the computational load and emphasizes structural details in expressions, which are key for recognizing emotions. The frames are then organized into sequences, capturing the temporal flow of expressions and facilitating the model's ability to recognize changes over time, which is essential for accurately detecting emotions in videos. Once the data is pre-processed, it moves to the Dataset Preparation stage. Here, the dataset is divided into training and testing subsets to allow for effective model evaluation. The labels, which indicate different emotion categories, are encoded to make them compatible with the model, often using a numerical representation that simplifies training. This stage is crucial for preparing the video data for model training. It involves extracting individual frames from the video, resizing them, and converting them to grayscale to simplify the data and reduce computational load. These frames are then organized into sequences, which capture the temporal progression of emotions across frames, making the data suitable for training temporal models like TCNs. This organized, preprocessed data is essential for effective emotion recognition from videos.

III.3 MODEL BUILDING

In the model-building phase, the architecture is constructed with layers suited to spatial and temporal feature extraction. Conv3D layers are used to capture spatial details within each frame, identifying subtle features related to facial expressions. Temporal Convolutional Network (TCN) layers are then applied to learn temporal patterns across frames, helping the model understand how emotions evolve. During Model Training, the model is trained on the prepared data, learning to recognize and classify emotions. It is evaluated on test data to ensure its generalization ability. In the Prediction from Video stage, the trained model can now extract frames from new videos and predict emotions based on learned patterns, producing the final output, which is the predicted emotion, concluding the pipeline.

III.3.1 CNN

Deep Learning (DL), a subfield of Machine Learning (ML), centers on the process of learning progressively more abstract representations through multiple layers. A model's depth indicates the total number of layers used in data processing. The three primary components of a standard neural network.

The model's depth corresponds to the number of layers that contribute to transforming the data. Convolutional Neural Networks (CNNs) differ from traditional neural networks by using convolution operations in their layers instead of relying solely on matrix multiplication. The convolution operation is mathematically represented by equation (1).

$$s(t) = (x * w)(t) \quad (1)$$

Here,

x – input

w – kernel filter

$s(t)$ – output is called feature map

Here t might be spatial location or time space

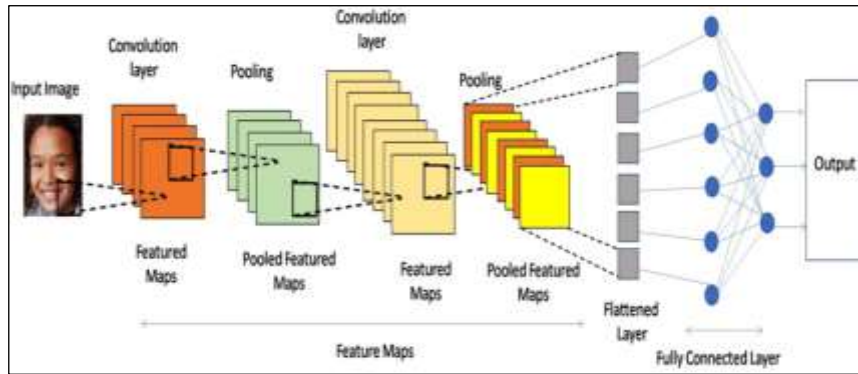


Figure 3: CNN Architecture.

Source: Authors, (2026).

Figure 3 depicts the convolutional neural network model's architecture. There are three steps to CNN, which are to perform several convolutions using filters concurrently. A nonlinear activation, such as a rectified linear activation function, is carried out in the second stage. If the input is positive, the output is generated directly; if not, it is set to zero. In the third step, a pooling technique is used to minimize the feature map's size. Max pooling, which chooses inside a window; average pooling, which substitutes the average value for window values; and weighted average pooling are common pooling techniques. Other variations of pooling processes can also be used. The result from the final pooling layer is sent to the FC layer at the end of a CNN. The input is categorized into multiple groups by each node in this layer, which is connected to every other node in the layer above.

III.3.2 Tcn (Temporal Convolution Network)

TCNs are CNs designed for analysing applications with causal restrictions [11]. The input sequence $x_0, \dots, x_t$ and its matching output sequence $y_0, \dots, y_t$ are supplied for training a sequence analysing network. The network can learn to predict the output sequence $y_0, \dots, y_t$ by evaluating the difference between the expected sequence and the actual output using a loss function. The network's causal constraint guarantees that each prediction $y_0, \dots, y_t$ is solely impacted by the inputs that came before it $x_0, \dots, x_t$ and not by inputs that came after it $x_{t+1}, \dots, x_t$. The noisy frames serve as the input sequence in speech enhancement software, while the clean frames create the output sequence.

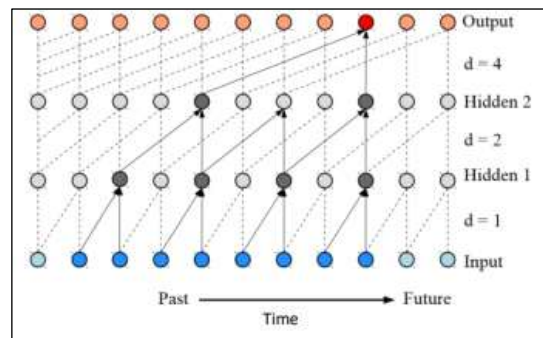


Figure 4: Expanded Causal Convolution Using a Size 2 Filter.

Source: Authors, (2026).

TCNs are built with a combination to enforce the causal constraint. The causal convolutions prevent future information from impacting previous data, maintaining temporal order. In contrast, the network can capture longer-term temporal correlations because of dilated convolutions, which improve the receptive field. The network's ability to analyse historical data increases as the receptive field expands. An opened and causal convolution with a filter size of two is shown in Figure 4.

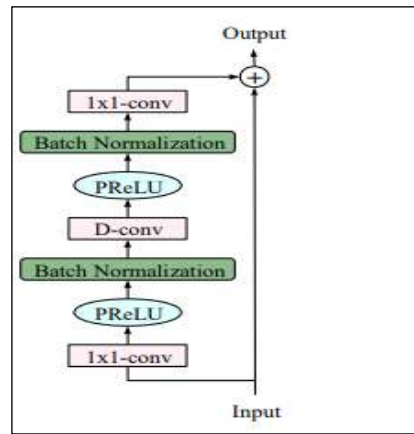


Figure 5: Residual Block.
Source: Authors, (2026).

A TCN incorporates residual blocks to facilitate the effective training of deep networks through residual learning. The residual block employed in this study is very similar to the one used here, as Figure 5 illustrates. Three different kinds of convolutions are included in this block: depth-wise, 1x1 input, and 1x1 output convolutions. In order to provide compatibility between the input and output through element-wise addition, the input convolution increases the quantity of entering networks. In contrast, the yield convolution restores the channel count to its initial value. By applying one filter per input channel and maintaining the channel count, the depth-wise convolution aids in lowering the sum of parameters. In contrast, traditional difficulty operations use as many filters as there are input channels for each output channel after the input and intermediate convolutions to further refine the model.

IV. ENHANCED CNN-TCN MODEL

The Enhanced CNN-TCN model manages the temporal and spatial dimensions of the video. The classifier first incorporates 3D Convolution layers. The result should then be reformatted to be fed into the 1D Convolution layers of the TCN section.

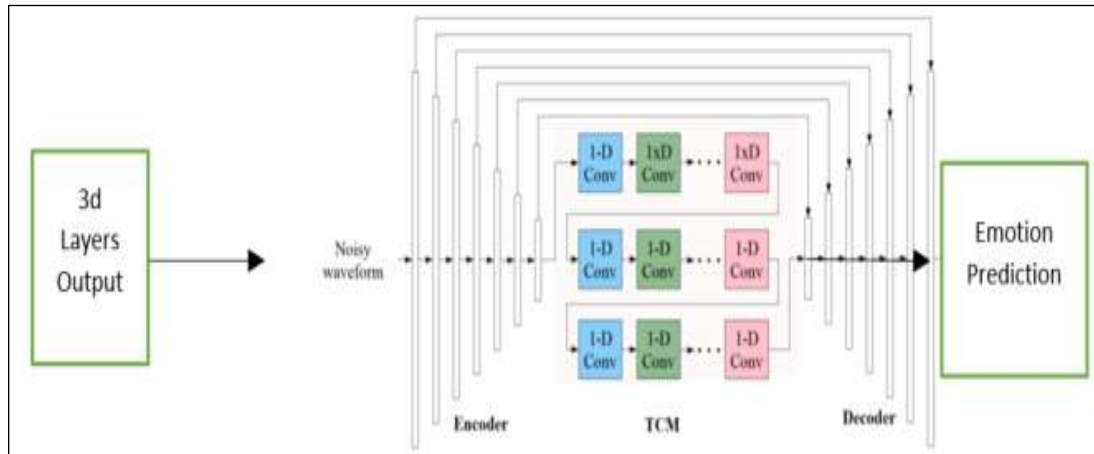


Figure 6: Enhanced CNN-TCN Model.
Source: Authors, (2026).

This model effectively combines Spatial Feature and Temporal Sequence Modeling using CNN and TCN. The three main parts of the Enhanced CNN-TCN architecture are the encoder, decoder, and Temporal Convolution Module (TCM). While the encoder and decoder employ 1D causal convolution layers, the TCM is constructed with 1D Convolution layers. A block form of the framework is shown in Figure 5. When T is the number of frames, 320 is the frame width, and 1 is the number of input channels, the encoder's input shape for the noisy frame sequence is $T \times 320 \times 1$. The first layer of the encoder enlarges the input channels from 1 to 64 to produce an output form of $T \times 320 \times 64$. Subsequent layers of the encoder execute convolutions with a stride two over seven layers, thus decreasing the frame dimension. $T \times 4 \times 64$ is the encoder's final output. The time dimension stays constant during this procedure, guaranteeing that the output's frame count corresponds to the inputs. Batch and parametric normalization, with ReLU as the activation function, come after each encoder layer. The encoder produces a one-dimensional signal of $T \times 256$. This reshaped signal is then processed by the Temporal Convolutional Module (TCM), which maintains the same output size. The TCM is structured with three stacked dilation blocks; each made up of six residual blocks, where dilation rates increase exponentially across layers. Specifically, the dilation rates used are 1, 2, 4, 8, 16, and 32. The decoder, designed as a mirrored version of the encoder, incorporates multiple 1D layers. At each layer, the decoder's output is combined with the corresponding encoder layer's output. Dropout at a rate of 0.3 is applied to the encoder's skip connections during training. Each decoder layer also includes batch normalization and uses a parametric ReLU activation function for enhanced nonlinearity. The encoder and decoder's hyperparameters are specified as Filter Height $\times$ Filter Width, with strides listed for both time and frame. Small braces around entries mark residual blocks, and the TCN's hyperparameters include filter size, dilation rate, and output channels.

Input:
Training dataset
Testing dataset
Initial values for CNN and TCN hyperparameters
Steps:
1. For each new population, do the following:
2. For each chromosome, do the following:
Test the combined CNNTCN model with Nadam optimizer and calculate accuracy
Next
3. Perform mutation:
(i) Select a random particle (hyperparameter)
(ii) Select a random position (hyperparameter, such as filter size or dilation rate)
4. Perform crossover:
(i) Select two random particles (hyperparameter sets)
(ii) Swap their values
(iii) Calculate the fitness of each set
(iv) Select the best hyperparameter set
5. Calculate the best values for CNN and TCN hyperparameters with the highest accuracy (H_Acc)
6. If termination conditions are satisfied:
Return the optimized CNN and TCN hyperparameters
Break
8. End If
9. Next
Output:
Optimized CNN and TCN hyperparameters

Figure 7: Pseudocode for Enhanced CNN-TCN model with Nadam optimizer.
Source: Authors, (2026).

V. RESULTS AND DISCUSSION

V.1 TRAINED OUTPUT OF ENHANCED CNN-TCN

The proposed system layer architecture is shown in Figure 8. Deep learning architectures consist of layers that process data in phases to extract significant characteristics and provide predictions. In a deep learning model, every layer serves a distinct purpose and usually works with inputs from layers before it. Let's review the typical layer types found in deep learning systems, as well as their functions and interactions with one another.

Layer (type)	Output shape	Param #
conv2d_2 (Conv2D)	(None, 8, 20, 40, 32)	400
max_pooling2d_2 (MaxPooling2D)	(None, 8, 10, 20, 32)	0
conv2d_3 (Conv2D)	(None, 8, 10, 20, 32)	400
max_pooling2d_3 (MaxPooling2D)	(None, 8, 5, 10, 32)	0
time_distributed_1 (TimeDistributed)	(None, 8, 1000)	0
conv1d_3 (Conv1D)	(None, 8, 32)	2,048
conv1d_4 (Conv1D)	(None, 8, 32)	2,048
conv1d_5 (Conv1D)	(None, 8, 32)	2,048
flatten_3 (Flatten)	(None, 256)	0
dense_3 (Dense)	(None, 256)	65,536
dropout_3 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 1)	256

Total params: 1,112,448 (9.93 MB)
Trainable params: 1,112,448 (9.93 MB)
Non-trainable params: 0 (0.00 B)

Figure 8: Trained output of Enhanced CNN-TCN Model.
Source: Authors, (2026).

V.2 PERFORMANCE AND ERROR METRICS OF THE ENHANCED CNN-TCN WITH EXISTING DEEP LEARNING MODELS.

Depending on the particular objective (e.g., classification, regression, time-series forecasting), different metrics are used to assess a Deep Convolutional Neural Network (Pulse DCNN) in conjunction with Temporal Convolutional Networks (TCN) with Nadam optimizer. Consider the following important assessment measures.

$$Accuracy = \frac{\text{correct prediction}}{\text{Total prediction}} \quad (2)$$

$$precision = \frac{\text{True Positives}}{\text{True positive} + \text{False positive}} \quad (3)$$

$$Recall = \frac{\text{True Positives}}{\text{True Positive} + \text{False Negative}} \quad (4)$$

$$F1 \text{ score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

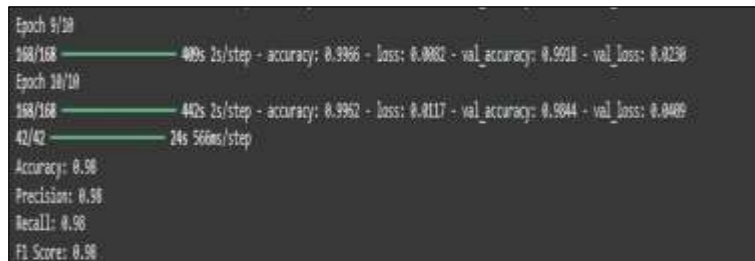


Figure 9: Performance metrics of Enhanced CNN-TCN algorithm with NADAM optimizer.
Source: Authors, (2026).

Figure 9. displays the performance attained with the suggested Enhanced CNN-TCN with Nadam optimizer, which has a value of 98 percent for all parameters, including F1 score, accuracy, precision, and recall calculated using equation (2)-(5).

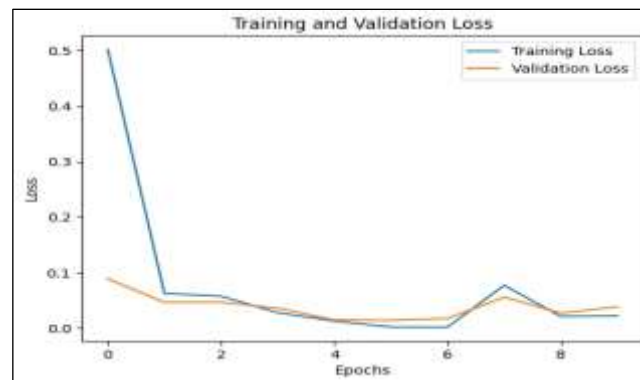


Figure 10: CNN-TCN model's training and validation loss.
Source: Authors, (2026).

The Enhanced CNN-TCN with Nadam optimizer training losses and testing accuracy is shown in Figure 10. The hyperparameter of our proposed system is set by using the Nadam optimizer. The losses of training and validation are shown in the figure for ten epoch steps. Figures 11 and 12 show the emotion recognition from the video frames using the Enhanced CNN-TCN model. The TCN model with CNN reduces the processing time.

Table 1: Performance metrics of the Enhanced CNN-TCN with Existing Model.

Models	Precision (%)	Recall (%)	F1 score (%)	Accuracy (%)
ResNet	95.69	95.69	95.82	95.67
Inception v3	99.64	97.38	98.51	99.65
GoogleNet	99.64	98.44	99.04	99.65
Hybrid CNN-TCN	98	98	98	98

Source: Authors, (2026).

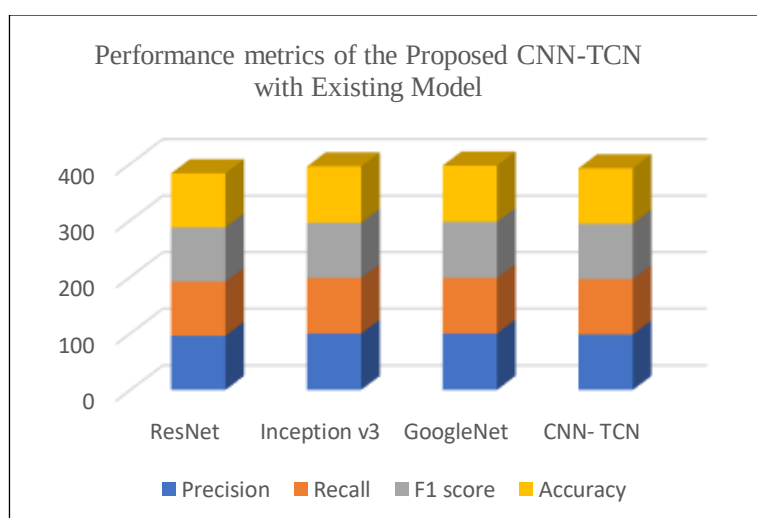


Figure 11: Performance Metrics Comparison of Proposed with Predefined Deep Learning Model.
Source: Authors, (2026).

CNN's performance utilizing the data without any improvements is shown in Figure 10. As shown in Table 1 and Figure 10, the efficacy of the suggested approach (TCN) is evaluated in comparison to the formerly widely used pre-trained deep learning models, including ResNet, Inception v3, and GoogleNet.

As the CNN's performance improved after the Nadam adjustment of its hyperparameters, Table 1 and Figure 11 show that the CNN-TCN classifier generated the best results. CNN-TCN outperforms earlier techniques with accuracy, precision, recall, and f1 score of 98%.

Table 2: Compare Error metrics of the Enhanced CNN-TCN with Existing Model.

Models	MAE	MSE	R^2
ResNet	0.021	0.059	0.901
Inception v3	0.019	0.061	0.898
GoogleNet	0.022	0.058	0.902
Enhanced CNN-TCN	0.02	0.06	0.95

Source: Authors, (2026).

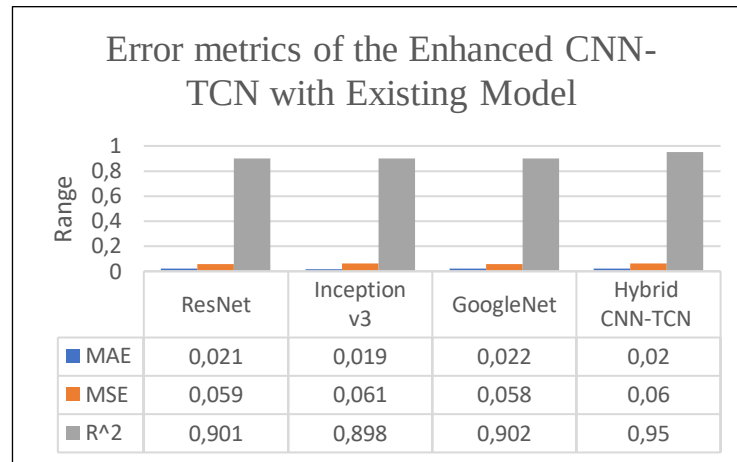


Figure 12: Error metrics of the Enhanced CNN-TCN with Existing Model.

Source: Authors, (2026).

The table 2 and figure 12 compares the performance metrics (MAE, MSE, and R^2) of ResNet, Inception v3, GoogleNet, and the proposed Hybrid CNN-TCN model. The Hybrid CNN-TCN outperforms other models with an MAE of 0.02, MSE of 0.06, and a superior R^2 of 0.95, reflecting its high accuracy and robust fit. In comparison, ResNet, Inception v3, and GoogleNet show slightly higher error rates and marginally lower R^2 values, indicating the Enhanced CNN-TCN's effectiveness in achieving minimal errors and maximum data fit.

V.3 EMOTION PREDICTION

Here, sample of predicted emotion

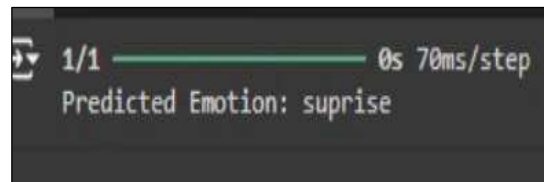


Figure 13: Surprise emotion recognition.

Source: Authors, (2026).

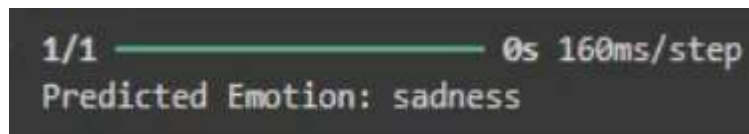


Figure 14: Predict sadness emotion from the facial model using CNN-TCN with Nadam optimizer.

Source: Authors, (2026).

V.4 ERROR PERFORMANCE OF PROPOSED MODEL

Mean Squared Error (MSE):

Calculates average squared difference between predicted and actual values.

$$MSE = \frac{\sum (y_i - p_i)^2}{n}$$

y_i = ith observed value

p_i = corresponding predicted value for y_i ,

n = number of observations.

Mean Absolute Error (MAE):

Calculates average absolute difference between predicted and actual values.

$$\text{MAE} = (1/n) \sum_{i=1}^n |y_i - \hat{y}_i|$$

y_i = the actual value of the i th observations

$\hat{y}_i$ = the predicted value of the i th observations

n =number of observations

R-squared (R^2):

A higher R^2 indicates better fit between the model and the data.

$$R^2 = 1 - \text{SSE} / \text{SST}$$

SSE = Sum of Squared Errors

SST = Total Sum of Squares

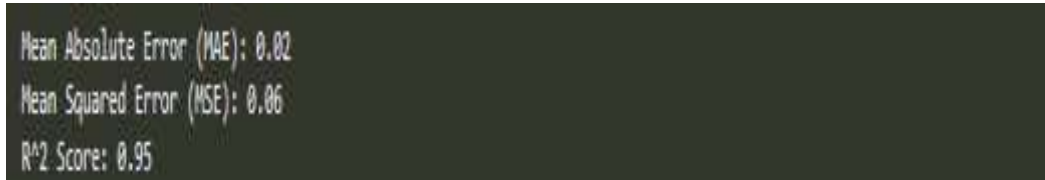


Figure 15: MAE, MSE and R^2 value for 10 epoch training steps.

Source: Authors, (2026).

The figure 15 shows the error performance such as Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error of proposed system is achieved 0.02, 0.06 and 0.95 for 10 epoch steps.

VI. CONCLUSION AND FUTURE WORK

For face motion recognition, this study proposed a convolutional neural network (CNN) with a temporal convolutional network (TCN) architecture. This study suggests using dynamic video frames to recognize emotions with minimal processing time. The three steps of emotion recognition are video frame extraction, preprocessing following video frame extraction, and face expression detection using CNN-TCN with less processing time. In future to assist families and medical professionals in the early identification of autism, this study proposes a CNN-TCN architecture for facial emotion detection. The performance, like accuracy, precision, recall, f1 score, training loss, and testing accuracy, is calculated and compared with pre-trained deep learning models like ResNet, Inception v3, and GoogleNet. The Error performance MAE and MSE and Root Mean Square Calculated. This study attempts to fill the gap in the current datasets, in future identify hybrid emotion and Emoji.

VII. AUTHOR'S CONTRIBUTION

Conceptualization: B. Maidhili, Dr.R. Thirumalai Selvi.

Methodology: B. Maidhili, Dr.R. Thirumalai Selvi.

Investigation: B. Maidhili, Dr.R. Thirumalai Selvi.

Discussion of results: B. Maidhili, Dr.R. Thirumalai Selvi.

Writing – Original Draft: B. Maidhili, Dr.R. Thirumalai Selvi.

Writing – Review and Editing: B. Maidhili, Dr.R. Thirumalai Selvi.

Resources: B. Maidhili, Dr.R. Thirumalai Selvi.

Supervision: B. Maidhili, Dr.R. Thirumalai Selvi.

Approval of the final text: B. Maidhili, Dr.R. Thirumalai Selvi.

VIII. REFERENCES

- [1] Ketan Sarvkar, R. Senkamalavalli, S. Raghavendra, J. Santosh Kumar, R. Manjunath, Sushma Jaiswal, Facial emotion recognition using convolutional neural networks, Materials Today: Proceedings, Volume 80, Part 3, 2023, Pages 3560-3564, ISSN 2214-7853, <https://doi.org/10.1016/j.matpr.2021.07.297>.
- [2] Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, Xiaoming Zhao, Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects, Expert Systems with Applications, Volume 237, Part C, 2024, 121692, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2023.121692>.
- [3] L. Atymtayeva, M. Kanatov, A. M. A Musleh, and G. Tulemissova, "Fast Facial Expression Recognition System: Selection of Models", Appl. Math. Inf. Sci. 17, No. 2, 375-383 (2023).
- [4] Wafa Mellouk, Wahida Handouzi, "Facial emotion recognition using deep learning: review and insights", Procedia Computer Science, Volume 175, 2020, Pages 689-694, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2020.07.101>.
- [5] Ketan Sarvkar, R. Senkamalavalli, S. Raghavendra, J. Santosh Kumar, R. Manjunath, Sushma Jaiswal, "Facial emotion recognition using convolutional neural networks", Materials Today: Proceedings, Volume 80, Part 3, 2023, Pages 3560-3564.
- [6] Bakariya, B., Singh, A., Singh, H. et al. Facial emotion recognition and music recommendation system using CNN-based deep learning techniques. Evolving Systems 15, 641–658 (2024). <https://doi.org/10.1007/s12530-023-09506-z>
- [7] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, Aibin Huang, Yigang Wang, "POSTER++: A simpler and stronger facial expression recognition network", Pattern Recognition, Volume 157, 2025, 110951, ISSN 0031-3203, <https://doi.org/10.1016/j.patcog.2024.110951>.
- [8] Kumari, N., Bhatia, R. Saliency map and deep learning based efficient facial emotion recognition technique for facial images. Multimed Tools Appl 83, 36841–36864 (2024). <https://doi.org/10.1007/s11042-023-16220-0>.

- [9] Kumari, N., Bhatia, R. Saliency map and deep learning based efficient facial emotion recognition technique for facial images. *Multimed Tools Appl* 83, 36841–36864 (2024). <https://doi.org/10.1007/s11042-023-16220-0>
- [10] R. Dong and K. -M. Lam, "Bi-Center Loss for Compound Facial Expression Recognition," in *IEEE Signal Processing Letters*, vol. 31, pp. 641-645, 2024, doi: 10.1109/LSP.2024.3364055
- [11] Talaat, F.M., Ali, Z.H., Mostafa, R.R. et al. Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children. *Soft Comput* 28, 6695–6708 (2024). <https://doi.org/10.1007/s00500-023-09477-y>
- [12] Jie Wei, Guanyu Hu, Xinyu Yang, Anh Tuan Luu, Yizhuo Dong, Learning facial expression and body gesture visual information for video emotion recognition, *Expert Systems with Applications*, Volume 237, Part A, 2024, 121419, ISSN 0957-4174, <https://doi.org/10.1016/j.eswa.2023.121419>.
- [13] Weijun Gong, Yurong Qian, Weihang Zhou, Hongyong Leng, Enhanced spatial-temporal learning network for dynamic facial expression recognition, *Biomedical Signal Processing and Control*, Volume 88, Part C, 2024, 105316, ISSN 1746-8094, <https://doi.org/10.1016/j.bspc.2023.105316>
- [14] Khattak, A., Asghar, M.Z., Ali, M. et al. An efficient deep learning technique for facial emotion recognition. *Multimed Tools Appl* 81, 1649–1683 (2022).
- [15] Akhand, M.A.H.; Roy, S.; Siddique, N.; Kamal, M.A.S.; Shimamura, T. Facial Emotion Recognition Using Transfer Learning in the Deep CNN. *Electronics* 2021, 10, 1036. <https://doi.org/10.3390/electronics10091036>
- [16] Sharmeen M.Saleem Abdullah Abdullah, Siddeeq Y. Ameen Ameen, Mohammed A. M. Sadeeq, Subhi Zeebaree, "Multimodal Emotion Recognition using Deep Learning", *JASTT*, vol. 2, no. 01, pp. 73–79, May 2021, doi: 10.38094/jastt20291.
- [17] E. Pranav, S. Kamal, C. Satheesh Chandran and M. H. Supriya, "Facial Emotion Recognition Using Deep Convolutional Neural Network," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 317-320, doi: 10.1109/ICACCS48705.2020.9074302.
- [18] Shaik Asif Hussain and Ahlam Salim Abdallah Al Balushi, "A real time face emotion classification and recognition using deep learning model" 2020 *J. Phys.: Conf. Ser.* 1432 012087 DOI 10.1088/1742-6596/1432/1/012087.
- [19] Ekaterina Ivanova, Georgii Borzunov, Optimization of machine learning algorithm of emotion recognition in terms of human facial expressions, *Procedia Computer Science*, Volume 169, 2020, Pages 244-248, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2020.02.143>.
- [20] Aya Hassouneh, A.M. Mutawa, M. Murugappan, Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods, *Informatics in Medicine Unlocked*, Volume 20, 2020, 100372, ISSN 2352-9148, <https://doi.org/10.1016/j.imu.2020.100372>.
- [21] Khan, A.R. Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges. *Information* 2022, 13, 268. <https://doi.org/10.3390/info13060268>
- [22] Song Zhenjie, "Facial Expression Emotion Recognition Model Integrating Philosophy and Machine Learning Theory", *Frontiers in Psychology*, Vol.12, 2021, 10.3389/fpsyg.2021.759485
- [23] Chowdary, M.K., Nguyen, T.N. & Hemanth, D.J. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Comput & Applic* 35, 23311–23328 (2023). <https://doi.org/10.1007/s00521-021-06012-8>
- [24] Mukhiddinov, M.; Djuraev, O.; Akhmedov, F.; Mukhamadiyev, A.; Cho, J. Masked Face Emotion Recognition Based on Facial Landmarks and Deep Learning Approaches for Visually Impaired People. *Sensors* 2023, 23, 1080. <https://doi.org/10.3390/s23031080>
- [25] Kumar, S., Sagar, V. & Punetha, D. A comparative study on facial expression recognition using local binary patterns, convolutional neural network and frequency neural network. *Multimed Tools Appl* 82, 24369–24385 (2023). <https://doi.org/10.1007/s11042-023-14753-y>
- [26] Agung, E.S., Rifai, A.P. & Wijayanto, T. Image-based facial emotion recognition using convolutional neural network on emognition dataset. *Sci Rep* 14, 14429 (2024). <https://doi.org/10.1038/s41598-024-65276-x>
- [27] Haposan Vincentius Manalu, Achmad Pratama Rifai, "Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm", *Intelligent Systems with Applications*, Vol 21, 2024, 200339, ISSN 2667-3053, <https://doi.org/10.1016/j.iswa.2024.200339>.
- [28] Saganowski, S., Komoszyńska, J., Behnke, M. et al. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Sci Data* 9, 158 (2022). <https://doi.org/10.1038/s41597-022-01262-0>