DOI: 10.21917/ijdsml.2025.0172

# REAL-TIME FACIAL EXPRESSION RECOGNITION USING DEEP LEARNING

**Purva Kalambate, Seema Hanchate, Poonam More and Janhavi Askar**
*Department of Electronics and Communication Engineering, SNDT Women's University, India*

***Abstract***

***In many applications like emotional analysis, human- computer interaction, mental health monitoring, sentiment analysis and surveillance systems, real-time facial expression detection has become a vital role. Real-time facial expression recognition systems recognize human emotions which improves user experiences and system reactions. The Convolutional Neural Network (CNN) algorithm with three convolution layers is used for human expression recognition. Two different datasets, FER2013 and CK +48 are used for training the proposed system. This dataset provides a diverse range of facial expressions for training and evaluation. The proposed system has trained for seven distinct emotions: anger, sadness, happiness, neutral, disgust, surprise, and fear. Many existing systems are accurate but suffer from complexity in their model architecture and code model implementation. The proposed system achieves a notable accuracy of 92%, outperforming many existing models in the field. The proposed model has high accuracy with less complexity which is suitable for real-time deployment. The proposed solution streamlines the design while preserving performance, resulting in greater ease of utilization and less computation requirements.***

***Keywords:***

***Facial Expression, CNN, Feature Extraction, Emotions Facial Landmarks, Emotion Detection***

## 1. INTRODUCTION

Facial expression recognition (FER) is crucial for understanding human emotions and intentions without words. Researchers have been exploring various methods, including deep learning (DL), to teach computers how to recognize facial expressions like humans do [1] [2]. However, it's challenging because faces can look different under different conditions, making it hard for computers to understand them well [3]. To improve FER, scientists are trying different approaches. Some are developing new computer programs to understand facial features correctly, while others are combining different Artificial Intelligence (AI) techniques to analyze facial expressions in videos more accurately [4] [5]. The FER (Facial Expression recognition) system consists of several critical steps. First, facial expressions are detected using a Convolutional Neural Network (CNN), which identifies faces within images. Once a face is detected, key facial landmarks: such as the eyes, nose, mouth and eyebrows are identified. These landmarks serve as essential reference points for analyzing facial expressions. Features like texture, motion vectors, and geometric attributes are extracted from these landmarks and used to differentiate between various expressions. The CNN algorithm then trains the models using these extracted attributes.

Thus, this research aims to advance the field of facial expression recognition by developing a robust system that can accurately classify expressions across varying image qualities, hence enhancing to understand the human emotions and helping to raise the applications of FER technology in real-world settings. The paper is arranged as follows: The introduction is followed by a second section. Section 2 presents Related Work, focusing on emotion/expression recognition and the various approaches considered by researchers. Next, section 3 provides a Background, detailing the main components of the proposed architecture. Section 4 summarizes the datasets used in this study. In section 5, the architecture is presented, followed by the experiments and results in Section 6. Finally, Section 7 concludes the paper, summarizing the key findings and outlining future research directions.

## 2. RELATED WORK

The work of Ekman and Friesen in 1978 served as the foundation for facial expression recognition (FER), which has undergone substantial development. These psychologists from America identified six fundamental types of human facial expressions. The foundation for current FER research which mainly concentrates on feature extraction and classification, was established by their creation of the Facial Action Coding System (FACS), which recognizes facial expressions by examining facial muscle movements and combinations of separate motor units [6, 7]. The use of Convolutional Neural Networks (CNNs) has revolutionized FER by automating the feature extraction process, improving accuracy and eliminating the tedious manual efforts required by traditional methods [8]. EfficientNet and MobileNet architecture-based Lightweight FER models are proposed for emotional feature extraction from facial images [9]. This proposed model uses the concept of robust data mining and modifies the SoftMax function for training. In facial videos, an effective neural network model simultaneously detects involvement and recognizes emotions at the individual and group levels [10]. CNN pulls unified emotional elements from every frame of the previous item. The statistical functions are used to combine the information of multiple frames into a video descriptor. CNN-based models incorporate convolutional and pooling operations to capture spatial hierarchies in facial images They integrate critical parameters such as weights and partial values, into excitation and loss functions to enhance classification [11]. The work of Mliki et al. emphasizes extracting geometric features like the eyes, eyebrows and mouth which are automatically detected and segmented, followed by computations of interest points for prediction rules [12]. Szu-Yin Lin proposed a model that analyzes continuous facial expressions and mood changes, enhancing recognition accuracy over traditional image-based methods [13]. Similarly, Yanti Liliana focused on simple geometric feature extractions, enabling applications such as pain, lie and cheat detection [14]. For addressing the computational challenges, Burkert's DeXpression network introduces the FeatEx structure, combining multiple convolutional and pooling layers with minimal effort, delivering competitive accuracy [15]. Competitions such as Emotion Recognition in the Wild (EmotiW) showcase advanced FER techniques. Their winners often use

ensembles of deep CNNs combined with multi-modal features like audio, facial and body pose data. However, these complex models are unsuitable for real-time applications in low resource environments [16]. In evaluating cross-quality face recognition, it has been observed that improving deep learning's ability to handle significant quality changes between face images is crucial [17]. Given the subjective nature of emotions and high Bayes error rates, FER systems must contend with images that can yield multiple valid interpretations [18]. Emotion recognition from diverse facial views, especially in uncontrolled environments, is increasingly major for secure and smart living [19]. Khaireddin et al. demonstrated the potential of optimization techniques, such as Cosine Annealing, to improve accuracy from 73.06% to 73.28% [20]. Meanwhile, attention mechanisms allow shallower networks to compete with deeper ones by focusing on key facial regions, further improving efficiency and accuracy [21]. The deep learning methods used by the author to interpret facial emotions through human's expressions. Six emotions were identified using three benchmark data sets: Google Facial Expression Comparison Dataset, Extended Yale Face Database B and LFW [22].

## 3. PROPOSED MODEL

After performing multiple steps, facial expressions are carried out from the images. All these stages work together to recognize and interpret emotions. In the first step, face detection takes place from the image with the help of different algorithms typically Haar cascades or Convolutional Neural Networks (CNNs).

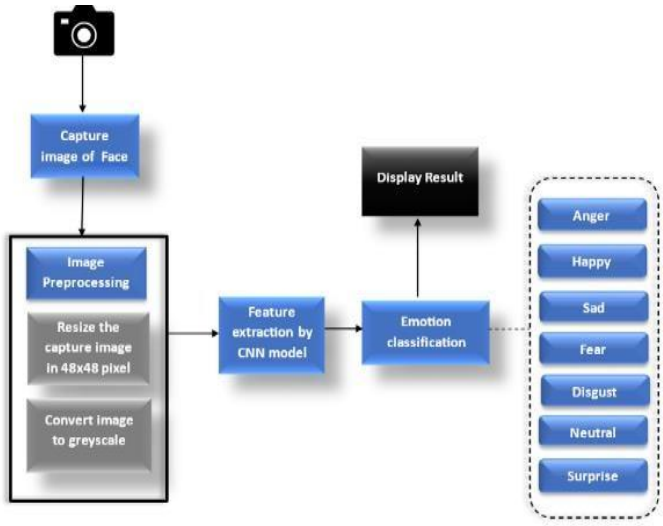


Fig.1. Proposed System Architecture

These algorithms locate faces and determine their position within the image. Once a face has been identified, the next phase involves the recognition of key facial landmarks, such as the eyes, nose, mouth and eyebrows. CNN-based algorithms are employed to perform this step, allowing crucial reference points to be identified for further analysis.

After the landmarks are detected, specific regions of interest (ROIs) such as the mouth or the area around the eyes and eyebrows, are isolated for more detailed scrutiny. These regions are essential for providing insight into particular emotions, with the mouth indicating smiling or frowning and the eyebrows revealing surprise or anger. These regions are isolated through geometric measurements based on the detected facial landmarks. Once these regions have been extracted, features such as texture, motion vectors or geometric attributes are gathered. These features are used to distinguish between different expressions, as characteristics like the distance between the eyebrows or the curvature of the lips can reveal specific emotions. Finally, machine learning techniques, particularly CNNs are utilized to train models on the extracted features. The Fig.1 shows the steps which is involved in system design.

## 4. IMPLEMENTATION

### 4.1 METHODOLOGY

The methodology for real-time facial expression recognition using a CNN model involves several steps:

#### *4.1.1 Data Collection and Preprocessing:*

The Fig.2 shows the pre-processing of images present in the dataset. The images are pre-processed to enhance the quality and suitability for input into the CNN model. Each image is resized to a standardized dimension of 48x48 pixels to ensure uniformity across the dataset. The images are converted to grayscale to simplify processing and reduce computational complexity while preserving essential facial expression information.

Fig.2. Pre-processing of image

#### *4.1.2 CNN Model Architecture Design:*

A convolutional neural network (CNN) architecture is designed to complete the task of facial expression recognition (Fig.3). The architecture consists of multiple layers, including convolutional layers for feature extraction and pooling layers for spatial down sampling.

The Fig.3 represents the CNN Architecture. Convolutional Layers (Conv2D) apply convolutional filters to the input. Each convolutional layer typically extracts different features from the input image. Here, three convolutional layers are used:

The first convolutional layer has 16 filters of size 5x5, using ReLU (Rectified Linear Activation) as the activation function. The second convolutional layer has 32 filters of size 5x5, also using ReLU activation. The third convolutional layer has 64 filters of size 3x3, again using ReLU activation.

Pooling Layers (MaxPooling2D) down-sample the input representation, reducing its dimensionality. Max- pooling takes

the maximum value from each patch of the feature map. After each convolutional layer, there is a max-pooling layer with a pool size of (2, 2). The flattened layer converts the multi-dimensional features map into a one-dimensional array. This prepares the data for input into the fully connected layers. Fully Connected Layers (Dense) connect every neuron in one layer to every neuron in the next layer. There are two fully connected layers in this model:

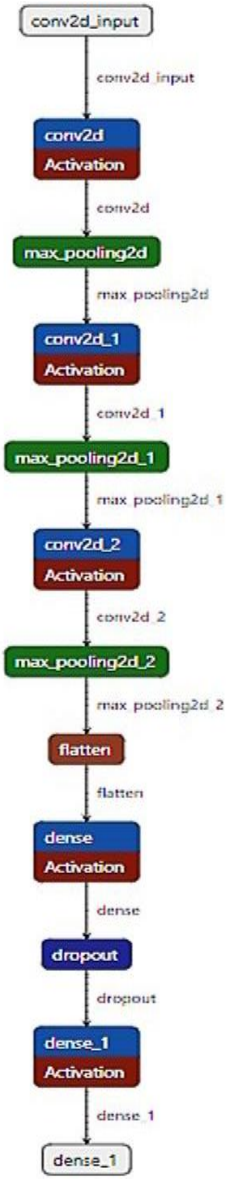


Fig.3. CNN Architecture

The first dense layer has 128 neurons with ReLU activation. The second dense layer has 7 neurons (since it's a classification task with 7 classes) with softmax activation which gives the probability distribution over the classes.

Compile () method is called on the model to configure its learning process. The loss function is categorical cross-entropy which works well for multi-class classification tasks. The Adam optimizer, an adaptive learning rate optimization technique, employed. Accuracy is used as the metric for assessing the model during training.

The Fig.4 represents CNN Architecture Output in which convolutional layers are used for feature extraction and pooling layers for spatial down-sampling. The CNN model is trained on a training set of facial images. During training, the model learns to extract discriminative features from facial expressions and map them to the appropriate emotion labels.

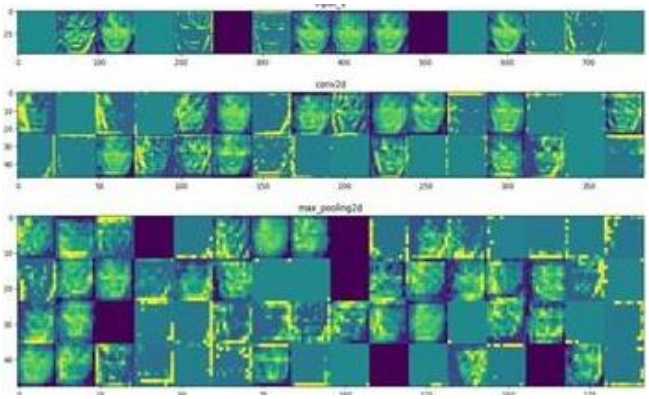

Fig.4. CNN Architecture Output for Happy face

This approach uses forward and backward gradient propagation to update model parameters (weights and biases) utilizing Adam optimization algorithms. During the training, the model learns to identify patterns and features associated with different facial expressions across the seven emotion classes. The training process involves iteratively presenting the images to the model and adjusting its parameters using optimization techniques such as gradient descent to minimize the error between predicted and actual emotion labels. Once the model is trained, its performance is tested on the unseen testing set. The model predicts the emotion labels for each test image, and these predictions are compared to the ground truth labels to calculate performance metrics such as accuracy, precision and recall, which provide a comprehensive assessment of the model's effectiveness.


Fig.5. Area of interest in a face for different emotions

The Fig.5 shows the area of interest in a face for different emotions. These facial landmarks are identified within the image to facilitate emotion analysis. These landmarks, which include distinguishing features like the corners of the eyes, nose and mouth are automatically discovered during the preprocessing step.

## 4.2 IMPORTANCE OF CNN ALGORITHM

Convolutional Neural Networks (CNNs) are a specialized class of deep learning models designed to process grid-like data, such as images. CNNs are particularly effective for applications such as image identification, object detection and segmentation because they can automatically recognize and learn spatial hierarchies in data. The core component of a CNN is the convolutional layer which uses filters to scan over the input image, detecting patterns such as edges, textures and complex shapes. These enable CNNs to capture the local information, while pooling layers reduce the dimensionality of the data, making the network more efficient and robust to small translations or distortions. CNNs use activation functions such as ReLU (Rectified Linear Unit) to introduce nonlinearity, which aids in learning complex patterns.

After numerous convolutional and pooling layers, the network typically flattens the collected feature maps into a 1D vector before passing it through fully connected layers for final classification or prediction. Dropout layers are used during training to prevent overfitting by randomly deactivating neurons and encouraging the network to generalize better. The performance of CNN based models compared with accuracy of proposed CNN Architecture in Table.1.

Table.1. Comparison of different CNN based model with Proposed model

| References | Dataset | Methodology | Accuracy |
|---|---|---|---|
| [7] | FER2013 | (7, Conv set) | 72.49 |
| [8] | FER2013 | (64, 128, 512) | 64 |
| Proposed system | FER2013 and CK+48 | (16, 32, 64) | 92.14 |

## 4.3 DATASET

The Fig.6 represents random images of different persons with different emotions from dataset. The dataset used in this study consists of 1400 images, evenly distributed across seven distinct emotion classes with each class containing 200 images shown in Fig.7. These images are sourced from two datasets, FER2013 and CK+48, providing a diverse range of facial expressions for training and evaluation.

Fig.6. Random images of different persons and emotions

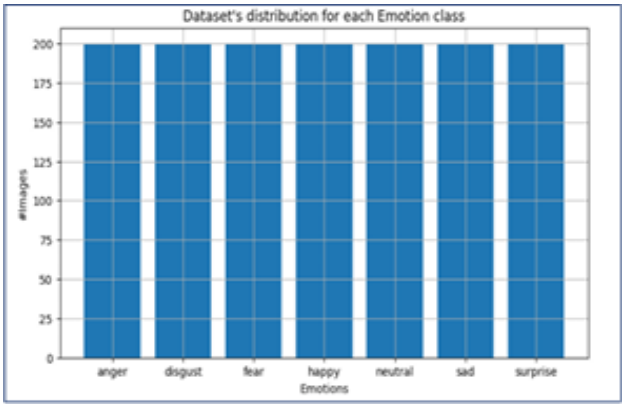


Fig.7. Dataset for proposed system

The dataset is divided into training and testing sets, with 90% of the images allocated for training and 10% for testing. his partitioning ensures that the model can access a substantial portion of the data during training while reserving a separate subset for evaluating its performance on unseen images.

# 5. RESULT ANALYSIS

## 5.1 ACCURACY AND LOSS

In Fig.8, the test accuracy of 0.9214285612106323 represents the proportion of correctly classified instances from the total test dataset. A higher test accuracy indicates that the model effectively distinguishes between different classes. The achieved test accuracy demonstrates that the model correctly predicts the classes of approximately 92.14% of the samples in test dataset, highlighting its strong performance in accurately recognizing facial expressions.

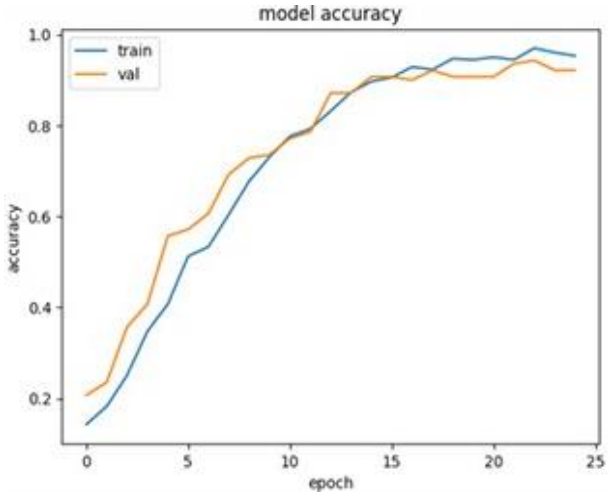


Fig.8. Epoch Vs Accuracy

In Fig.9, The test loss of 0.3318672776222229 indicates the average discrepancy between the predicted probabilities and the actual labels for the test dataset. A lower test loss suggests the model's predictions are closer to the true values. In this case, the obtained test loss value reflects a relatively low-level error, indicating that the model performs well in its classification task.

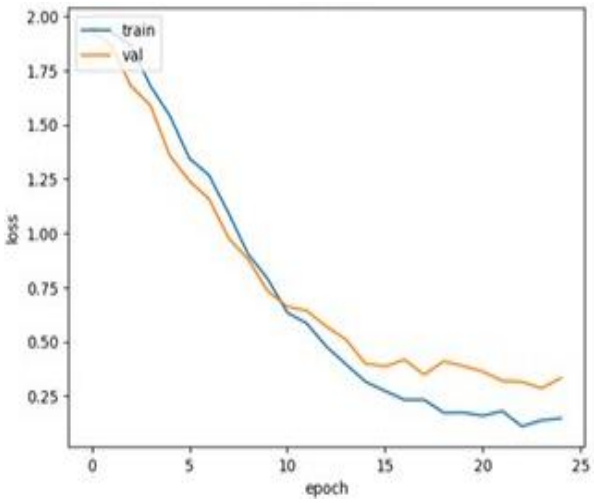


Fig.9. Epoch Vs Loss

The performance of various models with different dataset is compared with proposed system in Table.2. The accuracy of Proposed system is more as compared to existing systems [9] [11] [14] [16].

## 5.2 PERFORMANCE MATRIX

The proposed system achieves an impressive overall accuracy of 92%, indicating its ability to correctly classify the majority of instances in the dataset. This metric is fundamental as it provides a general overview of the model's performance across all classes. Additionally, the F1-score, which combines precision and recall, provides insights into the model's performance on individual classes. The F1 score values range from 0 to 1, with higher values indicating better performance. For this system, the F1 scores range from 0.85 to 1.00 across different classes, with an average macro F1-score of 0.93. This demonstrates the system's ability to balance both precision and recall, achieving high accuracy in classifying each emotion category.

Table.2. Comparison of ML models with proposed model

| References | ML Model | Dataset | Percentage |
|---|---|---|---|
| [9] | VGGFace | IJB-A | 85.8 |
| [9] | Light CNN | IJB-A | 90.5 |
| [9] | CenterLoss | IJB-A | 85.9 |
| [9] | FaceNet | IJB-A | 58.6 |
| [11] | RNN/CNN | FER2013 | 82.4 |
| [14] | DCNN | FER+ | 63.08 |
| [16] | Transfer Learning | FER2013 | 75.8 |
| Proposed system | | FER2013/CK+48 | 92.14 |

Moreover, examining precision and recall individually provides further insights into the system's performance. Precision represents the proportion of true positive predictions out of all positive predictions, while recall represents the proportion of true positive predictions out of all actual positive instances. The precision values range from 0.84 to 1.00, and recall values range from 0.81 to 1.00 across different emotion classes. These metrics indicate that the system achieves high precision and recall rates, effectively identifying and classifying each emotion category with minimal false positives and false negatives. Thus, the proposed system demonstrates robust performance with high accuracy, balanced F1-scores, and high precision and recall rates across multiple emotion classes, indicating its effectiveness in facial expression recognition.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.89 | 0.81 | 0.85 | 21 |
| 1 | 0.89 | 0.89 | 0.89 | 18 |
| 2 | 0.92 | 0.82 | 0.87 | 28 |
| 3 | 1.00 | 1.00 | 1.00 | 20 |
| 4 | 0.84 | 1.00 | 0.91 | 21 |
| 5 | 1.00 | 1.00 | 1.00 | 17 |
| 6 | 0.94 | 1.00 | 0.97 | 15 |
| accuracy | | | 0.92 | 140 |
| macro avg | 0.93 | 0.93 | 0.93 | 140 |
| weighted avg | 0.92 | 0.92 | 0.92 | 140 |

Fig.10 Performance matrix

## 5.3 CONFUSION MATRIX

A confusion matrix is a table that visualizes the performance of a classification model by comparing actual class labels with predicted class labels. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. For the proposed system with an accuracy of 92%, F1-scores, precision, and recall values across different emotion classes, the confusion matrix would provide a detailed breakdown of how the model performs in classifying each emotion category.

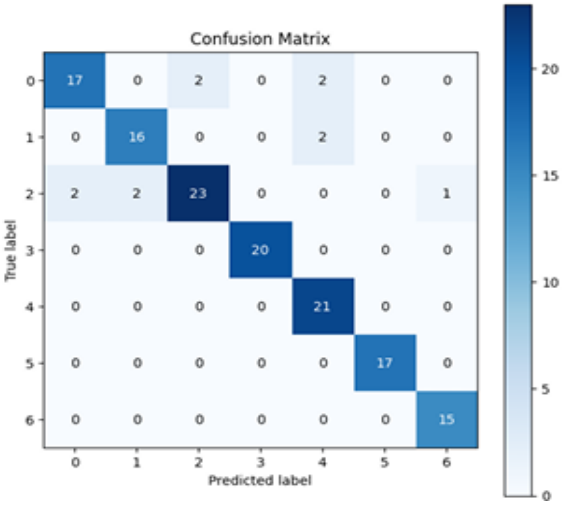


Fig.11. Confusion Matrix

1. ***Diagonal Elements (True Positives)***: These elements represent the number of instances that are correctly classified for each emotion class. A higher value along the diagonal indicates that the model correctly identifies the corresponding emotion category.

2. ***Off-Diagonal Elements (False Positives and False Negatives)***: These elements represent instances that are misclassified. False positives occur when the model incorrectly predicts an instance as belonging to a certain class, while false negatives occur when the model fails to predict an instance as belonging to a certain class.

By analyzing the confusion matrix, we can identify patterns of misclassification and areas where the model may need improvement. For example, Based on the provided class labels (0 to 6) and corresponding support values (21, 18, 28, 20, 21, 17, 15).

## 6. CONCLUSION AND FUTURE WORK

The proposed facial expression recognition system presents a compelling solution characterized by its efficiency, accuracy and balanced performance across seven distinct emotion classes. Notably, the system achieves high F1 scores, indicating a harmonious blend of precision and recall, essential for accurate classification. With a focus on less complexity, the system offers streamlined architecture and implementation, ensuring efficient resource utilization and reduced code complexity. The incorporation of seven emotion classes underscores the system's versatility in capturing a diverse range of human expressions, catering to various real-world applications effectively. Moreover, the system's efficient design facilitates swift inference, making it suitable for real-time deployment scenarios. The attained accuracy and precision further validate the system's effectiveness, affirming its ability to reliably recognize and classify facial expressions with minimal false positives. To enhance emotion detection in the future, hand gestures, body posture, and facial expression recognition may be combined. The accuracy and robustness of emotion recognition systems will be improved by incorporating these extra layers of data, allowing for more meaningful human-computer interactions and customized experiences.

## REFERENCES


[1] S. Pandey, S. Handoo and Yogesh, "Facial Emotion Recognition using Deep Learning", *Proceedings of International Conference on Mobile and Embedded Technology*, pp. 248-252, 2022.

[2] B. Fang, X. Li, G. Han and J. He, "Facial Expression Recognition in Educational Research from the Perspective of Machine Learning: A Systematic Review", *IEEE Access*, Vol. 11, pp. 112060-112074, 2023.

[3] V.T. Dang, H.Q. Do, V.V. Vu and B. Yoon, "Facial Expression Recognition: A Survey and its Applications", *Proceedings of International Conference on Advanced Communication Technology*, pp. 359-367, 2021.

[4] S.Ansari, P. Kulkarni, T. Rajesh and V.R. Gurudas, "Facial Emotion Detection using Deep Learning: A Survey", *Proceedings of International Conference on Contemporary Computing and Communications*, pp. 1-4, 2023.

[5] Z. Zhu and R. Jiao, "Real-Time Facial Expression Recognition Research based on Blazeface Face Detection and Resnet Emotion Classification", *Proceedings of International Conference on Computer Vision, Image and Deep Learning*, pp. 401-408, 2024.

[6] Yue Luo, Jiaxin Wu, Zhuhao Zhang, Huaju Zhao and Zhong Shu, "Design of Facial Expression Recognition Algorithm based on CNN Model", *Proceedings of International Conference on Power, Electronics and Computer Applications*, pp. 580-583, 2023.

[7] Ying Zhang, Di Peng, Yi Wang and Jiaqi Wang, "Research on Facial Expression Recognition Algorithm based on Deep Learning", *Proceedings of International Conference on Mechanical Engineering and Intelligent Manufacturing*, pp. 1010-1013, 2022.

[8] A.V. Savchenko, L.V. Savchenko and I. Makarov, "Classifying Emotions and Engagement in Online Learning based on a Single Facial Expression Recognition Neural Network", *IEEE Transactions on Affective Computing*, Vol. 13, No. 4, pp. 2132-2143, 2022.

[9] P. Xanthopoulos, P.M. Pardalos and T.B. Trafalis, "*Robust Data Mining*", 2012.

[10] S.A. Bargal, E. Barsoum, C.C. Ferrer and C. Zhang, "Emotion Recognition in the Wild from Videos using Images", *Proceedings of International Conference on Multimodal Interaction*, pp. 433-436, 2016.

[11] G. Guo and N. Zhang, "What is the Challenge for Deep Learning in Unconstrained Face Recognition?", *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pp. 436-442, 2018.

[12] H. Mliki, N. Fourati, S. Smaoui and M. Hammami, "Automatic Facial Expression Recognition System", *Proceedings of International Conference on Computer Systems and Applications*, pp. 1- 4, 2013.

[13] Szu-Yin Lin, Yi-Wen Tseng, Chang-Rong Wu, Yun- Ching Kung, Yi-Zhen Chen and Chao-Ming Wu, "A Continuous Facial Expression Recognition Model based on Deep Learning Method", *Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 1-6, 2019.

[14] D.Y. Liliana, M.R. Widyanto and T. Basaruddin, "Geometric Facial Components Feature Extraction for Facial Expression Recognition", *Proceedings of International Conference on Advanced Computer Science and Information Systems*, pp. 391-396, 2018.

[15] Burkert Peter, Trier Felix, Afzal Muhammad Zeshan, Dengel Andreas and Liwicki Marcus, "DeXpression: Deep Convolutional Neural Network for Expression Recognition", *Pattern Recognition Letters*, Vol. 112, pp. 1-9, 2015.

[16] Barsoum Emad, Zhang Cha, Ferrer Cristian and Zhang Zhengyou, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced-Label Distribution", *Proceedings of International Conference on Multimodal Interaction*, pp. 279-283, 2016.

[17] Marriwala, Nikhil and Vandana, "Facial Expression Recognition using Convolutional Neural Network", *Proceedings of International Conference on Artificial Intelligence Trends and Pattern Recognition*, pp. 1-11, 2022.

[18] Khanzada Amil, Bai Charles and Celepcikay Ferhat, "Facial Expression Recognition with Deep Learning", *Computer Vision and Pattern Recognition*, pp. 1-6, 2020.

[19] M.A.H. Akhand and Roy Shuvendu, Siddique Nazmul, M.A.S. Kamal and Shimamura Tetsuya, "Facial Emotion Recognition using Transfer Learning in the Deep CNN", *Electronics*, Vol. 10, pp. 1-13, 2021.

[20] Khaireddin Yousif and Chen Zhuofa, "Facial Emotion Recognition: State of the Art Performance on FER2013", *Computer Vision and Pattern Recognition*, pp. 1-9, 2021.

[21] Minaee Shervin, Minaei Mehdi and Abdolrashidi Amirali, "Deep-Emotion: Facial Expression Recognition using Attentional Convolutional Network", *Sensors*, Vol. 21, pp. 1-8, 2021.

[22] R. Gill and J. Singh, "A Deep Learning Approach for Real Time Facial Emotion Recognition", *Proceedings of International Conference on System Modeling and Advancement in Research Trends*, pp. 497-501, 2021.