IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Human Emotion: The Detection of Human Facial Expressions Using Convolutional Neural Network and Bidirectional Long Short-Term Memory

PILLI LAKSHMI ASHOK

M.Tech, Computer Science and Engineering, UCEK, JNTU Kakinada, Andhra Pradesh, India

Abstract **:** Facial expressions of people are crucial in their daily lives as they determine the emotions of a person. Those are often the most visible and reliable indicators of emotion. Detecting human emotions is easy, but using a computer algorithm to accomplish the same objective is quite difficult.The existing model, the custom Lightweight CNN-based Model (CLCM), is utilized to analyze the facial expressions with lower complexity. However, the architecture might be simplified at the expense of the ability to model the subtle patterns in facial expressions and would be less able to recognizing the complex emotions. It may also lead to a loss in predictive power, particularly when dealing with complex datasets or emotions. The model learns to memorize training data patterns instead of generalizing to new data. This can lead to reduced performance on new or unseen facial expressions, especially if the training dataset is small or not representative of the target application domain. The lightweight design may struggle to capture contextual information or subtle cues, such as background context or nonverbal cues, in facial expressions. The Convolutional neural network (CNN) and Bidirectional long-short term memory (BI-LSTM) model performed better in detecting facial expressions. This ability enables them to capture complex patterns and structures in raw input data, making them highly effective for emotion recognition and classification tasks. The facial emotion recognition research results using the CNN+BI-LSTM method on FER-2013 reached 72% accuracy.

IndexTerms **- Convolutional neural Networks, Bidirectional Long Short-Term Memory, facial expressions, computer vision, Deep learning.**

I. INTRODUCTION

Facial emotions are crucial in human experience. They influence how we think, behave, and communicate. Among the numerous methods of showing our emotions, facial expressions are the most natural and instant. A smile, a frown, or a raised eyebrow can convey volumes without a single word. The rapid advancement of artificial intelligence and machine learning, teaching machines to recognize these expressions, what we call facial emotion recognition (FER), has become a growing interest in both research and industry.

Today, emotion-aware systems are no longer science fiction. From virtual assistants that sense user frustration, in-classroom Artificial Intelligence (AI) tutors that adapt to student engagement, and healthcare platforms that monitor patient mood, FER is becoming a key technology in building emotionally intelligent systems(Zen et al, 2007). (Ko 2018). Despite these advancements, accurately identifying human emotions from facial images remains a tough challenge. This is mainly because human emotions are complex, subtle, and vary widely across individuals, cultures, lighting conditions, and occlusions.

Early attempts at emotion recognition were based on manually designed features such as (C. Shan, S. Gong, and P. W. McOwan, 2009) Local Binary Patterns (LBP), Gabor filters, or facial landmarks, which aimed to capture facial textures and geometries. These handcrafted methods often broke down in real-world applications. The limitations led to the rise of deep learning, particularly CNNs, which transformed the field by learning rich, discriminative features directly from data without manual intervention (Mollahosseini et al, 2017).

CNNs (Bhagat et al. 2023) have shown excellent results in extracting spatial features from facial images, such as how the mouth curves in a smile or how the eyebrows furrow in anger. But facial expressions are not static. They unfold over time. A person's emotion might start as neutral, develop into surprise, and then blend into happiness within a few frames. CNNs alone cannot fully capture this temporal flow of emotions, especially when subtle changes are critical for interpretation.

To address this issue, researchers have turned to Recurrent Neural Networks (RNNs)and their advanced versions like Long Short-Term Memory (LSTM) and Bidirectional LSTM (BI-LSTM) networks. (Hasani and Mahoor 2017). These models are designed to understand sequences; they remember patterns over time, making them ideal for learning how expressions evolve from

moment to moment. BI-LSTM reads input sequences in both directions, providing a more profound framework for analyzing emotion.

Proposed a hybrid deep learning model that combines the strengths of CNN and BI-LSTM (R. Febrian et al. 2023). The CNN part of the model is responsible for learning spatial characteristics of the input facial images. At the same time, the BI-LSTM layer captures temporal relationships and emotional context from those features. This joint model allows for a better understanding of emotions, especially in cases where a static appearance is not enough, for example, to tell the difference between a nervous smile and a genuine smile.

To evaluate the proposed model, we use the FER2013 dataset. (Goodfellow et al, 2015), It is widely used in facial emotion recognition. It contains over 35,000 grayscale images labelled with seven core emotions: Happy, Disgust, Fear, Angry, Neutral, Surprise, and Sad (Scott et al, 2013). This dataset is challenging due to its low resolution and real-world noise.

Figure 1: Facial expression images with labeled emotions from the FER-2013 dataset

Further validate our approach by comparing it with other popular deep learning models, including CNN, CNN+LSTM, CNN+BI-LSTM+ATTENTION, MobileNetV2, DenseNet121, and ResNet50.Through this comparative analysis, we demonstrate how integrating spatial and temporal learning, via a CNN-BI-LSTM hybrid, leads to improved performance, enhanced emotion classification accuracy, and greater robustness in real-world expressions.

II. RELATED WORK

Previous research used handmade features, including Gabor filters (Lakshmi et al. 2021), Histogram of Oriented Gradients (HOG), and Local Binary Patterns (LBP) to extract facial cues. For instance, Shan et al. (2009) investigated LBP-based FER systems and showed respectable performance in controlled settings. The methods performed poorly in difficult conditions of real-life situations with different lights, changing poses and occlusions.

The advent of CNN significantly boosted the performance of FER systems by automating spatial feature extraction. Introduced one of the earliest CNN-based FER systems using the FER2013 dataset, which laid the foundation for subsequent deep learning advancements. Mollahosseini et al. (2017) later proposed a deeper CNN model optimized for affective databases, achieving notable accuracy improvements across diverse datasets such as AffectNet and Extended Cohn-Kanade (CK+).

While CNNs effectively capture spatial features from static images, they often fail to recognize temporal dependencies in emotion sequences. Researchers have introduced hybrid models that combine CNNs with Long Short-Term Memory (LSTM) or Bidirectional LSTM (BI-LSTM) networks to address this. For instance, Wu et al. (2021) proposed a CNN+BI-LSTM framework that integrates spatial and temporal features, reporting improved accuracy on the FER2013 and Japanese Female Facial Expression (JAFFE) datasets.

(Kumar et al. 2023) Integrated attention mechanisms with CNN+BI-LSTM models, permissible tuning of the network to concentrate on emotionally relevant areas of the face. Their work reported higher classification performance, distinguishing similar expressions such as fear and surprise.

The most critical issue in the FER2013 dataset is class imbalance, where emotions like "disgust" are severely underrepresented. To overcome this, researchers such as Lin et al. (2017) introduced Focal Loss, which assigns higher weights to hard-to-classify samples. Similarly, Xu et al. (2022) used data augmentation and synthetic oversampling to boost minority class representation, which improves model fairness.

The recent studies have also explored different deep architectures to push the boundaries of facial emotion recognition performance. (Zhao et al. 2025) DenseNet-based model for FER that promotes feature reuse and alleviates the vanishing gradient problem, yielding better generalization. However, DenseNet architectures are computationally intensive, which makes them less suitable for applications in real-time. Nguyen et al. (2024) explored MobileNetV2-based FER models, achieving real-time inference speeds with modest compromises in accuracy.

III. METHODOLOGY

3.1 System Overview

The suggested Facial Emotion Recognition (FER) system uses a hybrid deep learning model (Tahri, M., Arfaoui, 2024) to identify human emotions from facial expressions reliably. By merging a Convolutional Neural Network (CNN) with a Bidirectional Long Short-Term Memory (BI-LSTM) network, the system combines the extraction of spatial and temporal features, allowing it to learn dynamic emotional cues and complex representations of facial patterns.

3.2 Dataset Description

The FER2013 dataset is a well-known benchmark in facial emotion recognition. It contains a total of 35,887 grayscale images, each image with a resolution of 48x48 pixels. These images represent facial expressions captured under various conditions, each emotion labeled with basic emotions: surprise, disgust, neutral, happiness, sadness, anger, and fear. The images in this data set often include noisy backgrounds, varied lighting, occlusions, and facial orientations, making it a challenging but practical choice for training and evaluating emotion recognition systems. The data set is split into two primary sets: the Training set, comprising 28,709 images, which is used to train the model and learn patterns associated with different facial expressions. Test set containing 7,178 images

split equally between a train and a test set. This evaluates the model's performance and generalization on new data. This simplified split between training and testing ensures effective learning of the data for unbiased performance evaluation. This structured split ensures that models trained on FER2013 are rigorously evaluated, allowing a fair assessment of their generalization performance on unseen data.

Table 1: FER_2013 Dataset basic details

Dataset Name	FER-2013
Number of images	35,887
Number of images in train	28,709
Number of images in test	7,178
Number of emotion classes	7
Colour of images	Grayscale images

3.3 Data Augmentation

Data augmentation makes the proposed model more robust by presenting different variations of the same image. This helps the model recognize facial expressions more accurately, even when the face is slightly rotated, zoomed, or shifted. Applied Random rotation (up to 15 degrees) to simulate tilted or blurred faces. Horizontal flipping, so the model sees both left-facing and right-facing human facial expressions. Zooming in on the faces to make them appear closer or farther away. Cropping the unnecessary background from the image and shifting to help the model handle complex faces, and applying these changes only to training images, not validation data, to avoid biasing the results. This makes training faster, also learn more accurate features, and is more stable by keeping all input values within a small, consistent range.

3.4 Handling Class Imbalance

The FER2013 dataset often suffers from class imbalance. For example, classes such as happy and neutral are overrepresented compared to disgust or fear. More importantly, disgust is the smallest emotion compared to all the other emotions; it leads to a modal focus on emotion that contains more images. Class weights training loss is adjusted using class weights that are computed inversely proportional to the frequency of each class. That is, with fewer samples, so more attention is given to those classes.
These class weights are passed to the training function, where the loss contribution of each sample is scaled accordingly, thus improving performance on minority classes.

Table 2: Class weights for each emotion

Emotions	**Class weight**
Class 0 (Angry)	1.0352
Class 1 (Disgust)	9.3637
Class 2 (Fear)	1.0010
Class 3 (Happy)	0.5703
Class 4 (Sad)	0.8437
Class 5 (Surprise)	1.2809
Class 6 (Neutral)	0.8272

The class weights of each emotion were assigned according to the data imbalance in the FER-2013 dataset (as shown in Table 2). Here, the disgust emotion is lower, so the class weight is assigned to 9.3637. This also helps the model to give importance to the minority classes, along with the primary emotions, which have more images. It will reduce the data imbalance and overfitting.

3.5 Feature Extraction using CNN

Convolutional neural networks (CNNs) extract valuable information from images. The CNN serves as the front-end feature extractor. It learns to automatically recognize patterns like edges, textures, and facial structures to identify human emotions. A grayscale image of a face is fed into the CNN. As the image moves through each convolutional layer for pattern extraction, the CNN uses filters, also known as kernels, to find local patterns. These patterns let the model identify a range of facial expressions, such as smiles, frowns, and raised eyebrows. Pooling layers then decrease the spatial dimensions while emphasizing the salient features. The outcome is a reduced, abstract feature map that captures the most important visual information from the image.

3.5.1 Bidirectional LSTM (Bi-LSTM)

After extracting spatial features using CNN, treat those features as a sequence and feed them into a Bidirectional Long Short-Term Memory (BI-LSTM) layer. This is essential for capturing temporal dependencies or positional relationships among different parts of the face. Traditional LSTM learns from the past to the future. BI-LSTM processes the progression in both past-to-future and

future-to-past directions. This helps the model understand context from both sides, which is especially useful for recognizing subtle expressions that depend on multiple facial regions.

3.6 Data Preprocessing

The system begins with preprocessing the dataset, which consists of facial images. Preprocessing steps include batch normalization, resizing the image, grayscale conversion, and data augmentation techniques such as horizontal flipping, rotation, and random zoom. To address the significant class imbalance present in FER-2013, techniques such as class weighting and other methods (mentioned in Fig. 2) are also applied.

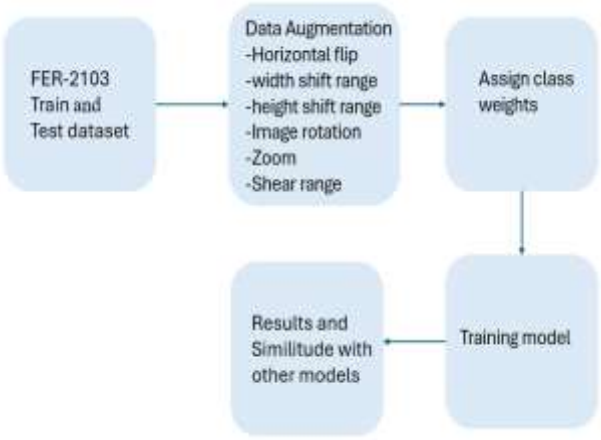

Figure 2 The workflow

3.7 Model Design and Selection

While proficient in learning spatial features, the traditional convolutional architectures often fall short in modeling temporal or contextual patterns, especially when dealing with sequences or subtle variations in expression. A hybrid model combining CNN with a BI-LSTM network is proposed to address this. The proposed design model is shown in Fig.3. The input layer consists of the shape of the input as (48x48x1). The fer-2013 dataset images go through the training process from layer to layer in the CNN+BI-LSTM model to extract the rich facial features.

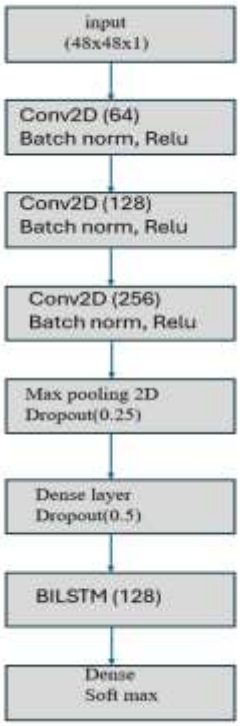

Figure 3: The architecture of the CNN+BI-LSTM model

3.8 Model Candidates

a) CNN

Convolutional Neural Networks (CNNs) play an important role in facial emotion recognition by automatically learning and extracting meaningful features from facial images. Instead of relying on manually crafted features, CNNs directly identify important patterns like smiles, frowns, or eyebrow movements from the image data. This makes them highly effective for recognizing emotions such as happiness, sadness, or anger. Their layered architecture allows them to detect simple and complex facial cues, making CNNs a strong foundation for many FER systems today.

b) LSTM and BI-LSTM

Long Short-Term Memory (LSTM) networks are a special type of recurrent neural network (RNN) designed to learn from sequences and retain critical information over time. In facial emotion recognition, LSTMs are used to understand how different facial features relate to each other across a sequence, such as the connection between eye, Nose, chin movement, and mouth shape.

Bidirectional LSTM (BI-LSTM) is the advanced version of the LSTM model. It extends this idea by processing the data in both forward and backward directions, allowing the model to simultaneously consider past and future context. This is especially helpful in FER, where emotions often depend on the relationships between various facial regions. By combining CNN for spatial feature extraction with BI-LSTM for temporal context, the model becomes more sensitive to subtle and complex emotional expressions.

c)RESNET-50

ResNet-50 (B. Li and D. Lima, 2023)is a powerful deep learning model(K. He, X. Zhang, S. Ren and J. Sun, 2016) known for using residual connections. This makes it possible to train deep networks without vanishing gradient issues. In facial emotion recognition, ResNet-50 is effective at learning detailed and high-level facial features crucial for distinguishing between subtle emotions. Its 50-layer architecture allows it to capture complex patterns from facial images, often leading to higher accuracy than traditional CNNs. However, due to its depth, ResNet-50 requires more computational resources and may not be ideal for real-time or lightweight applications.

d)DENSNET-121

DenseNet-121(Bin Li , 2021) is a deep convolutional neural network with feed-forward connections between each layer and the following layers. This dense connectivity promotes feature reuse and improves gradient flow, resulting in a highly efficient and accurate model with fewer parameters. DenseNet-121, in facial emotion recognition, can capture fine-grained facial details by leveraging features learned at multiple levels. The DenseNet-121 requires more memory and computational power.

e) MOBILENETV2

A lightweight deep learning model called MobileNetV2 was created to operate effectively on portable and low-resource devices. Inverted residual blocks and depth-wise separable convolutions are used to lower computational costs without appreciably compromising accuracy. MobileNetV2 is helpful for real-time applications in facial expression identification when speed and resource efficiency are crucial. However, because of its compact structure, it could not capture complex face traits better than deeper models like DenseNet-121 or ResNet-50, which could lead to slightly lower accuracy in some situations.

IV. RESULTS AND DISCUSSION

Seven models were used, trained, and evaluated using a stratified 80-20 train-test split. The accuracy of each model (as listed in Table 4) determines how well those models work on human face expressions.

Table 3: Model Accuracy Comparison

Model	Accuracy(%)
CNN	56%
CNN+LSTM	62%
CNN+BI-LSTM	72%
CNN+BI-LSTM+ATTENTION	66%
RESNET-50	48%
DENSENET-121	58%
MOBILENETV2	54%

This table represents a comparative analysis of model accuracies achieved by seven different deep learning algorithms. The lightweight models are easy to use and offer faster performance, but the accuracy and key features may be missed.

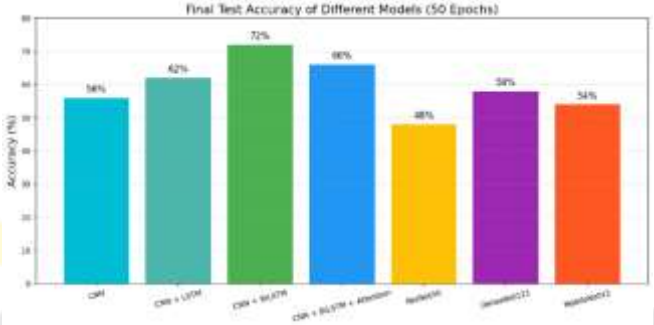

Figure 4 Model accuracy Comparison

Explored and compared the performance of multiple deep learning models for facial emotion recognition to determine which model could best capture the subtle variations in human expressions (shown in Fig.4). The basic CNN model achieved an accuracy of 56%.It extracts local spatial features but shows limitations in modeling complex facial cues.

The CNN + LSTM model attained a higher accuracy of 62%, benefiting from its ability to model sequential spatial patterns, which improved its ability to decipher emotional dynamics across facial regions. The CNN+LSTM model performed slightly better than the CNN model. The proposed CNN + BI-LSTM model outperformed all others, achieving an accuracy of 72%, demonstrating a solid ability to extract spatial features (through convolutional layers) and temporal dependencies (via BI-LSTM layers) from facial images.

For further experimentation with an enhanced version of the attention mechanism in the CNN + BI-LSTM architecture. This variant achieved an accuracy of 66%. While attention mechanisms typically improve focus on relevant features, the added complexity results in the model's performance being lower than that of the proposed model, resulting in a slightly lower performance of the attention model compared to the CNN+BI-LSTM.

The ResNet-50, known for its deep architecture and residual connections, achieved an accuracy of 48%. Despite its ability to learn deeper representations, it lacked temporal modeling, which limited its effectiveness in capturing emotional dynamics. DenseNet-

121, which has dense connectivity that encourages feature reuse, performed better at 58%. This model's strong gradient flow and effective feature propagation made stable performance even with fewer parameters possible.

Lastly, MobileNetV2, optimized for mobile and edge devices with a compact design, reached an accuracy of 54%. While it excels in efficiency, the lightweight architecture may limit its capacity to fully capture human facial expressions.

In summary, the CNN + BI-LSTM model achieved the best results. This comparative study clarifies how various architectural decisions affect facial emotion recognition in terms of depth, connectivity, and efficiency. These insights are especially helpful for selecting appropriate models in real-world applications where complexity and computational cost must be carefully balanced.

V. CONCLUSION

This study focused on comparing various deep learning models for recognizing emotions from human faces, particularly emphasizing the proposed CNN+BI-LSTM model. After extensive testing, we discovered that combining convolutional layers with bidirectional LSTMs enabled the model to capture both spatial and temporal features of facial expressions, resulting in an accuracy of 72%. While incorporating attention mechanisms did not enhance our results, comparing the proposed model with others provided valuable insights. Despite their robust design, models such as ResNet-50 and DenseNet-121 still fell short in certain aspects.

The proposed model seems promising in the healthcare domain, where there is a need to capture patients' emotions through real-time facial expressions. Future work could concentrate on enhancing the model, utilizing larger datasets, 3D facial expressions, and enhancing accuracy even further.

VI. REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," in *Proc. 9th Int. Conf. Multimodal Interfaces (ICMI)*, Nagoya, Japan, pp. 126–133, 2007.

[2] B. C. Ko, "A brief review of facial emotion recognition based on visual information," Sensors, vol. 18, no. 2, p. 401, 2018.

[3] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.

[4] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009. doi: 10.1016/j.imavis.2008.08.005.

[5] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affective Comput., vol. 10, no. 1, pp. 18–31, 2017.

[6] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," in Proc. 12th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2017), pp. 790–795, 2017.

[7] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015.

[8] N. Scott, N. Kasabov, and G. Indiveri, "NeuCube neuromorphic framework for spatio-temporal brain data and its Python implementation," in Proc. Int. Conf. Neural Information Processing, pp. 78–84, 2013.

[9] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," IEEE Trans. Affective Comput., vol. 6, no. 1, pp. 1–12, 2015.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Computer Vision (ICCV), pp. 2980–2988, 2017

[11] K. Xu, L. Zhang, Y. Chen, and M. Wang, "Addressing data imbalance in emotion recognition using synthetic oversampling," Journal of Real-Time Image Processing, vol. 19, no. 4, pp. 987–1002, 2022.

[12] H. Zhao, J. Liu, R. Chen, and L. Wang, "Facial emotion recognition using DenseNet and transfer learning," Journal of Visual Communication and Image Representation, vol. 88, p. 103456, 2025.

[13] A. S. Mohammad, T. G. Jarullah, M. T. Al-Kaltakchi, J. Alshehabi Al-Ani, and S. Dey, "IoT-MFaceNet: Internet-of-Things-based face recognition using MobileNetV2 and FaceNet deep-learning implementations on a Raspberry Pi-400," Journal of Low Power Electronics and Applications, vol. 14, no. 3, p. 46, 2024.

[14] L. Zahara, P. Musa, E. P. Wibowo, I. Karim, and S. B. Musa, "The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi," in 2020 5th Int. Conf. Informatics and Computing (ICIC), pp. 1–9, 2020.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, pp. 4510–4520, 2018.

[16] B. Li and D. Lima, "Facial expression recognition via ResNet-50," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 57–64, 2021. doi: 10.1016/j.ijcce.2021.02.002.

[17] Bin Li (2021). "Facial expression recognition by DenseNet-121", in Multi-Chaos, Fractal and Multi-Fractional Artificial Intelligence of Different Complex Systems, pp. 263–276, 2021.

[18] D. Lakshmi and R. Ponnusamy, "Facial emotion recognition using modified HOG and LBP features with deep stacked autoencoders," *Microprocessors and Microsystems*, vol. 82, p. 103834, 2021.

[19] S. Bhagat, "Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN)," *Procedia Computer Science*, vol. 235, pp. 2079–2089, 2024, International Conference on Machine Learning and Data Engineering (ICMLDE 2023).

[20] R. Febrian, B. M. Halim, M. Christina, D. Ramdhan, and A. Chowanda, "Facial expression recognition using bidirectional LSTM - CNN," *Procedia Computer Science*, vol. 216, pp. 39–47, 2023, doi: 10.1016/j.procs.2022.12.109.

[21] Tahri, M., Arfaoui, N. (2024). E-Learning Facial Emotion Recognition Using Deep Learning Models. In: Abraham, A., Bajaj, A., Hanne, T., Hong, TP. (eds) Intelligent Systems Design and Applications. ISDA 2023. Lecture Notes in Networks and Systems, vol 1047. Springer, Cham. https://doi.org/10.1007/978-3-031-64836-6_22