


Real-Time Facial Expression Recognition: Advances, Challenges, and Future Directions

Christine Dewi ^{*}, Lanyta Setyani Gunawan[†] and Sastra Gangga Hastoko[‡]

*Department of Information Technology
Satya Wacana Christian University Salatiga, Indonesia*

**christine.dewi@uksw.edu*

†lanytagunawan5@gmail.com

‡hastokosastra@gmail.com

Henoch Juli Christanto [§]

*Department of Information System
Atma Jaya Catholic University of Indonesia Jakarta 12930, Indonesia
henoch.christanto@atmajaya.ac.id*

Received 19 August 2023

Revised 13 November 2023

Accepted 15 November 2023

Published 22 December 2023

Facial emotion recognition (FER) is the technology or process of identifying and interpreting human emotions based on the analysis of facial expressions. It involves using computer algorithms and machine learning techniques to detect and classify emotional states from images or videos of human faces. Further, FER plays a vital role in recognizing and understanding human emotions to better interpret someone's feelings, intentions, and attitudes. In the present time, it is widely used in various fields such as healthcare, human–computer interaction, law enforcement, security, and beyond. FER has a wide range of practical applications across various industries including Emotion Monitoring, Adaptive Learning, and Virtual Assistants. This paper presents a comparative analysis of FER algorithms, focusing on deep learning approaches. The performance of different datasets, including FER2013, JAFFE, AffectNet, and Cohn–Kanade, is evaluated using convolutional neural networks (CNNs), deep face, attentional convolutional networks (ACNs), and deep belief networks (DBNs). Among the tested algorithms, DBNs outperformed other algorithms, reaching the highest accuracy of 98.82%. These results emphasize the effectiveness of deep learning techniques, particularly DBNs, in FER. Additionally, outlining the advantages and disadvantages of current research on facial emotion identification might direct future research efforts in the direction of the most profitable directions.

Keywords: Face emotion recognition; comparative analysis; deep learning; deep belief network.

[§]Corresponding author.

This is an Open Access article published by World Scientific Publishing Company. It is distributed under the terms of the Creative Commons Attribution 4.0 (CC BY) License which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Facial expressions play a crucial role in communication, serving as a non-verbal means of showing emotions.¹ Accurate emotion recognition helps people maintain important relationships necessary for survival, working together, and reproduction.² The human face consists of numerous dynamics of emotional states including happiness, sadness, anger, fear, surprise, and disgust. These facial expressions are characterized by changes in facial muscle movements, shape deformations, and overall appearance. FER is a field of study where experts in computer vision, affective computing, human-computer interaction, and human behavior focus on forecasting emotions by analyzing facial expressions within images or videos.³ In recent years, significant progress has been made in developing techniques for facial emotion recognition (FER). Traditional approaches have relied on handcrafted features and machine learning algorithms to extract relevant information from facial images and classify emotions. However, with the emergence of deep learning techniques, particularly convolutional neural networks (CNNs), there has been a paradigm shift in the field. Deep learning models have shown remarkable success in automatically learning discriminative features directly from raw facial images, leading to improved accuracy in emotion recognition tasks.⁴

Despite the advancements made, several challenges persist in FER. One of the primary challenges is dealing with the inherent ambiguity and subjectivity of facial expressions, as the same facial configuration can be associated with multiple emotions. Additionally, individual differences in expressing and perceiving emotions present a significant hurdle, as cultural and personal factors influence the interpretation of facial expressions. Moreover, variations in lighting conditions, head poses, occlusions, and facial attributes further complicate the recognition process, making it challenging to develop robust and generalizable systems. Furthermore, research in FER can be categorized based on the foundation of the emotional model being either discrete emotional states or continuous dimensions such as valence and arousal.⁵ In the former, there's a shared understanding among researchers regarding emotions as distinct states, although there are variations in defining these fundamental emotions. For instance, Ref. 6 identified happiness, anger, disgust, sadness, and fear/surprise as the five primary emotions. In Ref. 7, delineated play, panic, fear, rage, seeking, lust, and care as fundamental emotions. On the other hand, the continuous dimension model characterizes emotions through two or three dimensions, encompassing valence or pleasantness as one dimension and arousal or activation as the other dimension.⁸

FER has a wide range of practical applications across various industries. Here are some specific examples of how this technology can be applied to Healthcare and Mental Health: (1) Emotion Monitoring: FER can help in monitoring patients' emotional states in real-time, enabling healthcare professionals to provide timely interventions and support for those struggling with mental health issues. (2) Autism Therapy: Emotion recognition technology can assist therapists in working with

individuals on the autism spectrum by helping them recognize and understand emotions through interactive exercises. (3) Depression and Anxiety Detection: By analyzing changes in facial expressions, this technology can aid in detecting early signs of depression or anxiety in patients. Examples of face emotion application in education are as follows: (1) Adaptive Learning: Educational platforms can use FER to adapt their content and teaching methods based on students' engagement levels and emotional states, thereby enhancing the learning experience. (2) Assessment and Feedback: Emotion recognition can be used to analyze students' reactions to learning materials, allowing educators to tailor their teaching strategies and provide targeted feedback. Applications of FER include (1) Human-Computer Interaction: Emotion recognition can be used to create more intuitive and responsive human-computer interfaces. For example, an application could adjust its behavior based on the detected emotions of the user. (2) Market Research: Businesses can use emotion recognition to gauge consumer reactions to advertisements, products, or services by analyzing their facial expressions. (3) Healthcare: Emotion recognition technology can be used to monitor and assist individuals with conditions such as autism or depression, helping caregivers and therapists better understand their emotional states. (4) Security: FER can be integrated into security systems to enhance access control and identify suspicious behaviors. (5) Education: Educational software can use emotion recognition to adapt learning materials based on students' emotional states, optimizing the learning experience.

This paper aims to provide a comprehensive review of the techniques and challenges in FER. We will delve into the underlying theories of emotion, explore the physiological basis of facial expressions, and discuss the importance of feature extraction and selection. Additionally, we will survey the existing methodologies, ranging from traditional machine learning algorithms to deep learning models, and analyze their strengths and limitations. By addressing the challenges and discussing the techniques, the main goal of this paper is to serve as a valuable resource for researchers and practitioners in the field of FER. The insights gained from this study will contribute to the development of more accurate and robust systems capable of understanding and responding to human emotions effectively. We collected the journal paper from the Mendeley database and searched around 70 papers with the keyword FER from 2020 to 2023. Next, we eliminated our paper to become 40 for further analysis.

The next section discusses some relevant research in the field of FER using deep learning approaches. Section 3 discusses the methodology used in FER. More specifically, information is given about the datasets used, the algorithms used, as well as the experimental steps performed. The results of the experiments conducted in FER using the Convolutional Neural Network (CNN) algorithm are discussed in Sec. 4. From the experimental results and analysis conducted with the various algorithms, Sec. 5 describes the implications and limitations of the findings. Finally, Sec. 6 shows the conclusions and avenues for future research.

2. Related Work

This section covers several significant studies in the field of FER using deep learning approaches. Research by Alessandro Chiurco *et al.*⁹ provides a broad overview of deep learning approaches applied to FER, covering the advancements, challenges, and trends. The DeepFace algorithm has demonstrated its effectiveness in the real-time detection of facial expressions, especially in large industries, by utilizing datasets such as CK+ and FER2013. Another notable work by Parashakthi *et al.*¹⁰ present Adaboost Algorithm, a machine learning algorithm that combines multiple weaker classifiers such as the SVM classifier to create a stronger classifier. This work reveals a new music recommendation system based on FER that provides music recommendations based on a person's emotional expressions including happiness, sadness, anger, fear, disgust, and neutrality. Moreover, prominent research by Samira Ebrahimi *et al.*¹¹ provides a hybrid CNN-RNN architecture for facial expression analysis. The author of Ref. 12 focuses on identifying medical face masks from images taken in real-life situations. Yolo V4 CSP SPP model scheme is performed on masked faces within the FMD and MMD datasets. Among various commonly used techniques, Yolo V4 stands out as a more efficient and successful method. In Ref. 13, their work introduced a framework utilizing a multi-task cascaded CNN. This model adopts a three-stage cascaded architecture to enhance the effectiveness of face detection. The performance of the FER is successful within the controlled image datasets. However, it did not perform well in images with different variations. Instant FER through the utilization of the Local Binary Point (LBP) algorithm is conducted in Ref. 14. They extracted features from captured video using the LBP method and fed them into K-Nearest Neighbour (KNN) regression along with dimension labels. The achieved accuracy for the LBP algorithm stood at 51.28%.

Additionally, Gabrielle Simcock's research investigates the application of CNNs with some data processing techniques such as the Kessler Psychological Distress Scale (K-12) and Somatic and Psychological Health Report (SPHERE-12) to enhance performance in handling various expressions.¹⁵ By detecting these diverse expressions, potential mental health problems can be identified during early adolescence, allowing for early prevention and solutions. Furthermore, research by Lourn Bourke *et al.*¹⁶ offers comprehensive reviews of deep learning approaches, covering deep neural networks along with their evaluation metrics utilizing survey tests and correlation analysis. Using the FACES datasets,¹⁶ it was observed that children have a higher tendency to interpret negative emotions when individuals are wearing face masks, although overall emotion recognition accuracy is yet low. Lastly, a distinguished work by Guiping Yu *et al.*¹⁷ explores the use of a new feature descriptor called "Oriented gradient histogram" including the HOG-3D (Histogram of Oriented Gradients) in combination with LSTM networks to characterize different facial expressions. The method has demonstrated superior effectiveness compared to recent approaches in addressing facial expression recognition challenges. Collectively, these studies add valuable insights into the advancements and techniques

engaged in deep learning for FER. In the study by Ref. 18, an investigation was conducted on three distinct models: Network-in-network (NiN-CNN), convolutional network (All-CNN), and very deep convolutional network (VD-CNN). They extracted beneficial features from these models and introduced an enhanced model. In this enhanced model, a multi-layer perceptron substituted the linear convolutional layer, several max-pooling layers were replaced by an additional convolutional layer, and the network's depth was modified with small (3×3) convolutional filters. The performance of engagement detection was evaluated, and the proposed model's effectiveness was compared against the performance of the three baseline models using the DAiSEE dataset¹⁸ in electronic environments. In Ref. 19, a new method for facial expression recognition using a multi-kernel convolution block is proposed. This approach utilized three separable convolutions to extract features, capturing various kernel sizes, and details from facial expressions. The resulting lightweight facial expression network achieved a 73.3% accuracy on FER-2013 and CK+ datasets according to experimental results.

2.1. Appearance-based features

This approach extracts visual descriptors from facial images, such as texture, color, and statistical features.¹⁷ A new feature descriptor called HOG-3D is used in characterizing different facial expressions from 3D data, such as videos or volumetric data.²⁰ The utilization of HOG-3D allows the recognition of six basic emotions (happiness, sadness, anger, disgust, fear, and surprise) in facial expression recognition.²¹ This method involves a few steps to compute histograms of oriented gradients. Initially, voxelization is performed to divide 3D data into small volumetric cells.²⁰ Then, it calculates gradients in the x , y , and t dimensions using the derivative filter to capture spatial variations. The gradients are then categorized into orientation bins, creating a 3D histogram. Lastly, all the histograms are concatenated to create a high-dimensional feature vector representing the 3D data, capturing the spatial appearance.

2.2. Deep learning approaches

2.2.1. Convolutional neural networks (CNNs)

One of the most highly successful FER is through the utilization of CNNs. This algorithm is a deep-learning approach that enables an increasing level of precision in recognition.²² One role of CNN is to reduce images into a form that is easier to process without losing features that are critical for good prediction.²³ CNNs can be employed to recognize and classify emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutrality. It is created upon multiple different building blocks or layers, such as convolution layers, pooling layers, and fully connected layers. These layers perform different transformation functions. Convolutional is the first layer that enables feature extractions from images.²² Operations like edge detection,

blur, and sharpening on an image can be performed by using different convolutional filters. The second layer is the pooling layer, which can reduce parameters in images that are excessively large. Moreover, spatial pooling, also called down sampling, is a technique that reduces the dimension of a certain image yet preserves crucial information. The last convolutional layers further deepen their understanding of object features.²³ Popular CNN architectures are VGGNet, ResNet, and Inception.⁹ Formula (1) contains the definition of the convolution operation between two functions f and g .

$$(f * g)(t) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau. \quad (1)$$

The variables and symbols used are as follows:

- $(f * g)(t)$: This represents the convolution of functions f and g at the point t .
- $f(\tau)$: This is one of the functions being convolved.
- $g(t - \tau)$: This is the second function being convolved, with a shift of τ .
- t : The variable at the point where the convolution result is evaluated.
- τ : The integration variable in the integral, representing a potential shift in the function g .

Formula (1) illustrates how the influence of the function g on f changes as g is shifted across f . This operation provides insights into how these functions interact and affect each other.

2.2.2. DeepFace network

DeepFace Networks is a deep learning approach that utilizes deep CNNs to extract features from facial images.⁹ This method consists of three layers, convolutional layers, pooling layers, and fully connected layers. The first layer is used to extract hierarchical features from the images and capture patterns and structures at different layers of abstraction. Then pooling layers are performed to reduce spatial dimensions while maintaining the vital information. Finally, the fully connected layers classify the features that are extracted and create predictions by mapping the high-dimensional representations to identify vectors. These steps allow the DeepFace network to perform accurate face image recognition. Deep Face contains four modules which are 2D alignment, 3D alignment, formalization, and neural network. The relation between 2D and 3D is established using the relation in Formula (2).

$$x2d = X3d\underline{P}. \quad (2)$$

The elaboration of Formula (2) is presented as follows:

- $x2d$: This signifies the fiducial points extracted from the 2D image.
- $x3d$: This signifies the fiducial points extracted from the 3D image.
- \underline{P} : This represents the reduced loss resulting from the conversion between a 2D image and a 3D image.

2.2.3. Visual geometry group

The Visual Geometry Group (VGG) is a deep learning approach that is commonly used to perform different computer vision tasks, including FER. The network is made from multiple layers including convolutional layers, max-pooling layers, and fully connected layers.²⁴ VGG utilizes a stack of smaller convolutional filters with a small size and padding of 1. This serves as a detailed capture of spatial information and allows learning of the complex features in images. VGG has different types depending on its depth, such as VGG16, and VGG19, which have 16 and 19 layers, respectively. The image classification performance is proven to be powerful compared to shallower networks.

2.2.4. Xception network

Xception network is a deep learning approach known as one of the most prominent approaches in artificial vision. It allows the extraction of features with lesser parameters as well as enables richer representation. The Xception network uses separable convolution to enhance its performance. It mainly divides the process of convolutions into two steps which are depth-wise convolutions and pointwise convolutions. Depth-wise convolutions perform one filter to each input. It then captures every spatial information from each input. On the other hand, pointwise convolutions perform 1×1 convolution, allowing the correlation of cross-channels to be captured.⁹

2.2.5. Improved RM-Xception

The RM-Xception emotion recognition algorithm initially selects the activation function called the RELU in neural networks or commonly known as the modified linear unit. It doesn't require exponential operations and requires small computations shown in Formula (3).

$$f(x) = \max(0, x). \quad (3)$$

The explanation of the Formula (3) is as follows:

- When x is a positive number or zero, the function returns $f(x) = x$, meaning the output is the same as the input.
- When x is a negative number, the function returns $f(x) = 0$, meaning the output becomes zero. This reflects the linear behavior of negative inputs in the function.

The overall structure of this network is mainly built of three parts: Entry flow, Middle Flow, and Exit Flow. Initially, the Entry flow performs 3×3 convolution on the face images that are imputed and normalize the batch after the activation from the RELU function is done. The normalization and RELU functions reduce the diversity of data and enhance the nonlinear expression capability. In the middle flow section, the convolved data is passed through four depth-separable convolution modules, each having a direct residual connection. Within each module, the convolutions are performed three times, followed by activation, batch normalization,

and then a 1×1 convolution with a direct residual connection. In the Exit flow section, the output of the final module, which undergoes a 1×1 convolution and a global mean pooling operation, is sent forward. This output is then passed to a SoftMax classifier to classify and determine seven emotions: Anger, disgust, fear, happiness, sadness, surprise, and neutral.⁹

2.2.6. Hybrid attention cascade network

The Hybrid Attention Cascade Network is commonly used to recognize facial expressions in videos.²⁵ The videos are processed in different parts, also called the *k parts*. Each of the parts is then extracted into frame sequences of F .

$$F = n.f1, f2, \dots, fko, \quad (k \in N). \quad (4)$$

From Formula (4), the fk is selected randomly from the divided parts. While N is the total number of frames that are utilized. The network is primarily divided into three modules which are the spatial feature extraction module, the temporal feature extraction module, and the hybrid attention module. Initially, the spatial feature extraction module performs a selection process of the residual network to extract spatial features and input the extracted features to the module. To form hybrid attention features, they need to be weighted to the spatial feature vector. The weighted hybrid attention feature is then selected by the temporal feature extraction module. A fully connected layer connects the temporal feature extraction module and the SoftMax layer is imputed to obtain the classification results.

2.2.7. Emotion recognition using meta-learning across occlusion, pose, and illumination

Emotion Recognition using Meta-Learning across Occlusion, Pose, and Illumination (ERMOPI) is a network designed for FER. It combines embedded networks and classifiers. Feature embedding refers to the process of converting features from their original space into a new, smaller dimension space that facilitates efficient learning. In the context of ERMOPI, the feature embedding network plays a crucial role in extracting complex features and low-level features from facial images.²⁶ The extracted images are then formed into residual blocks of 3×3 convolutions. According to Ref. 19, a residual block is defined as given in Formula (5).

$$y = F(x_i\{w_i\}) + w_s x, \quad (5)$$

where x represents the input within the residual blocks, y corresponds to the output within the residual blocks $F(x_i\{w_i\})$, presented the residual mapping that's learned and $w_s x$ depicts the shortcut connections between the blocks.

2.2.8. Deep belief network

Deep Belief Network (DBN) is an artificial neural network (ANN) that is made of multiple layers of nodes that are interconnected. It is utilized to learn data that is

input without any labels predefined. DBN is mainly formed of two layers, which are the visible layers and the hidden layers. Although these layers are fully connected, it performs specific functions. The visible layers represent the data that is input while the hidden layer captures abstract data. The network is then trained using the RBM technique.²⁴ Formula (6) was performed to transform the layers.

$$P(S_i = 1) = \frac{1}{1 + \exp(-b_i - \sum_j S_j W_{ij})}. \quad (6)$$

Explanation:

- $P(S_i = 1)$ represents the probability that the binary variable S_i takes the value of 1.
- b_i is the bias associated with node i in the network.
- S_j denotes the state of the binary hidden unit j in the network.
- W_{ij} represents the weight between the visible unit i and the hidden unit j .

The process of training is continued and the parameters are using Formula (7).

$$\text{update} \left(w_{ij} + \frac{n}{2} (\text{positive}(E_{ij}) - \text{negative}(E_{ij})) \right). \quad (7)$$

The elaboration of the Formula (7) is presented as follows:

- $\text{positive}(E_{ij})$ — Positive statistics of edge $E_{ij} = p(h_j = 1|v)$
- $\text{negative}(E_{ij})$ — Positive statistics of edge $E_{ij} = p(v_j = 1|h)$.

2.2.9. Attentional convolutional network

The attentional convolutional network (ACN) is a neural network that is a combination of convolutional layers and attention mechanisms. It will select only relevant parts from the data input and then it will go through the convolutional layers for further processing. This network allows adaptiveness to determine which task is most important. It will assign attention weights to different spatial channels in the data.²⁷ This model is then optimized using the loss function as described in Formula (8).

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{classifier}} + \lambda \|\omega(fc)\|_2^2. \quad (8)$$

From Formula (8) the explanations are as follows:

- $\mathcal{L}_{\text{overall}}$: This represents the total loss function or objective function that is being optimized. The function comprises two components: $\mathcal{L}_{\text{classifier}}$ and $\lambda \|\omega(fc)\|_2^2$.
- $\mathcal{L}_{\text{classifier}}$: This component represents the loss function associated with classification or the main task being performed. It's a part of the loss function that focuses on measuring how well the model classifies the data.
- λ : This is a parameter that controls the magnitude of the penalty component $\|\omega(fc)\|_2^2$. The parameter λ is often referred to as a control factor or regularization factor, and it can be adjusted to determine how much the penalty will influence the optimization outcome.

- $\|\omega(fc)\|_2^2$: This represents the L2 Euclidean norm of the weight vector (ω) associated with a fully connected layer (fc). The L2 norm is used here to measure the magnitude and complexity of the weight vector. Adding the L2 norm in the loss function aims to prevent overfitting and encourages the weight vector to approach zero, which can contribute to a more generalizable model.

2.2.10. Spatial Pyramid Zernike Moments

Spatial Pyramid Zernike Moments (SPZM) is a feature descriptor to analyze images and recognize patterns.¹⁹ It is a combination of two strength techniques which are Zernike moments and spatial pyramid representation. These two combined enables capturing local and global information from images that are input. Specifically, Zernike is formed of complex orthogonal moments that can indicate the texture of objects. The images are projected to a set of orthogonal base functions that are defined in a circular region. The properties include rotation invariance, scale invariance and robustness to noise. Meanwhile, the spatial pyramid representation enables dividing images hierarchically based on their sub-regions. Fine-grained details and coarse-grained global structures are both captured.²⁸ The facial images $I(x, y)$ are divided into set of grids at levels 0–2. Each level has two sub regions or cells $C(x, y) = I(x, y)_k^l$. To extract features, computation is performed on the images for each cell $C(x, y)$ ZM is calculated for every cell across various levels of the image pyramid, and these results are combined to create the SPZM feature, as illustrated in Fig. 1.

$$[Anm_k^0], \quad k = 1, \quad [Anm_k^1]k = 1 : 4, \quad [Anm_k^2]k = 1 : 16. \quad (9)$$

The SPZM feature vector is shown in Formula (9). The explanation of the formula is as follows: n represents the order of the Zernike polynomial and m represents the repetition factor such that $|m| < n$ and $n - |m|$. These terms represent specific quantities computed using Zernike moments or other techniques within the SPZM framework. Overall, the Spatial Pyramid Zernike Moments feature descriptor provides a way to capture both local and global information from images by combining

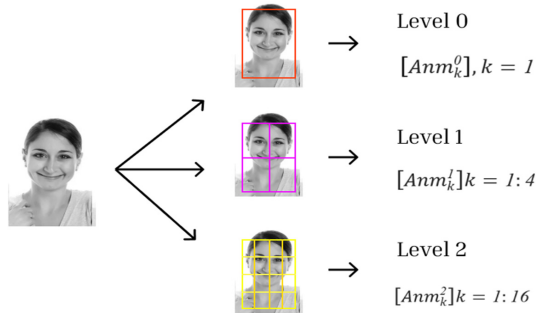


Fig. 1. SPZM features sample images.

Zernike moments and spatial pyramid representation, making it effective for various image analysis and pattern recognition tasks.

2.2.11. *You only look once*

You Only Look Once (YOLO), a popular real-time object recognition system, utilizes a deep learning framework and has evolved through various versions, including YOLOv1, v2, v3, v4, and v5 series. In Ref. 29, the author highlighted that YOLOv3 has three detectors, resulting in more precise detection. Conversely, YOLOv4, with its spatial pyramid pooling, enhances feature extraction for more accurate object classification. The latest iteration, YOLOv5, introduces several versions, namely YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. YOLOv5 employs the anchor strategy, focus structure, and cross-state partial connection (CSP). Notably, YOLOv5 offers an advantage in terms of run speed, particularly in real-time detection scenarios within clinical workloads.

2.2.12. *Long short-term memory recurrent neural networks*

This study discusses the incorporation of Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) as a crucial component of the proposed emotion recognition model. LSTM-RNNs are widely recognized for their ability to capture temporal dependencies in data, making them suitable for processing electroencephalography (EEG) signals in the context of emotion recognition.³⁰ This research demonstrates the significance of LSTM-RNNs in handling EEG data for emotion recognition, especially when combined with CNNs to form a fused model for improved accuracy.

LSTM is used to extract temporal features from EEG signals for emotion recognition. It comprises of five key components: The memory cell ($\hat{C} < t >$), a candidate value ($\check{C} < t >$), and three gates — the update gate (\hat{g}_u), forget gate (\hat{g}_f), and output gate (\hat{g}_o). The memory cell variable is computed at each time step, and the candidate value is used for memory cell replacement. During training, the memory cell retains a sequence of values, and the weight matrix and bias for the gates are updated based on values between 0 and 1. The forget gate (\hat{g}_f) determines which information to discard, the update gate (\hat{g}_u) decides whether to replace the memory cell value with the candidate value, and the output gate (\hat{g}_o) generates the activation function for the current time step. The LSTM behavior is defined by a set of equations. The extraction of Temporal Feature Vectors (TFV) using LSTM involves capturing the entire sequence of time-step features.

A modified RNN, known as LSTM-RNN, is utilized to compute temporal features from EEG input signals. This model comprises of two stacked RNN layers, each with 'S' units. The output of the first LSTM layer is used as an input sequence to the second layer, enabling the transfer of information from previous time steps. The goal is to identify and classify emotions from EEG data, and a pseudocode for training a parallel CNN-LSTM cross-subject EEG-based emotion detection model is also presented in the study.

2.2.13. Deep reinforcement learning

Utilizing EEG signals and the Deep Reinforcement Learning approach for emotion recognition, EEG signals represent the electrical brain activity, offering insights into a person’s emotional state. In this context, Deep Reinforcement Learning is employed to process EEG data and identify patterns associated with various emotional conditions. This approach creates an innovative way to combine knowledge of brain activity with advanced machine learning technology, aiding in a deeper understanding of emotional expression through EEG signals.³¹ Moreover, Deep Q Network (DQN) is a reinforcement learning (RL) that utilizes the neural network architecture. RL allows Machine Learning (ML) an agent to learn and maximizes performance through trial and error.³²

3. Methodology

3.1. Datasets

Three different datasets have been utilized to test which contributed to better outcomes. FER dataset was created in 2013, utilizing APIs from the Google search engine and datasets from the Kaggle Challenge.⁹ The test was run during a challenge. The images in the FER2013 dataset are 48 × 48 in size and black and white image. The FER2013 dataset contains images with different perspectives, lighting, and scales. Figure 2 shows some sample images from the FER2013 dataset, and Table 1 gives a description of the dataset.

CK+: The Cohn–Kanade (CK+) dataset is an emotion recognition dataset that remains publicly used. The dataset contains 593 videos of 123 different subjects with age ranging from 18 to 50 years. Moreover, approximately 69% were female, 81% were Euro-American, 13% were African American, and 6% were from other groups.



Fig. 2. Sample images from the FER2013 dataset.²⁴

Table 1. Description of FER2013 dataset.

Label	Number of images	Emotion
0	4,953	Angry
1	547	Disgust
2	5,121	Fear
3	8,989	Happy
4	6,077	Sad
5	4,002	Surprise
6	6,198	Neutral



Fig. 3. Sample images from the CK+ dataset.²⁴

The CK+ dataset is mainly labeled with one of seven expression classes: Anger, disgust, fear, happiness, sadness, and surprise. Figure 3 shows some sample images from the CK+ dataset and Table 2 gives a description of the dataset.

AffectNet: AffectNet is a large-scale facial expression dataset containing about 400,000 images, which were manually analyzed according to eight facial expressions (neutral, happy, angry, sad, fearful, surprised, disgusted, and contempt) and their intensity of valence and arousal. This includes over 1 million facial images of him collected from the web by querying three major search engines using 1,250 emotion-related keywords in six different languages. Approximately half of the acquired

Table 2. Description of CK+ dataset.

Label	Number of images	Emotion
0	5,941	Angry
1	9,735	Disgust
2	4,125	Fear
3	12,420	Happy
4	3,696	Sad
5	14,619	Surprise
6	2,970	Contempt



Fig. 4. Sample images from the AffectNet Dataset (0: neutral; 1: happy; 2: sad; 3: surprise; 4: fear; 5: disgust; 6: anger; 7: contempt).²⁴

images (~440 KB) were manually annotated for the presence of seven distinct facial expressions (categorical model) and intensity of valence and arousal (dimensional model). AffectNet is by far the largest existing database of facial expressions, valence, and excitement, enabling the study of automatic facial expression recognition in two different emotional models. Figure 4 shows some sample images from the AffectNet dataset and Table 3 gives a distribution of the eight classes in the training set.

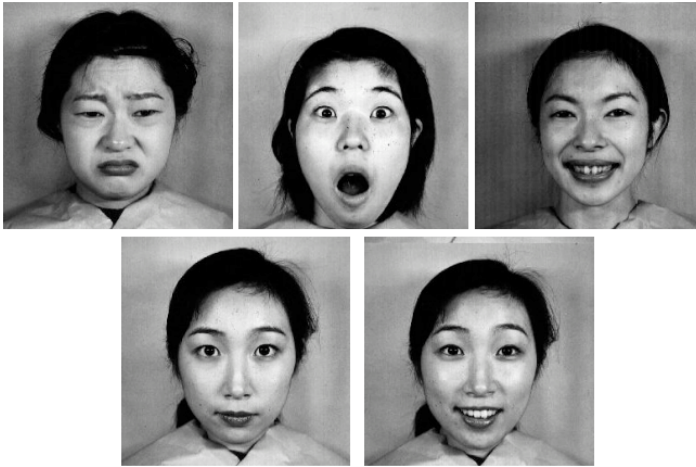
JAFPE: The JAFPE dataset comprises of 213 pictures showcasing various facial expressions portrayed by 10 Japanese women as shown in Table 4 and Fig. 5. Each participant was instructed to display seven different facial expressions, including

Table 3. Distribution of the eight classes in the training set.

Label	Emotion	Percentage accuracy (%)
0	Neutral	26.2
1	Happy	47.2
2	Sad	8.8
3	Surprise	4.8
4	Fear	2.1
5	Disgust	1.2
6	Anger	8.6
7	Contempt	1.2

Table 4. Description of JAFFE dataset.

Label	Number of images	Emotion
0	4,840	Angry
1	4,840	Disgust
2	4,842	Fear
3	4,842	Happy
4	4,841	Sad
5	4,840	Surprise
6	4,840	Neutral

Fig. 5. Sample images from the JAFFE dataset.²⁴

happiness, surprise, fear, disgust, anger, sadness, and neutrality. To provide further insights, the images were meticulously annotated with average semantic ratings for each facial expression by a panel of 60 annotators.

3.2. Experimental procedure

The overall process of face emotion recognition consists of three phases: Preprocessing, face detection, and sentiment classification. In the preprocessing phase, the dataset is prepared to work with generalized algorithms and generate efficient results. The face detection phase involves detecting faces in real-time captured images. Finally, the sentiment classification step implements a CNN algorithm to classify an input image into one of six emotion classes. These phases illustrated in the flowchart are presented in Figs. 6 and 7.

3.3. Preprocessing

Input images to FER are noisy and may show variations in lighting, size, and color. To get several preprocessing operations were performed on the images to get more

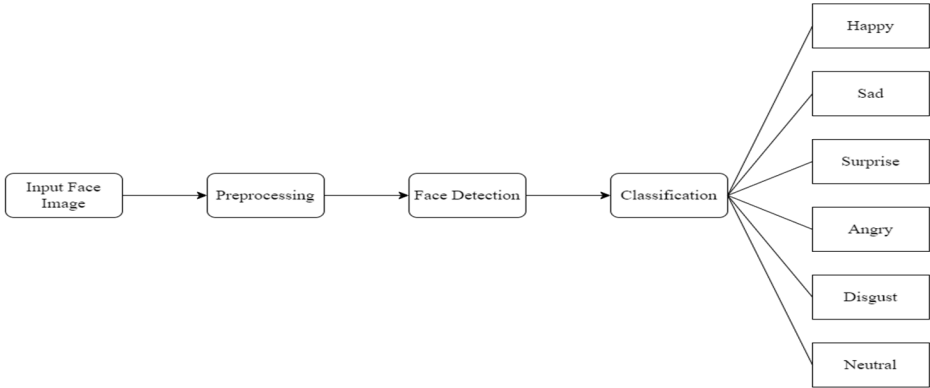


Fig. 6. Flowchart face classification phases.

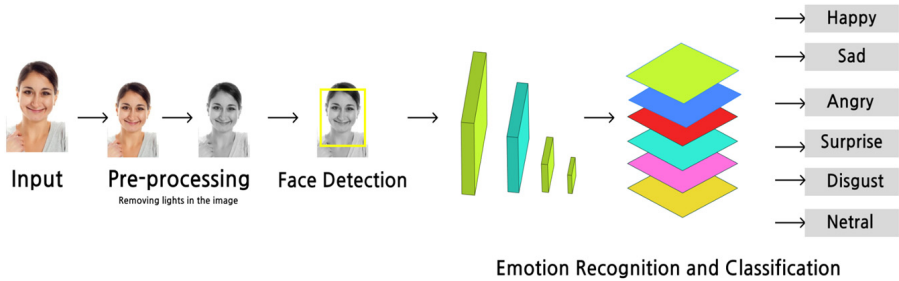


Fig. 7. System architecture block diagram of face emotion recognition.

accurate and faster results from the algorithm. The preprocessing strategies used include image conversion to grayscale, image normalization, and resizing.

- (1) Normalization — Image normalization is done to eliminate lighting variations and yield improvements in face image.
- (2) Grayscale — Grayscale is the process of converting a color input image into a pixelated image. The value depends on the intensity of light on the image. Color images are difficult to display, resulting in grayscale algorithmic processing.
- (3) Resize — The image is resized to remove unnecessary parts of the image. This reduces memory is required, which speeds up the computation.

The application of this preprocessing stage was performed on each of the datasets used, namely FER2013, CK+, AffectNet, and JAFFE.³³ However, it should be noted that the exact preprocessing method may vary depending on the specific characteristics of each dataset and algorithm used. Adjustments and optimizations of the preprocessing methods were made to suit the image characteristics of each dataset and provide better results in FER.

3.4. Face detection

The face detection phase is a crucial step in real-time FER systems as it involves detecting faces in captured images. In this study, the Haar cascades method, also known as Viola–Jones detectors, was employed for face detection.³³ Haar cascades are classifiers that have been trained to detect specific objects in images or videos, and they have shown high accuracy and efficiency in object detection. Haar cascades utilize Haar-like features to detect specific patterns on the face, such as the presence of dark and light regions. These features are trained using both positive (faces) and negative (non-faces) facial images. The detection process involves rapidly calculating pixel comparisons within the image to identify the presence of these features. By using Haar cascades, the algorithm can effectively filter out unnecessary background data and accurately detect the facial region. The face detection process using Haar cascade classifiers was implemented in the OpenCV library, a widely used computer vision library. This method, originally proposed by Papageorgiou *et al.*³⁴ utilizes rectangular features to identify facial patterns, as shown in Fig. 8. By employing the Haar cascades method for face detection, this study aims to accurately locate and extract facial regions from the captured images, which will serve as input for subsequent emotion recognition algorithms. The use of Haar cascades in face detection has been proven effective in numerous studies, making it a suitable choice for this research.

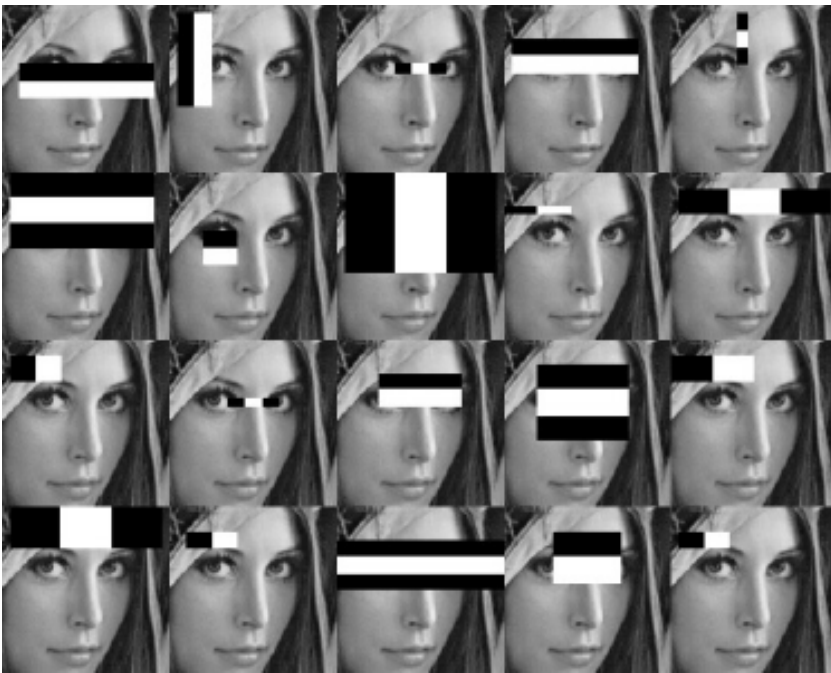


Fig. 8. Haar features examples for face detection.³⁵

3.5. *Emotion classification*

In this stage, the system performs classification of the image into one of the seven universal expressions: Happiness, Sadness, Anger, Surprise, Disgust, Fear, and Neutral, as labeled in the FER2013 dataset. The classification is done using CNN which have proven to be effective in image processing. The dataset is divided into training and test datasets, and the training is conducted on the training set without any prior feature extraction. To improve accuracy and minimize overfitting, different CNN architectures are experimented with during the emotion classification step. The process involves the following phases:

- (1) **Data Splitting:** The FER2013 dataset is split into three categories based on the “Usage” label: Training, PublicTest, and PrivateTest. The Training and PublicTest sets are used for model generation, while the PrivateTest set is used for model evaluation.
- (2) **Training and Model Generation:** The neural network architecture includes the following layers:
 - **Convolution Layer:** A learnable filter is convolved over the input to generate multiple feature maps.
 - **Max Pooling:** Reduces the spatial size of the input layer to decrease computation cost.
 - **Fully Connected Layer:** Neurons from the previous layer are connected to the output neurons, and the size of the final output layer corresponds to the number of emotion classes.
 - **Activation Function:** ReLu activation function is used to mitigate overfitting.
 - **Softmax:** Normalizes the output vector into a range of values between 0 and 1.
 - **Batch Normalization:** Speeds up the training process and maintains the mean activation close to 0 and the activation standard deviation close to 1.
- (3) **Model Evaluation:** The generated model is evaluated on the validation set, consisting of 3,589 images.
- (4) **Real-time Image Classification:** The concept of transfer learning is employed to detect emotions in real-time images. The pretrained weights and values from the generated model can be used for implementing facial expression detection in new scenarios, resulting in faster real-time image processing. In this context, “processing” encompasses the entire pipeline of capturing, analyzing, and classifying real-time images. The term “processing” involves various steps, including image capture, preprocessing for analysis, feeding the images through the deep learning model, and generating rapid classification results, such as identifying emotions on individuals’ faces. Transfer learning expedites this processing by leveraging pre-trained models that bring prior knowledge from extensive training on large datasets. These models can quickly adapt to the new task of emotion detection, making real-time image analysis more efficient, accurate, and

timely. The CNN architecture of Real Time Facial Expression Recognition is depicted in Fig. 9.

The following is an explanation of the parameters in the CNN architecture image above:

- **Features (F):** This refers to the number of filters or kernels used in the convolution layer. These filters detect specific features in the input data, such as edges or texture.
- **Kernel Size (K):** The kernel size represents the dimensions of the filters applied during the convolution operation. In 2D CNNs for image processing, the kernel size is usually expressed as $K \times K$, where K is an odd number such as 3 or 5. It defines the area where the filter slides across the input.
- **Stride (S):** Stride indicates how much the filter shifts after each convolution operation. Stride 1 means the filter moves one pixel at a time, while stride 2 makes it move two pixels at a time. A smaller stride results in a larger output feature map.
- **Padding (P):** Padding is used to control the spatial dimension of the output feature map. “Valid” padding means no padding is added, resulting in a smaller output size, while “Equal” padding adds zeros to the input to maintain the same spatial dimension in the output.
- **Activation (A):** The activation function introduces nonlinearity to the CNN. Common activation functions include Rectified Linear Unit (ReLU) and Sigmoid.

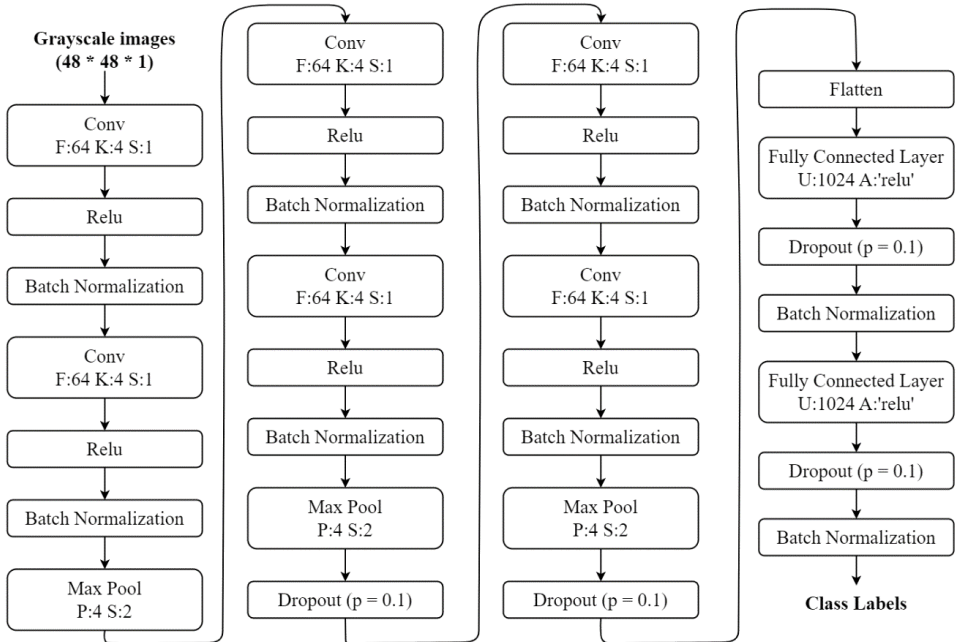


Fig. 9. CNN architecture.³⁵

These functions help the network learn complex patterns and relationships in the data.

- Units (U): In the context of CNNs, “Units” may refer to the number of neurons or units in a fully connected layer, also known as a dense layer, in the network.

These parameters are essential for configuring a CNN’s architecture and impact its ability to learn and extract meaningful features from input data.

4. Review of the Applications of Face Emotion Recognition Studies

This section specifies the algorithms that were evaluated to recognize facial emotions and the outcomes obtained. The algorithms were tested using four main datasets: FER2013, CK+, AffectNet, and JAFFEE. This research analyzes and compares the accuracy of the algorithms and other six basic facial expressions including, happiness, sadness, fear, disgust, surprise, and neutral expressions. Among these datasets, the Hybrid Attention Cascade Network performed exceptionally well on the Cohn–Kanade dataset, achieving an impressive overall accuracy of 98.46%. Table 5 illustrates the mapping of datasets and algorithms that are used, and the overall accuracy achieved by each algorithm in face emotion recognition:

Several studies have made notable advancements in the CNN algorithm that is evaluated using the FER2013 dataset. In Ref. 9, Deepface algorithm achieved an initial accuracy of 72.02%, followed by 39.13% accuracy in a controlled lab environment, and 26.6% accuracy in a real-world factory setting. In comparison, Ref. 36 reported a slightly higher accuracy of 74.0% and demonstrated exceptional performance in happiness (91.0%) and surprise (83.0%) expressions using the CNN algorithm. Furthermore, a work by Ref. 45 employed data augmentation techniques and a newly proposed method called RM-Exception Network and achieved an accuracy of 73.32% after 127 training epochs, which reveals that the accuracy will increase as the training algorithm increases.

In the evaluation of various algorithms using the Cohn–Kanade (CK+) dataset,²⁵ introduced the Hybrid Attention Cascade Network, which showed considerable improvements in accuracy compared to previous methods like CNN and VGG-16. This approach outperforms other algorithms and dataset combinations tested in prior studies, achieving an exceptional accuracy rate of 98.46%. Moreover, it reveals that the expressions of contempt and surprise reached the highest accuracy rate, reaching a perfect score of 100%. The data processing involved extracting images, reducing noise, and normalizing them to a resolution of 24×24 pixels. The training process in this study utilized 100 epochs. These findings demonstrate the superior performance of the Hybrid Attention Cascade Network in FER tasks.

The study in Ref. 40 that used the AffectNet dataset reported accuracy values for different emotions as follows: Anger (70%), disgust (65%), happiness (70%), neutral (65%), and surprise (70%). Introducing weight decay initially resulted in an overall

Table 5. Results Comparison of FER2013, CK+, AffectNet and JAFFEE dataset in different algorithms.

No.	Dataset	Algorithm	Result
1	FER2013	DeepFace	DeepFace achieved an overall accuracy of 72.0% on the FER2013 dataset, accurately predicting 18/25 photos. ⁹
2	FER2013	Convolutional Neural Network	The CNN model achieved an overall accuracy of 74.0% on FER2013. Specific accuracies were 66% for anger, 69% for fear, 64% for disgust, 91% for happiness, 63% for sadness, 83% for surprise, and 77% for neutral expressions. ³⁶
3	FER2013	RM-Xception network	After 127 Epochs, the RM-Xception network achieved 73.32% accuracy. ³⁷
4	Cohn–Kanade	Convolutional Neural Network	CNN achieved an impressive 98.0% accuracy on the Cohn–Kanade dataset. ³⁸
5	Cohn–Kanade	Hybrid Attention Cascade Network	Recognition accuracy rate from CK+ dataset was 98.46%. Expressions of contempt and surprise achieved perfect 100% accuracy. ²⁵
6	Cohn–Kanade	VGG-16	VGG-16 reached 52.49% accuracy. Detailed accuracies: Anger 66%, disgust 93%, sadness 64%, and fear 48%. ³⁹
7	AffectNet	ERMOPI	Anger, disgust, happiness, neutrality, and surprise achieved accuracies of 70%, 65%, 70%, 65%, and 70%, respectively. Weight decay improved overall accuracy from 55% to 68%. ⁴⁰
8	AffectNet	Deep Belief Network	Deep Belief Network displayed a remarkable 95.34% accuracy. ⁴¹
9	AffectNet	Convolutional Neural Network	CNN achieved 69.3% accuracy on AffectNet. Highest accuracies: Happiness 90.3%, Surprise 80.7%. Lower rates for Fear, Sadness, and Anger due to minor landmark adjustments. Fear and disgust reached 72.5%. ⁴²
10	JAFFEE	CNN+SVM	The combined model achieved around 95.31% accuracy on JAFFE. ³⁸
11	JAFFEE	Attentional Convolutional Network	CNN achieved 92.8% accuracy on the JAFFE dataset. ⁴³
12	JAFFEE	Spatial pyramid Zernike moments-based shape features and Law’s texture features	The algorithm achieved 95.86% accuracy. MLPNN was 94.35% and RBFNN was 95.86%. ^{35,44}

accuracy of 55%. However, after implementing the weight, the overall accuracy significantly improved to 68%. These findings highlight the impact of weight decay on enhancing the accuracy of the algorithm. However, a significant improvement is achieved using the Deep Belief Network,⁴¹ where it reached an accuracy of 98.82%. In this paper, the LBPNet and particle swarm optimization-based feature extraction technique is used to optimize the method and extract the features from the AffectNet dataset. It then utilized a deep belief network for image classification. With optimal features and efficient classification algorithms, the result reveals a higher rate of accuracy and low error rates in FER.

Table 6. Advantages and disadvantages of face emotion recognition technology.

Advantages	Disadvantages
<ul style="list-style-type: none">● Accuracy: Deep learning-based FER models, such as Deep Belief Network, have shown high accuracy in detecting and classifying facial expressions. They can automatically learn discriminative features from raw facial images, leading to improved accuracy compared to traditional handcrafted feature-based approaches.● Real-time Processing: Deep learning models can perform real-time FER, enabling applications in real-world scenarios where immediate responses are required. This is particularly beneficial in areas such as human-computer interaction, virtual reality, and emotion-based marketing.● Robustness: CNN models have proven to be robust in handling variations in lighting conditions, head poses, occlusions, and facial attributes. They can generalize well to different individuals and diverse environments, making them suitable for practical applications.● Feature Extraction: Deep learning models can automatically extract relevant features from facial images, eliminating the need for manual feature engineering. This reduces the dependency on domain knowledge and allows the models to learn complex patterns and representations directly from the data.● Deep Belief Networks (DBNs) provide valuable advantages in emotion recognition due to their hierarchical feature learning, unsupervised pre-training for effective utilization of limited labeled data, efficient dimensionality reduction, adaptability to diverse data types, the capability to handle complex emotional patterns, and enhanced robustness in recognizing emotions across varying individuals and contexts. Their hierarchical structure allows the extraction of multi-level features making them adept at capturing nuanced emotional expressions.	<ul style="list-style-type: none">● Data Requirements: Deep learning models for FER typically require large amounts of labeled data for training. Acquiring and annotating such datasets can be time-consuming and labor-intensive, especially when considering the need for diverse facial expressions and cultural variations.● Overfitting: Deep learning models are prone to overfitting, especially when the training dataset is limited or imbalanced. Overfitting can lead to poor generalization and performance degradation on unseen data, impacting the accuracy and reliability of the FER system.● Hardware and Computational Resources: Deep learning models, particularly complex CNN architectures, require significant computational resources and specialized hardware, such as high-performance GPUs, for efficient training and inference. This can pose limitations for deployment on resource-constrained devices or in real-time applications with strict latency requirements.● Environment condition: Variations in conditions of images such as lighting, background, and noise also may impact the overall result accuracy of the recognition, particularly in real-time face emotion recognition.● DBNs offer powerful neural network capabilities but come with several limitations. Training DBNs can be computationally expensive and time-consuming, often requiring layer-wise pre-training and substantial labeled data. DBNs tend to overfit, and their operation as black-box models can result in a lack of interpretability. Their performance is heavily data-dependent and inadequate or noisy data can lead to suboptimal results. Additionally, access to high-performance computing resources is often necessary. Selecting appropriate hyperparameters can be challenging and DBNs may not be universally applicable. They pose challenges in terms of transparency, particularly in applications where interpretability is essential. Finally, scalability concerns emerge when building very deep networks often associated with vanishing, and exploding gradient problems.

The last dataset that is being tested is JAFFEE, where in Ref. 35, it shows a significant accuracy of 95.86%. In the paper, each image in the dataset was computed using the Spatial Pyramid Zernike Moments (SPZM) combined with texture features called Local Texton Models (LTExM). Furthermore, the SPZM feature went on

extractions using the ZM orders of $n = 1, 2, 3$. Then, the SPZM experienced normalization and is also combined with LTexM. The results are then used to train the neural network classifier, which enables the model to predict unseen samples in face expressions. Although the SPZM algorithm stands out from CNN and ACN, both algorithms also performed well, where it reached 95.31% and 92.8% accuracy rates, respectively.

Overall, FER203 and CNN achieved the highest accuracy of 74.0%. In contrast, the hybrid cascade network demonstrated remarkable performance with an almost perfect accuracy score of 98.46% on the Cohn–Kanade dataset. Similarly, the deep belief network outperformed previous algorithms on the AffectNet dataset, achieving a slightly higher accuracy rate of 98.82%. Lastly, the combination of the JAFFEE dataset with Spatial Pyramid Zernike moments-based shape features and Law's texture features yielded the highest accuracy rate of 95.86%.

5. Discussion

The discussion section of this paper aims to critically analyze the advantages and disadvantages of face emotion recognition technology. Face emotion recognition has emerged as a prominent field of research, offering the potential to revolutionize various sectors, including healthcare, security, and human–computer interaction. This technology utilizes sophisticated algorithms to identify and interpret facial expressions, providing valuable insights into an individual's emotional state. However, as with any technological advancement, face emotion recognition also presents certain limitations and challenges that need to be addressed. This discussion explores the advantages offered by face emotion recognition, such as improved emotional understanding and personalized experiences, while also examining its disadvantages, including privacy concerns, accuracy limitations, and potential biases. By examining both the positive and negative aspects, a balanced perspective on the implications and prospects of face emotion recognition technology can be provided. Table 6 describes the benefits and limitations of Face Emotion Recognition Technology.

6. Conclusions and Future Work

In conclusion, this paper has provided a comprehensive review of the advances and challenges in FER. It also provided a comparison of the performances of various deep learning algorithms that are evaluated using the four main datasets including FER2013, Cohn–Kanade (CK+), AffectNet, and JAFFEE datasets. The utilization of deep learning techniques, particularly CNNs has revolutionized the field by enabling the automatic extraction of discriminative features from raw facial images. A new field algorithm called the DBN however outperformed the algorithms from previous studies, illustrating the highest accuracy rate of 98.82%. Nonetheless, challenges such as the inherent ambiguity of facial expressions, individual differences in expressing and perceiving emotions, and variations in lighting conditions and


facial attributes still pose significant obstacles. Future research focuses on addressing these challenges by developing robust and generalizable systems that can accurately recognize and interpret facial expressions in real-world scenarios. Additionally, the integration of face emotion recognition has the potential to enhance various fields, including healthcare systems, mental health, safety measures, and crime investigations, among others. This technology offers opportunities for advancements and benefits that extend beyond these areas as well.

Acknowledgment

This research is sponsored by the DIREKTORAT RISET DAN PENGABDIAN MASYARAKAT Satya Wacana Christian University, Indonesia.

ORCID

Christine Dewi  <https://orcid.org/0000-0002-1284-234X>

Henoch Juli Christanto  <https://orcid.org/0000-0003-0276-295X>

References

1. K. Kaulard, D. W. Cunningham, H. H. Bülthoff and C. Wallraven, The MPI facial expression database — A validated database of emotional and conversational facial expressions, *PLoS One* **7**(3) (2012) e32321, doi: 10.1371/journal.pone.0032321.
2. J. Chen, Exploring the impact of teacher emotions on their approaches to teaching: A structural equation modelling approach, *Br. J. Educ. Psychol.* **89** (2019) 57–74, doi: 10.1111/bjep.12220.
3. D. K. Jain, P. Shamsolmoali and P. Sehdev, Extended deep neural network for facial emotion recognition, *Pattern Recognit. Lett.* **120** (2019) 69–74, doi: 10.1016/j.patrec.2019.01.008.
4. O. Mohamad Nezami, M. Dras, L. Hamey, D. Richards, S. Wan and C. Paris, Automatic recognition of student engagement using deep learning and facial expression, in *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, Vol. 11908 (Springer, 2020), pp. 273–289, doi: 10.1007/978-3-030-46133-1_17.
5. D. C. Rubin and J. M. Talarico, A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words, *Memory* **17** (2009) 802–808, doi: 10.1080/09658210903130764.
6. A. I. Staff, M. Luman, S. van der Oord, C. E. Bergwerff, B. J. van den Hoofdakker and J. Oosterlaan, Facial emotion recognition impairment predicts social and emotional problems in children with (subthreshold) ADHD, *Eur. Child Adolesc. Psychiatry* **31** (2022) 715–727, doi: 10.1007/s00787-020-01709-y.
7. J. Panksepp, *Affective Neuroscience: The Foundations of Human and Animal Emotions* (Oxford University Press, Oxford, 2005).
8. A. Mollahosseini, B. Hasani and M. H. Mahoor, AffectNet: A database for facial expression, valence, and arousal computing in the wild, *IEEE Trans. Affect. Comput.* **10** (2019) 18–31, doi: 10.1109/TAFFC.2017.2740923.
9. A. Chiurco, J. Frangella, F. Longo, L. Nicoletti, A. Padovano, V. Solina, G. Mirabelli and C. Citraro, Real-time detection of worker's emotions for advanced human-robot

- interaction during collaborative tasks in smart factories, *Procedia Comput. Sci.* **200** (2022) 1875–1884.
10. M. Parashakthi and S. Savithri, Facial emotion recognition-based music recommendation system, *Int. J. Health Sci.* **6**(S4) (2022) 5829–5835, <https://doi.org/10.53730/ijhs.v6nS4.9419>.
 11. S. E. Kahou, V. Michalski, K. Konda, R. Memisevic and C. Pal, Recurrent neural networks for emotion recognition in video, in *Int. Conf. Multimodal Interaction 2015* (ACM, 2015).
 12. C. Dewi and R. C. Chen, Automatic medical face mask detection based on cross-stage partial network to combat COVID-19. *Big Data Cogn. Comput.* **6** (2022) 104, doi: 10.3390/bdcc6040106.
 13. K. Zhang, Z. Zhang, Z. Li and Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **23**(10) (2016) 1499–1503, doi: 10.1109/LSP.2016.2603342.
 14. S. Turabzadeh, H. Meng, R. Swash, M. Pleva and J. Juhar, Facial expression emotion detection for real-time embedded systems, *Technologies (Basel)* **6** (2018) 17, doi: 10.3390/technologies6010017.
 15. G. Simcock, L. T. McLoughlin, T. De Regt, K. M. Broadhouse, D. Beaudequin, J. Lagopoulos and D. F. Hermens, Associations between facial emotion recognition and mental health in early adolescence, *Int. J. Environ. Res. Public Health* **17** (2020) 330, doi: 10.3390/ijerph17010330.
 16. L. Bourke, J. Lingwood, T. Gallagher-Mitchell and B. López-Pérez, The effect of face mask wearing on language processing and emotion recognition in young children, *J. Exp. Child Psychol.* **226** (2023) 105580, doi: 10.1016/j.jecp.2022.105580.
 17. G. Yu, Emotion monitoring for preschool children based on face recognition and emotion recognition algorithms, *Complexity* **2021** (2021) 1–12, doi: 10.1155/2021/6654455.
 18. C. Pabba and P. Kumar, An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition, *Expert Syst.* **39** (2022) 1–28, doi: 10.1111/exsy.12839.
 19. D. Li, L. Yu, J. He, B. Sun and F. Ge, Action recognition based on multiple key motion history images, in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, Chengdu, China, 2016, pp. 993–996, doi: 10.1109/ICSP.2016.7877978.
 20. A. Kläser, M. Marszałek and C. Schmid, A spatio-temporal descriptor based on 3D-gradients, in *Proc. British Machine Vision Conf., Sep. 2008* (Leeds, United Kingdom, 2008), pp. 275:1–10, inria-00514853.
 21. N. Nourbakhsh Kaashki and R. Safabakhsh, RGB-D face recognition under various conditions via 3D constrained local model, *J. Vis. Commun. Image Represent.* **52** (2018) 66–85, doi: 10.1016/j.jvcir.2018.02.003.
 22. N. A. S. Badrullisham and N. N. A. Mangshor, Emotion recognition using convolutional neural network, *J. Phys. Conf. Ser.* **1962** (2021) 012040.
 23. A. Oyedeki, Facial Emotion Detection: A Comprehensive Exploration of Convolutional Neural Networks, Vol. 11, No. 3, p. 130, doi: 10.22624/AIMS/DIGITAL/V11N4P1.
 24. G. E. Hinton, S. Osindero and Y. W. Teh, A fast learning algorithm for deep belief nets, *Neural Comput.* **18** (2006) 1527–1554, doi: 10.1162/neco.2006.18.7.1527.
 25. X. Zhu, S. Ye, L. Zhao and Z. Dai, Hybrid attention cascade network for facial expression recognition, *Sensors* **21** (2021) 2003, doi: 10.3390/s21062003.
 26. E. Golinko and X. Zhu, Generalized feature embedding for supervised, unsupervised, and online learning tasks, *Inf. Syst. Front.* **21** (2019) 125–142, doi: 10.1007/s10796-018-9850-y.

27. J. F. Hu, T. Z. Huang, L. J. Deng, T. X. Jiang, G. Vivone and J. Chanussot, Hyper-spectral image super-resolution via deep spatio-spectral attention convolutional neural networks, *IEEE Trans. Neural Netw. Learn. Syst.* **33** (2022) 7251–7265, doi: 10.1109/TNNLS.2021.3084682.
28. S. Rodtook and S. Makhanov, Rotationally invariant filter bank for pattern recognition of noisy images, *J. Intell. Fuzzy Syst.* **17** (2006) 71–82.
29. P. Ardhiyanto, B. Y. Liao, Y. K. Jan, J. Y. Tsai, F. Akhyar, C. Y. Lin, R. B. R. Subiakto and C. W. Lung, Deep learning in left and right footprint image detection based on plantar pressure, *Appl. Sci.* **12**(17) (2022) 8885.
30. M. Ramzan and S. Dawn, Fused CNN-LSTM deep learning emotion recognition model using electroencephalography signals, *Int. J. Neurosci.* **133**(6) (2023) 587–597.
31. D. Li, L. Xie, Z. Wang and H. Yang, Brain emotion perception inspired EEG emotion recognition with deep reinforcement learning, *IEEE Trans. Neural Netw. Learn. Syst.* (1 May 2023), doi: 10.1109/TNNLS.2023.3265730.
32. Y. T. Kim and S. Y. Han, Cooling channel designs of a prismatic battery pack for electric vehicle using the deep Q-network algorithm, *Appl. Thermal Eng.* **219** (2023) 119610, doi: 10.1016/j.applthermaleng.2022.119610.
33. D. M. Abdulhussien and L. J. Saud, Evaluation study of face detection by Viola-Jones algorithm, *Int. J. Health Sci. (Qassim)* **6** (2022) 4174–4182, doi: 10.53730/ijhs.v6ns8.13127.
34. A. Mohan, C. Papageorgiou and T. Poggio, Example-based object detection in images by components. *IEEE Trans. Pattern Anal. Mach. Intell.* **23** (2001) 349–361, doi: 10.1109/34.917571.
35. V. G. V. Mahesh, C. Chen, V. Rajangam, A. N. J. Raj and P. T. Krishnan, Shape and texture aware facial expression recognition using spatial pyramid Zernike moments and Law's textures feature set, *IEEE Access* **9** (2021) 52509–52522, doi: 10.1109/ACCESS.2021.3069881.
36. Z. Song, Facial expression emotion recognition model integrating philosophy and machine learning theory, *Front. Psychol.* **12** (2021) 759485, doi: 10.3389/fpsyg.2021.759485.
37. Y. Shang, M. Yang, J. Cui, L. Cui, Z. Huang and X. Li, Driver emotion and fatigue state detection based on time series fusion, *Electronics (Switzerland)* **12** (2023) 26, doi: 10.3390/electronics12010026.
38. U. Sabina and T. K. Whangbo, Edge-based effective active appearance model for real-time wrinkle detection, *Skin Res. Technol.* **27** (2021) 444–452, doi: 10.1111/srt.12977.
39. S. Ghafourian, R. Sharifi and A. Baniyadi, Facial emotion recognition in imbalanced datasets, *Academy and Industry Research Collaboration Center (AIRCC)*, May 2022, pp. 239–251, doi: 10.5121/csit.2022.120920.
40. S. Kuruvayil and S. Palaniswamy, Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning, *J. King Saud Univ. — Comput. Inf. Sci.* **34** (2022) 7271–7282, doi: 10.1016/j.jksuci.2021.06.012.
41. K. Babu, C. Kumar and C. Kannaiyaraju, Face recognition system using deep belief network and particle swarm optimization, *Intell. Autom. Soft Comput.* **33** (2022) 317–329, doi: 10.32604/iasc.2022.023756.
42. M. Mukhiddinov, O. Djuraev, F. Akhmedov, A. Mukhamadiyev and J. Cho, Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people, *Sensors* **23** (2023) 1080, doi: 10.3390/s23031080.
43. S. Minaee, M. Minaei and A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, *Sensors* **21** (2021) 3046, doi: 10.3390/s21093046.

44. V. G. V. Mahesh, C. Chen, V. Rajangam, A. N. J. Raj and P. T. Krishnan, Shape and texture aware facial expression recognition using spatial pyramid Zernike moments and Law's textures feature set, *IEEE Access* **9** (2021) 52509–52522, doi: 10.1109/ACCESS.2021.3069881.
45. Y. Shang, M. Yang, J. Cui, L. Cui, Z. Huang and X. Li, Driver emotion and fatigue state detection based on time series fusion, *Electronics (Switzerland)* **12** (2023) 26, doi: 10.3390/electronics12010026.