Intelligent Methods in Engineering Sciences 2(1): 005-021, 2023

IMIENS

INTELLIGENT METHODS
IN ENGINEERING SCIENCES

https://www.imiens.org

International Open Access

March, 2023

e-ISSN 2979-9236

Review Article

https://doi.org/10.58190/imiens.2023.7

Real-Time Emotion Recognition Using Deep Learning Methods: Systematic Review

Muthana ALISAWI [a,b*], ***Nursel YALÇIN*** [c]

[a] *Department of computer sciences, College of computer sciences and information technology, Kirkuk University, Iraq*
[b] *Institute of Information, Computer sciences, Gazi University, Ankara – Türkiye*
[c] *Department of Computer and Instructional Technologies Education, Gazi Faculty of Education, Ankara, Türkiye*

ARTICLE INFO

Article history:
Received 09 December 2022
Accepted 21 February 2023

Keywords:
Dataset
Deep learning
Emotion recognition
Facial Expression Recognition
Real-time
Systematic review

ABSTRACT

The seven basic facial expressions are the most important indicator of a person's psychological state, regardless of gender, age, culture, or nationality. These expressions are an involuntary reaction that shows up on the face for a short time. They show how the person is feeling—sad, happy, angry, scared, disgusted, surprised, or neutral. The visual system and brain automatically detect a person's emotion through facial expressions. Most computer vision researchers struggle to automate facial expression recognition. Human emotion-detection pioneers have also tried to mimic human automatic detection. Thus, deep learning techniques are the closest to mimicking human intelligence. Despite deep learning techniques, creating a system that can accurately distinguish between facial expressions is still difficult due to the diversity of faces and the convergence of some expressions that express different emotions. This systematic review presents a scientifically rich paper on deep learning-based facial expression emotion detection methods. From 2019 to the present, PRISMA was used to search and select research on real-time emotions. The study collected datasets from the mentioned period that were used to train, test, and verify the models presented in the relevant studies. Each dataset was fully explained in terms of number of items, type of data, etc. The study also compared relevant studies and identified the best technique. Furthermore, challenges to systems that detect emotions through facial expressions have been identified.

1. INTRODUCTION

At the present time, despite the development in the field of interaction between computers and humans, researchers in the field of identifying human emotions still face difficulties and challenges when working to detect these feelings in an accurate manner due to the complexity and diversity that emotions represent. Emotions can be defined as the way in which a person expresses himself through a specific reaction (sadness, joy, anger, fear, disgust, surprise, neutrality, etc.) to a particular situation [1]. In order to understand and ease the use of these emotions in related fields, the researchers provided emotional models that are mainly classified into two types, namely, the dimensional emotional model and the discrete emotional model. In the first model, one's emotions are classified into (valence, arousal, etc.), where the intensity of an emotion is quantified by its Arousal level, whereas Valence defines whether the emotion is positive or negative. While the second model categorizes human emotions into accurate emotions (sadness, anger, joy, frustration, etc.), as shown in Fig.1 [2] [3]. Also, these different types of emotions can be expressed by using the Valence-Arousal model, as shown in Fig.2 [4].

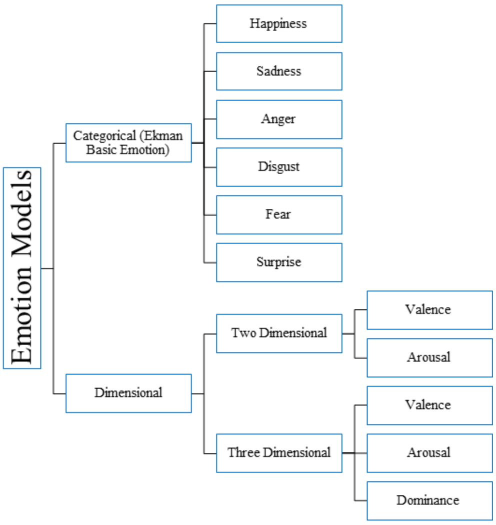

Figure 1. Emotional models and their classifications

[*] **Corresponding Author:** muthanayaseen@uokirkuk.edu.iq

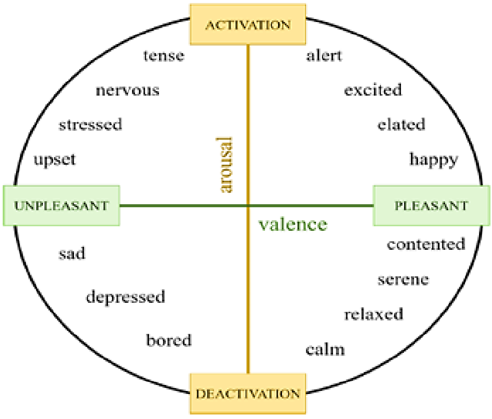

Figure 2. The Circumplex Model's Valence-Arousal dimensions

Nonetheless, in their studies of human emotions, most researchers have used the Ekman model, which represents six different emotions (sadness, disgust, anger, fear, surprise, and happiness) arising from the human neural network in response to an external stimulus. In addition to these six emotions, the researchers added another emotion, represented by the neutral, that indicates a person's state when there is no external stimulus[2]. Furthermore, these emotions have a significant impact on our daily lives and the activities we engage in and communicate with one another, as well as on our interaction and decisions in the areas of learning and work. In the framework of working to identify these human influences, Picard developed the concept known as "affective computing." From this concept, systems and devices were developed that interpret, process, and simulate human influences[5].

Emotion can be divided according to Albert Mehrabian into three main parts through which a person expresses himself with a specific reaction, where the visual information (emotional facial expressions and body language) is 55%, the vocal information (tone and volume of speech) is 38%, and the rest is a phrase for the verbal information (the content of the speech itself)[6]. In order to identify emotions, a set of vital information in the human body is used by researchers, depending on a number of methods. This vital information is employed in a system based on Human-Machine Interaction (HMI). This information is produced through a number of life activities in the human body, including breathing, such as Heart Rate Variability (HRV), Galvanic Skin Response signals (GSR), Facial Expression Recognition (FER), Speech Emotion Recognition (SER), perspiration (skin conductivity), body temperature, and neural activity as measured by Electroencephalography (EEG) signals[7]. Furthermore, we can categorize this information into three types: unimodal, bimodal, and multimodal, as shown in Fig. 3[8].

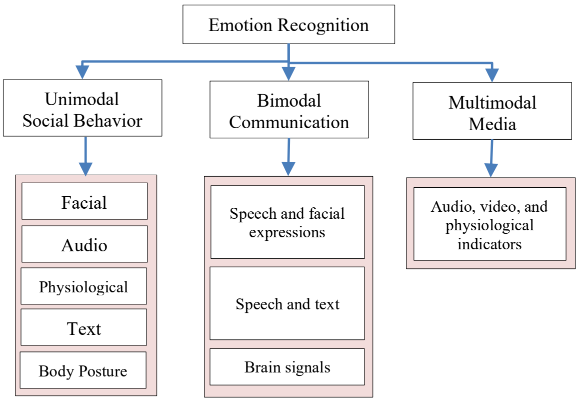

Figure 3. Information used in emotions recognition.

Emotion recognition, defined as, innate in human nature," is the process of recognizing feelings as a way of communicating through a range of emotional intentions (joy, sadness, etc.) regardless of gender and age, since the emotion of joy in an old person is the same as that expressed by a child[9] [10]. On the other hand, in the technological field, this process can be defined as the use of machine learning techniques (such as deep learning, etc.) to identify human emotions based on various visual cues, verbal expressions, and other indications. Also, among all these indications, emotion recognition based on facial expressions is one of the ideal methods used to identify one's feelings[11]. Moreover, the technique of recognizing emotions can be used in many fields in the context of human-computer interaction (HCI), including situation analysis of social interaction, affective computing, feedback during e-learning, psychology, contests, and the entertainment industries. Patient monitoring, forensics, medical aid, psychology, surveillance, the automotive industry, and human-robot interaction for autistic children are also areas of interest[12].

Furthermore, according to the most recent research, deep-learning feature-based approaches and hand-crafted feature-based techniques are the two categories used to classify facial expression detection. The preceding procedure is divided into two types (geometric features and appearance-based features)[13]. The study will include methods for detecting emotions (automatic, static images) and in particular based on facial expressions, where the internal feelings of humans can be understood through facial expressions, which play an important role in social interaction, as shown in the scope of the study in Fig.4.

This study relied on the standard of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (as shown in Fig. 5) in order to present a research paper with scientific abundance in the field of research. The research paper is organized as follows:

- The second section presents the research methodology.

- The third section contains background concepts.
- The fourth section contains the results of previous studies as well as related works.
- The fifth section includes the discussion.
- The sixth and final section includes the conclusion.

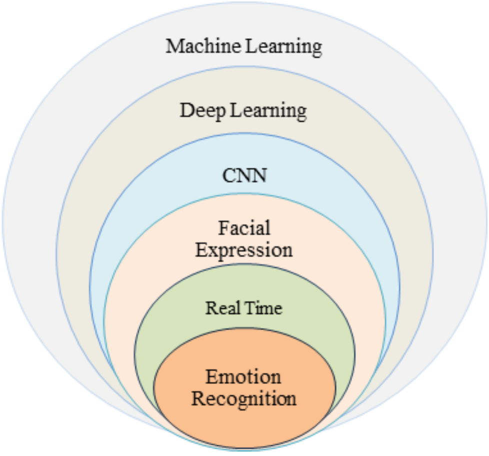

Figure 4. Scope of the systematic review

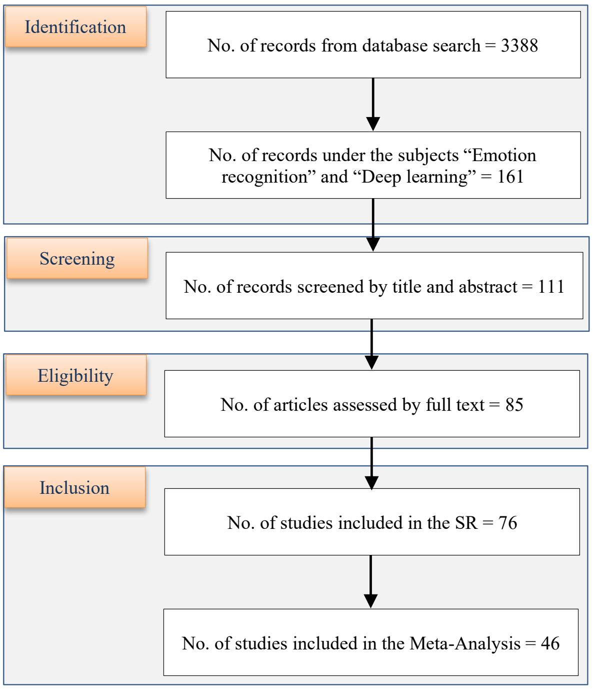
Identification
No. of records from database search = 3388
No. of records under the subjects "Emotion recognition" and "Deep learning" = 161
Screening
No. of records screened by title and abstract = 111
Eligibility
No. of articles assessed by full text = 85
Inclusion
No. of studies included in the SR = 76
No. of studies included in the Meta-Analysis = 46

Figure 5. Flowchart of the systematic review according to PRISMA standard

2. THE RESEARCH METHODOLOGY

In this study, the process of searching for sources was carried out by relying on the EBSCO Publishing Service, which provides advanced searches and gives access to many international publishing containers through an email linked to university institutions. The figure below (Fig. 6) shows the search process for our systematic review.

Fig.6 Emotion Recognition-Search definition for our systematic review.

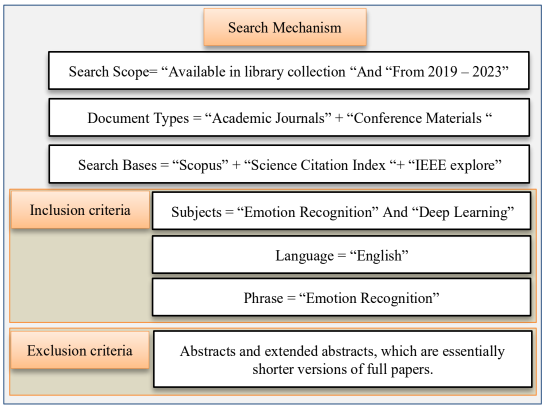
Search Mechanism
Search Scope= "Available in library collection "And "From 2019 – 2023"
Document Types = "Academic Journals" + "Conference Materials "
Search Bases = "Scopus" + "Science Citation Index "+ "IEEE explore"
Inclusion criteria
Subjects = "Emotion Recognition" And "Deep Learning"
Language = "English"
Phrase = "Emotion Recognition"
Exclusion criteria
Abstracts and extended abstracts, which are essentially shorter versions of full papers.

Figure 6. Emotion Recognition-Search definition for our systematic review.

2.1. Research questions

In this study, we try to find answers to a number of questions that can help researchers figure out how to use deep learning techniques in the field of emotion recognition.

RQ1: What are the latest deep learning techniques used to recognize emotions?

RQ2: Which dataset is used most often to train, test, and validate deep learning models?

RQ3: What are the challenges in recognizing emotions through facial expressions?

RQ4: What is the most commonly used technique in the process of detecting faces in the images of the dataset?

RQ5: What are the systems that have adopted deep learning techniques and have achieved good performance in detecting emotions?

3. BACKGROUND CONCEPTS

In this section, the terms related to the study's topic will be explained. We want to cover all aspects of the process of detecting emotions using deep learning techniques, which raises ambiguity and provides a resource for researchers in the field of detecting emotions using deep learning techniques.

3.1. Deep learning (DL)

Deep Learning (DL) is a machine learning technique gaining popularity in computer vision, speech recognition, and NLP. Deep learning learns features directly from data, unlike standard machine learning, which requires hand-engineered feature extraction. With large datasets and better processing capacity, these methods can produce high-performing models. DL models are based on biological neural networks. Backpropagating a corrected error signal across the network changes the weighted connections between nodes and neurons based on example inputs and desired outputs. In summary, deep learning is of great benefit in the field of computer vision, especially when the appropriate data set is available in terms of accuracy, completeness, compatibility, consistency, and validity [14].

In recent years, the DL has been used successfully in a wide range of traditional applications, such as cyber security, natural language processing, bioinformatics, robotics and control, and medical information processing. The well-known ML approaches haven't done as well as DL; Fig. 7 simplifies the difference between deep learning and machine learning [15]. Moreover, deep learning is the process of learning many levels of representations of the underlying distribution of the data that will be modelled automatically. In other words, a deep learning system automatically pulls out the low-level and high-level features that are needed for grouping. High-level features are those that depend on other features in a hierarchical way. In computer vision, for example, this means that a deep learning algorithm will learn its own low-level representations from a raw image (like edge detectors, Gabor filters, and so on), then build representations that depend on those low-level representations (like linear or non-linear combinations of those low-level representations), and then repeat the process. Automatic representation learning is a big part of this technique because it gets rid of the need to create features by hand, which could take a lot of time [16].

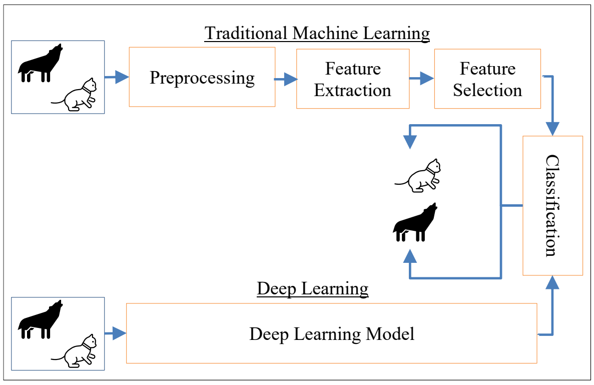

Figure 7. Deep learning vs. Machine learning

3.2. Convolution neural network (CNN)

CNN is one type of deep neural network, but there are many others. It was first suggested by Yann LeCun in 1988 and is often used in computer vision because it can handle a wide range of problems with image recognition. CNN can also beat humans in some of these challenges because it can find and categorize hidden patterns that are very hard for humans to see [17]. CNN is built so that it can get the most important parts of an image without changing the main parts of the image in a way that leads to the right prediction. After the image was put in, it went through several layers like convolution, pooling, flattening, and finally being fully connected as shown in Fig. 8 [18].

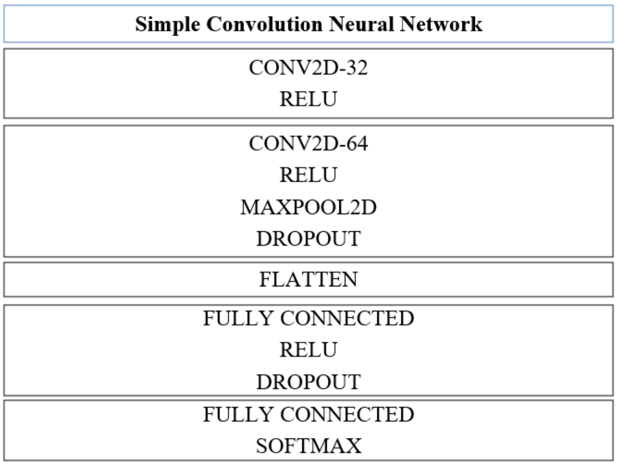

Figure 8. CNN simple architecture

Over the last decade, CNNs have made significant advances in pattern recognition, from processing images to recognizing voices. The most useful thing about CNNs is that they help cut down on the number of parameters in ANN. Because of this success, researchers and developers are now looking at bigger models to solve hard problems, which was not possible with classic ANNs. The most important assumption about problems that CNN can solve is that they shouldn't depend on where they are. In other words, an application that looks for faces doesn't have to worry about where the faces are in the images. The only thing that matters is finding them, no matter where they are in the images. It is also important to obtain abstract features as input moves toward the deeper layers of CNN. In image classification, for example, the edge might be found in the first layer, followed by simpler shapes in the second layer, and then higher-level features like faces in the layers after that[19]. Furthermore, CNN has shown that it is the best at getting features [20].

Also, CNNs use filters or "kernels" to convolve images to pull out features from them. When you use a filter on an image, the same feature shows up all over the image. After each action, the window moves, and the feature maps figure out what the features are. The feature maps consider the local receptive field of the image and work with weights and biases that are shared. Equation (1) can be used to figure out the size of the output matrix without padding, and Equation (2) shows how the convolution is done. Padding is used to keep the size of the image you send the same. "SAME" padding makes sure that the size of the output image is the same as the size of the input image, while "VALID" padding makes sure that the output image has no padding. The size of the output matrix with padding is shown in Equation (3) [21]. The training process of a convolutional neural network is shown by the back propagation, which is made up of four parts: the forward pass, the loss function, the backward pass, and the weights update [22].

$$N \times N \ * f \times f = N - F + 1 \tag{1}$$

$$O = \sigma(b + \sum_{i=0}^{2} \sum_{j=0}^{2} w_{i,j}\ h_{a+i,b+j}) \qquad (2)$$

$$N \times N\ * f * f = (N + 2P - f)/(s + 1) \qquad (3)$$

Where:

$O: Output$
$P: Padding$
$s: Stride$
$b: bais$
σ: sigmoidal activation function
w: 3×3 weight matrix of shared weights
h_{xy}: input activation at position x, y

As for the limitations of CNN technology, even though CNN-based FER has been used successfully, most CNN-based FER approaches only employ the input image to extract features from the external aspect of the face. Though, it's worth noting that geometric features are just as crucial for facial expression recognition [23].

To get a better understanding of each CNN layer, we will explain each layer separately as follows:

Convolution layer: The convolution layer is the first layer in a CNN model that an image will enter after the input layer. The two most important things that are used in this layer are the activation function and the filters. This layer has a group of filters that will be used on the image that comes in [24]. The convolution is made by these filters moving along the image. After the multiplication and addition, it makes a matrix that is smaller than the one it started with. The goal of that process is to find the best qualities. The result is then sent to the next layer. Moreover, filters are useful because they can find patterns in data. Each filter has two properties: stride and padding. Padding is the process of adding empty pixels to the image frame to make it bigger. The value of these empty pixels is zero. The stride is the number of pixels that the filter moves when it applies the convolution. The padding process is helpful because it keeps the image size from shrinking when filters are used (no loss of features in the image). Figure 9 shows an example of the filter that is used on the image that is read in[25].

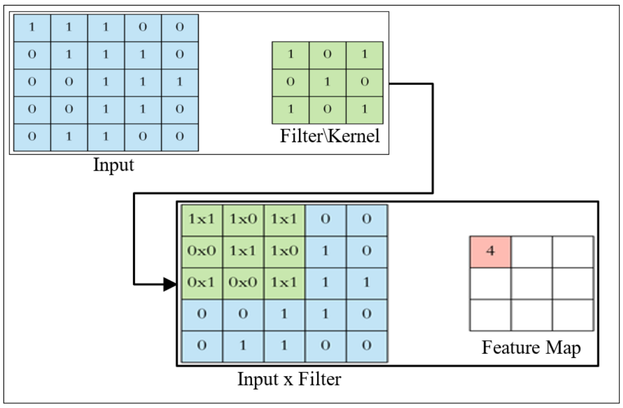

Figure 9. Make use of a filter on the image.

The different activation functions that represent one of the layers of the CNN model are as follows [26] [27]:

Rectified Linear Unit (ReLU): The ReLU activation function is the most common one used in deep learning models. In ReLU, the input pixel's value is kept if it's greater than zero and removed if it's less than zero, as shown in equation (4).

$$f\ (x) = max(0, x) \qquad (4)$$

SoftMax: A prediction of the distribution of multinomial probabilities can be made using the SoftMax function. It is an activation function that is implemented in the output layer of the neural network. In other words, SoftMax is used as an activation function for multi-class classification tasks that need class membership on more than two class labels. SoftMax returns a range of values between 0 and 1, with the sum of the probability being 1. SoftMax is computed by:

$$f\ (x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \qquad (5)$$

Sigmoid Function: In some literature, the sigmoid activation function is also known as the logistic function or the squashing function. Sigmoid function analysis has led to three variations of the sigmoid AF, all of which find usage in DL applications. Non-linear activation functions (AFs) like the Sigmoid are commonplace in feedforward neural networks. It's a smooth, bounded, differentiable real function with positive derivatives everywhere and a real-valued input range. The formula for the sigmoid function is:

$$f(x) = \left(\frac{1}{(1 + exp^{-1})}\right) \qquad (6)$$

Softsign: Softsign is another neural network (AF). Softsign non-linear AF for DL applications was introduced by Turian et al., 2009.Softsign is a quadratic polynomial with the following equation:

$$f(x) = \left(\frac{x}{|x| + 1}\right) \qquad (7)$$

Hyperbolic Tangent Function (Tanh): The hyperbolic tangent function is another activation function utilized in DL. The hyperbolic tangent function, or tanh function, is a smooth zero-centered function whose range is -1 to 1, as in the following equation:

$$f(x) = \left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right) \qquad (8)$$

Softplus Function: The Softplus Activation Function is a smooth variant of the ReLU function with smoothing and nonzero gradient features, boosting the stability and performance of deep neural networks containing Softplus units. The Softplus function was made by Dugas et al. in 2001. It is a basic form of the sigmoid function.

$$f(x) = log(1 + exp^x) \qquad (9)$$

Exponential Linear Units (ELUs): Clevert et al., (2015) suggest using exponential linear units (ELUs) to speed up deep neural network training. ELUs ease the vanishing gradient problem by employing identity for positive values and improving learning. Negative values reduce computational complexity and improve learning speed by moving mean unit activation closer to zero. ELU reduces bias shifts by pushing mean activation towards zero during training. The ELU equation is:

$$f(x) = \begin{pmatrix} x, & if\ x > 0 \\ \alpha\ exp(x) - 1, & if\ x \leq 0 \end{pmatrix} \quad (10)$$

Maxout Function: Maxout AF applies non-linearity to neural network weights and data as a dot product. The Maxout generalizes the leaky ReLU and ReLU in network computations where there are no dead neurons or saturation. Maxout is:

$$f(x) = max(w_1^T x + b_1, w_2^T x + b_2) \quad (11)$$

T = transpose, b = biases, w = weights

Exponential linear Squashing (ELiSH): Basirat and Roth, 2018, proposed the Exponential linear Squashing (ELiSH) AF. ELiSH and Swish share properties. The ELiSH function combines the ELU and Sigmoid functions.

$$f(x) = \begin{cases} \left(\dfrac{x}{1 + e^{-x}}\right), x \geq o \\ \left(\dfrac{e^x - 1}{1 + e^{-x}}\right), x < 0 \end{cases} \quad (12)$$

Pooling layer: The pooling layer reduces the array's dimensions using the convolved map or feature map as input. By using max-pooling or average pooling, the filter over the convolved feature will be smaller than the map. That is, the pooling layer will then do simple downsampling along the spatial dimensions of the input, thus lowering the number of parameters within the activation [25].

Flattening Layer: Flatness refers to the process of converting data into a one-dimensional matrix that can be sent to the next layer (the fully connected layer). After the convolution layer, the result is flattening, which makes a single long feature vector [25].

Fully connected layer (FC): The results of the convolution and pooling layers are crucial to the success of this layer. All the feature maps generated during the classification process are utilized and prepared in this final layer [25].

Loss Function: The loss function measures the deviation between the expected value and the actual value. The loss function finds the error value at the end of each iteration and then utilizes this value as feedback to update the model's weights [28].

3.3. Automated emotion recognition (AER)

Automated emotion recognition (AER) is the method through which computers automatically recognize human emotional responses. It is an essential research subfield within the field of human-computer interaction (HCI) and a developing field of application within artificial intelligence. AER has considerable potential in a variety of intelligent systems, such as education, marketing, and mental health monitoring [29] [9].

3.4. Facial emotion recognition (FER)

In human-machine interaction, facial expression recognition is fundamental. Real-world barriers like as lighting changes, major location variations, and partial or whole occlusions result in facial features with various degrees of clarity and completeness [30].

There are four distinct phases involved in facial expression recognition: image pre-processing, face detection, feature extraction, and facial expression classification. The most important and difficult part of face expression recognition is featuring extraction. Once the features have been recovered, the classifier will sort them into the appropriate emotional categories [31].

Moreover, there are two types of facial emotion recognition: FER based on machine learning and FER based on deep learning. In contrast to conventional FER methods, deep-learning-based methods (for example, convolutional neural networks) can be used from beginning to end, which greatly reduces the need for specialized training. Where Convolutional Neural Networks (CNN) are the most popular deep learning model for recognizing facial expressions [32].

3.5. Feature extraction techniques

Geometric feature-based and appearance-based methods are the two main types of feature extraction techniques. With so many potential obstacles, including changes in head attitude, lighting, the existence of optical obstructions, noise, and fuzzy images, facial feature extraction is a crucial component of any facial emotion detection system. Support Vector Machine (SVM), Neural Networks (NN), Random Forest, and K-Nearest Neighbor are only a few of the classifiers that have been studied in the literature. On the other hand, there are two ways to train a deep learning network: by transfer learning from an existing model or by creating a new model from scratch [12].

3.6. Transfer learning (TL)

The term "transfer learning" refers to the practice of applying a model trained on one activity to another. The core idea behind transfer learning is to apply the knowledge gained by a model trained on a large dataset to a smaller one. While transfer learning is computationally efficient and does not necessitate a large amount of data, training a convolutional neural network from scratch is time-consuming and data-intensive [3]. Therefore, time savings and improved accuracy are two primary benefits associated with transfer learning [33].

3.7. Optimizers

Adaptive Moment Estimation (Adam): One of the most widely used optimization methods for the purpose of training deep neural networks by combining momentum and RMS-prop.

Stochastic Gradient Descent (SGD): It has been empirically demonstrated that the stochastic gradient descent (also known as SGD) method for training deep neural networks is effective when used to huge datasets [34].

In some cases, adaptive optimization methods like Adam outperform stochastic gradient descent (SGD). Recent research indicates that Adam performs worse than SGD for training deep neural networks on image classification [35].

3.8. Face recognition systems

Face recognition systems are classified into three types based on their detection and recognition methods[36]:

Local approaches: People are put into groups based on the features of their faces instead of taking the face as a whole into account.

Holistic (subspace): This method takes data from the whole face, which is then projected onto a small subspace, called the correlation plane.

Hybrid approaches: This method can improve the accuracy of face recognition because it uses both local and global features.

3.9. Action units

In 1978, Ekman and Friesen developed the Facial Action Coding System. They were divided into 46 distinct face action units based on the facial muscular structure. These facial action units can be combined to create the six fundamental emotions. These AU were separated into the top and lower faces. The six fundamental emotions (joy, anger, sadness, fear, surprise, and disgust) are formed from these AU. Figure 10 is an illustration of a face with action units[37].

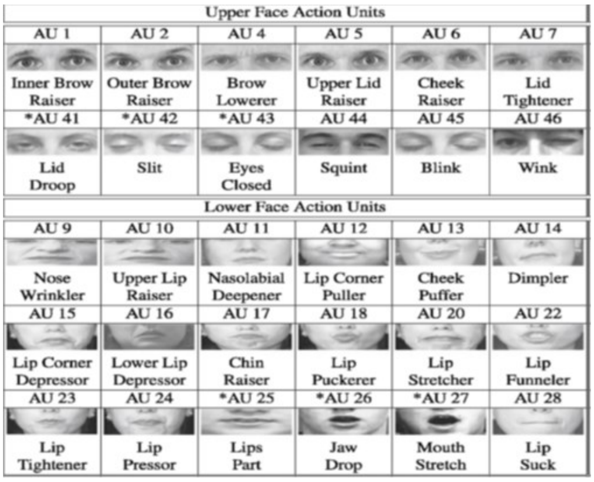

Figure 10. Action Units of Upper and lower face

3.10. Real-time emotion detection hardware

In order to implement the process of detecting emotions through facial expressions in real time and based on easy-to-carry devices such as glasses, a number of programmable boards that are easy to carry are required. In this section, we show a number of such programmable boards:

Field-programmable gate array (FPGA): Since they are changeable and may be programmed to implement any digital logic, field-programmable gate arrays are perfect for adaptive systems. Face recognition and online failure recovery are examples of applications for adaptive systems based on FPGAs. FPGAs are less efficient than application-specific integrated circuits (ASICs) as a result of the additional circuitry required to make them reconfigurable. Wherefore, FPGAs are projected to be three to four times slower, five to thirty-five times larger, and seven to fourteen times less energy efficient than ASICs, depending on the application and the FPGA's flexibility [38].

Raspberry Pi3 B: The Raspberry Pi 3 Model B represents the third version of the Raspberry Pi. This powerful credit-card-sized-single board computer can be used for a variety of applications and replaces the original Raspberry Pi Model B+ and Raspberry Pi 2 Model B. While maintaining the popular board format, the Raspberry Pi 3 Model B features a 10x faster processor than the original Raspberry Pi. In addition, it incorporates wireless LAN and Bluetooth connectivity, making it the optimal alternative for highly linked designs [39].

4. DATASETS

In this section, a group of the most famous datasets used in the field of emotion recognition (ER), based on facial expression recognition (FER), and different techniques are demonstrated, as shown in Table I.

Table I includes the datasets used by various studies to identify emotions using different techniques. Previous studies on emotion recognition used different datasets obtained by collecting them in special laboratories, in the usual external environment, or by using Internet of Things (IOT) technologies to obtain a dataset of different biological signals. This section will also include detailed information about the above datasets, as follows:

The Multimodal EmotionLines Dataset (MELD) based on the EmotionLines dataset derived from the TV series Friends TV. The dataset contained 1,400 dialogues and 13,708 speeches and was categorized into 7 emotions (Fear, Anger, disgust, neutrality, joy, surprise, sadness).

Table I. Detailed of Datasets

(ID.: IDENTITIES, N\A: NOT APPLICABLE, W: IN THE WILD, L: IN THE LABORATORY, E. NO.: NUMBER OF EMOTIONS)

Datasets	**Type of Samples**	**W*\L***	**E*.No.**	**No.**	**Id.***
[2] MELD	Utterances	W	N\A	*13708*	*N\A*
[2] IEMOCAP	Utterances	L	N\A	*10039*	*10*
[2] CIFE	Images	L	7	*14756*	*N\A*
[11] CK+	Videos	L	7	*593*	*123*
[11] JAFFE	Images	L	7	*213*	*10*
[11]RAF-DB	Images	L	6	*29672*	*N\A*
[40] RaFD	Images	L	8	*8040*	*67*
[40] AM-FED+	Videos	W	N\A	*1044*	*N\A*
[40] SFEW	Images	W	8	*1766*	*N\A*
[41][40] KDEF	Images	L	7	*4900*	*70*
[41] FER13	Images	W	7	*35887*	*N\A*
[42] KMU-FED	Images	W	6	*1101*	*12*
[13] RAVDESS	3 Modalities	W	8	*7356*	*1*
[32] LFW	Images	W	6	*13233*	*5749*
[32] Yale Face B	Images	L	6	*16128*	*28*
[32] GFEC	Images	L	30	*156000*	*N\A*
[43] SEED	EEG, Eye	L	3	*N\A*	*9*
[43] SEED-IV	EEG, Eye	L	4	*N\A*	*15*
[43] SEED-V	EEG, Eye	L	5	*N\A*	*16*
[43] DEAP	EEG, PPS	L	2-Bin.	*N\A*	*32*
[43] DREAMER	EEG, ECG	L	3-Bin.	*N\A*	*23*
[44] FDFB	Videos	L	4	*40*	*N\A*
[23] MMI	Videos\Image	L	6	*2900*	*75*
[45] UTKFace	Images	W	N\A	*20000*	*N\A*
[46] BAUM-1s	Videos	L	13	*1184*	*31*
[46] RML	Videos	L	6	*720*	*N\A*
[47] MUG	Images	L	6	*1462*	*86*
[48] CMU	Images	L	6	*750000*	*337*
[49] FEI	Images	L	2	*2800*	*200*
[49] IMFDB	Images	W	N\A	*34512*	*100*
[49] TFEID	Images	W	8	*7200*	*40*
[50] UMD	Images	W	N\A	*367888*	*N\A*
[50] CelebA	Images	W	N\A	*202595*	*N\A*
[51] MELD	Multi-M	W	N\A	*27000*	*6*
[52] CASMEII	Videos	L	7	*255*	*26*
[52] CASME	Videos	L	8	*189*	*19*
[52] FER+	Images	W	7	*35887*	*N\A*
[52] SAMM	Images	L	6	*133*	*N|A*
[53] AffectNet	Images	W	8	*291651*	*N\A*

The **Interactive Emotional Dyadic Motion Capture** (**IEMOCAP**) database are collected at the University of Southern California / SAIL Laboratory through the participation of 10 actors in five sessions for a period of 12 hours, of these sessions 10,039 articulations were obtained divided into 9 emotions (frustration, fear, neutrality, sadness, fear, happiness, anger, surprise, and others).

Candid Images for Facial Expression (**CIFE**) was generated with the aim of analysing real-time facial expression tasks by Li et al. Seven categories of emotions (neutrality, happiness, anger, sadness, disgust, fear, surprise) were selected based on the dataset obtained from the web and social media through web crawling methods. This dataset includes 14,756 images in addition to a number of images added manually to balance the dataset [2].

Cohn–Kanade plus (**CK+**) was collected through the participation of 123 people in creating 593 videos to create images with a resolution of 640 x 480 pixels and divided into seven emotional expressions (sad, anger, contempt, disgust, fear, happiness, surprise).

Japanese Female Facial Expression (**JAFFE**) Created with images of ten actresses to represent seven expressions (neutrality, fear, disgust, anger, sadness, happiness, surprise), this dataset contains 213 facial images with a resolution of 256 x 256 pixels.

Real-World Affective Face Database (**RAF-DB**) are consisting of 29,672 facial images of people of different races and ages, in addition to unbalanced lighting conditions, to compose seven different facial expressions (sad, joy, fear, anger, surprise, disgust, neutrality) [11].

Radboud Faces Database (**RaFD**) contains 67 models, and each model has 120 images, where the total dataset is 8040 images for eight facial expressions (neutrality, fear, anger, sadness, contempt, surprise, happiness, disgust). The characters (females, males, children) performed the required expressions under certain conditions, as they were in black uniform clothes, in addition to the condition of not having facial hair and not wearing jewelry.

Affective-MIT Facial Expression Dataset (**AM-FED+**) It is data that expresses real situations obtained in the wild and makes up 1,044 video clips of the face. Then, using the facial procedures coding system (FACS), a manual explanation was added for all frames, in addition to providing 34 locations for the facial features that were automatically detected.

Static Facial Expressions in the Wild (**SFEW**) it is a Dataset derived from a selection of fixed frames of facial expressions found in the AFEW database. This data set contains 1766 samples divided into three main groups (testing, verification, training) 372, 436, and 958 respectively, and the expressions are eight (happiness, sadness, fear, disgust, contempt, anger, neutrality, surprise) [40].

The **Karolinska Directed Emotional Faces** (**KDEF**) is a dataset is relatively balanced and contains 4900 images of the seven facial expressions (joy, sadness, anger, fear, surprise, neutrality, disgust), and the training and validation data are randomly selected.

Facial Expression Recognition 2013 (**FER2013**) Dataset is considered one of the most widely used wild datasets among researchers, as it contains 35,887 facial images with a size of 48 x 48 and for the seven expressions (anger, disgust, happiness, fear, sadness, surprise, neutrality), the distribution of expressions is unbalanced where the expression of disgust is less than 600 samples, while the rest of the expressions are distributed over 5000 images [41].

Keimyung University Facial Expression of Drivers (**KMU-FED**) database [42] represents 1600 and 1200 pixel images captured in a real driving environment, i.e. inside the vehicle, by a NIR camera mounted on the dashboard or steering wheel, it contains 1,101 images and six basic expressions. Also, it is useful for simulating systems designed to monitor the safety of drivers.

The **Ryerson Audio-Visual Database of Emotional Speech and Song** (**RAVDESS**) dataset contains emotional recordings (7356) which are divided into three modes (audio, video and audio channels, and finally video), where one actor represented eight emotions (neutral, sadness, joy, anger, fear, surprise, calm, disgust) [13].

Labelled Faces in the Wild (**LFW**) database contains 13,233 images of the faces of 5,749 people, measuring 250 x 250. This data was published in 2007 and then updated in 2014. It is the first large wild dataset to be used in the field of facial verification as a reference standard.

The **Yale Face Database B** Created by photographing 28 positions and human subjects under 64 different lighting conditions, it contains 16,128 images to represent the six basic emotional expressions.

The **Google facial expression comparison** (**GFEC**) dataset composed of a wide range of facial expressions (love, neutrality, pain, pride, realization, comfort, sadness, shame, surprise, sympathy, victory, amusement, anger, dread, boredom, focus, elation, embarrassment, fear, interest, disappointment , disgust, distress, doubt, ecstasy, confusion, contemplation, contempt, contentment, desire) are derived from Flickr.com and used for the purpose of biometrics analysis [32].

SJTU Emotion EEG Dataset (**SEED**) consists of electroencephalogram (EEG) signals with eye movement signals, which were elicited from participants' reactions to watching a Chinese movie (15 clips) simulating the three expressions (sadness, neutrality, happiness), and this data set was developed by Zheng and LU.

SJTU Emotion EEG Dataset for Four Emotions (**SEED-IV**) is complementary to the previous one from the SEED dataset with a difference in the number of feelings, the number of participants, and the number of stimulating film clips, which simulate four feelings (sadness, fear, neutrality, happiness) and the number of participants is 15 people who watched 72 movie clips in three sessions.

SJTU Emotion EEG Dataset for Five Emotions (**SEED-V**) As in SEED-IV, this dataset is considered complementary to the previous SEED dataset with a difference in the number of feelings, the number of participants, and the number of stimulating film clips, as it simulates five feelings (sadness, fear, neutrality, happiness, disgust) and the number of participants is 16 people who watched 15 clips Film equally divided by the number of feelings in three sessions.

Dataset for Emotion Analysis using EEG, Physiological and video signals (**DEAP)** depend on the recorded reactions of 32 participants while watching a number of musical pieces, and the reactions depended on the Peripheral Physiological Signals (PPS) with the EEG signals of the participants. The reactions reflected a number of emotions that were simulated in this dataset (familiarity, hatred, arousal, dominance, liking, valence).

In the **DREAMER** dataset, EEG and ECG signals are acquired during audio-visual effect elicitation. Participants were recorded 23 cues and subjective ratings as they watched eighteen films to elicit a number of different emotions (valence, arousal, dominance) [43].

A **Facial response Database including local Facial expression and Bio-signal responses** (**FDFB**): Researchers relied on several techniques to obtain the dataset, where these techniques used facial expressions by a camera, electrical skin activity (EDA), photogrammetry on the earlobe (PPG), electroencephalography (EEG) to record the emotional response that included the participants' facial expressions and their vitality signals. Then, by watching the participants watch 32 video clips, the expressions of the four emotions (fear, happiness, frustration, calm) were simulated [44].

The **MMI Facial Expression** has more than 2,900 high-resolution images and videos on 75 different subjects. The presence of AUs in the videos is fully annotated (event coding) and partially encoded at the frame level, showing for each frame whether the AU is in the neutral phase, start, apex, or offset. The six basic emotions help to explain a small part of the audio-visual laughter [23].

UTKFace dataset consists of a very large group of facial images of up to 20,000 thousand images with a set of comments that are used to indicate race, age (0 to 116 years) and gender [45].

Bahçeşehir University Multimodal Affective Database-1 (**BAUM-1s**) dataset consists of audio-visual snippets of acted and spontaneous (re-acted) facial emotive reactions. The audio-visual clips were recorded from 31 participants who spontaneously exhibit a variety of emotional and mental states in Turkish. The database includes synchronized face recordings of people captured by a frontal stereo camera and a half profile mono camera. It also contained 13 emotions (disgust, fear, surprise, boredom, disdain, confusion, hesitation, reflection, focus, interest (including curiosity), happiness, anger, sadness, and annoyance (including complaint).

Ryerson Multimedia Lab (**RML)** dataset collected 720 audio-visual examples of how people show how they feel in the RML emotion database. People can show anger, disgust, fear, happiness, sadness, surprise, and surprise. The samples were recorded with a digital video camera in a quiet, bright room with a simple background. The people who took part in the experiment were given a list of emotional sentences and told to say how they felt as naturally as possible by remembering an emotional event from their own lives. There was a total of ten different sentences for each type of emotion [46].

The **Multimedia Understanding Group** (**MUG**) dataset includes 1462 RGB image sequences of 86 persons, ranging in length from 50 to 160 images. There are just six emotions, five positive and one negative [47].

The **CMU Multi-PIE face** database was collected (305 GB of facial expressions) during five months with the participation of 337 people. The participants' faces were photographed in 15 positions and 19 lighting conditions to produce 750,000 images [48].

Federation Equestrian International (**FEI)** dataset was created and assembled in the Artificial Intelligence Laboratory in the Federation Equestrian International / Brazil 2005-2006, where 200 people participated (100 males and 100 females) and their ages ranged from (19-40 years) and 14 photos were taken for each person, bringing the total photos to (2800 facial photos) (Neutral, Smile).

Indian Movie Face Database (**IMFDB**): 34,512 images (various in expression, age, makeup and posture) of 100 actors were collected from 100 Hindi film clips.

Taiwanese Facial Expression Image Database (**TFEID**) consists of 7200 pictures in total, depicting 40 different persons, eliciting 8 distinct feelings (6 primary, neutral, contempt). Every emotion is depicted here from two perspectives (0 and 45). There were a total of 4800 photographs in the custom dataset, 3400 of which were candid shots of 35 persons displaying a range of emotions, and 1400 of which were staged portraits [49].

UMD Faces Dataset consists of still images (367,888 facial images of 8,277 subjects) with captions for the purpose of identifying the face. In addition to more than 22,000 videos on 3,100 topics to form 3.7 million video frames. UMD dataset collected using popular search engines like google, yahoo, etc.

CelebA is a dataset includes all the world's famous people, totaling 202,599 RGB photos. Moreover, face recognition, landmark recognition, and attribute recognition were all objectives in the development of this dataset [50].

Multimodal EmotionLines Dataset (**MELD)** Dataset includes facial photos and character sounds. Face and voice come from the video frame. Five TV seasons' worth of facial and voice data were used. MELD comprises seasons S01E03, S04E04, S05E05, S07E07, and S10E15. MELD has 120,000 face photos for six characters. Each character has 27,000 faces. The spoken fragment relates to facial photos [51].

Facial Expression Recognition Plus (**FER+**) to tackle the problem of noisy labels in FER2013, the photos were re-labelled, and a probability distribution was utilized instead of a single tag to determine the category of each image. Images and three sets are same in FER2013, whereas tags are else.

CASMA Dataset: Expressions in CASME can be either tense, happy, repressed, surprised, disgusted, fearful, contemptuous, or sorrowful. Overall, the dataset contains 189 films across 19 subjects. Tension (69 videos), joy (9 videos), suppression (38 videos), surprise (20 videos), disgust (44 videos), fear (2 videos), contempt (1 video), and depression (6 videos).

CASME II Dataset: Has seven different emotions: joy, disgust, suppression, surprise, fear, and sadness. The collection includes 26 subjects, 255 videos, and 16781 frames. There are 32 videos (2,319 frames) of joy, 99 videos (63,36 frames) of neutral expressions, 63 videos (4,153 frames) of disgust, 27 videos (2,150 frames) of suppression, 25 videos (1,514) of surprise, 2 videos (66 frames) of fear, and 2 videos (15,14) of grief (7 videos, 243 frames).

SAMM (Spontaneous Actions and Micro-Movements): The purpose of this dataset was to fill in the gaps left by previous efforts to study micro-expressions, such as a lack of demographic variety, ground truth Facial Action Coding System (FACS) coding, and spontaneous movements. Through an emotional inducement experiment, collected 159 micro-facial movements that were completely natural and unprompted. Thirty-two people's micro-movements were captured for this dataset, and they came from a wide range of backgrounds: 13 different nationalities, ages 19 to 57 (with a mean of 33.24 years and a standard deviation of 11.32), and an even split of 17 men and 16 women [52].

The AffectNet Dataset: There are 287651 training photos and 4000 validation photos in the dataset, all of which have been manually annotated. Researchers analyze their methods using the validation set, which consists of 500 photos for each of the following 8 emotion classes: anger, contempt, disgust, fear, happiness, neutrality, sorrow, and surprise (the test set is not publicly available)[53].

5. RELATED WORKS

This section includes relevant previous studies in the field of identifying emotions for the last four years (2019–2022). Therefore, the articles that are related to the technology of emotion recognition, which is mainly divided into vision-based techniques (camera) and bio-signals and physiological techniques, were addressed as elaborated in Table II. Since the scope of study in this systematic review is the detection of emotions in real time (Fig. 4), especially through facial expressions, the detailed study will be limited to related studies from Table II.

Sharma et al. [2], relied on multimodal datasets (visual, audio, and text) to be used in a model (EmoHD) based on delayed dataset fusion, transfer learning (TL), and deep learning (DL) to detect emotions identified in the MELD and IEMOCAP databases (disgust, happiness, fear, anger, neutrality, surprise, and sadness). Regarding the visual dataset, ResNet was used to adjust the model, where the model was trained using 75% of the datasets (MELD and IEMOCAP) and the rest for evaluating the model's performance. For the purpose of evaluating the model, the F1-score was used. The average accuracy of the model was 65% in IEMOCAP and 61% in MELD. The weakest performance in the model was the detection of disgust emotions (54% and 58%) in MELD and IEMOCAP, respectively. While the highest performance was the detection of happiness (71%) in IEMOCAP and sadness (66% in MELD), According to researchers, the proposed approach can be utilized for real-time video surveillance of patients to detect their unstable emotions automatically.

Gill and Singh [32] used facial datasets (LFW, Yale Face B, Google FEC) to train and evaluate a model based on a convolutional neural network (CNN) for the purpose of identifying the six emotions (happiness, sadness, anger, disgust, surprise, neutrality), and the Dlib technique to detect 68 landmarks in facial images. The model achieved an accuracy of 93% with an actual runtime of 656 milliseconds. The weakest performance of the model was when detecting the emotion of disgust (86%), while the highest performance was when detecting the emotion of sadness (95%).

B S and Rao [54] designed a Binary Neural Network (BNN) that is fed with pre-processed data by the Viola-Jones (VJ) algorithm, followed by Local Binary Pattern (LBP), which with Linear Fully Connected (LFC) classifies facial images into six emotions (Anger, Disgust, Fear, Happy, Sad, and Surprise), and then implemented on the Field Programmable Gate Array (FPGA) board. After training the LBP-BNN with the FER2013 dataset and through a webcam connected to a PYNQ-Z1 (which can process 3906.25 images per second), the system was set up to recognize emotions in real-time. In real time, the FPGA achieved an overall accuracy of 75%, which is comparable to the FPGA's results on the JAFFE dataset. Overall, according to the highest and lowest performance, the model classifying the emotion of happy to 89%, and disgust to 60%. According to the researchers the proposed technology could provide a solution to passenger safety issues in a globalized, mobile environment.

Saxena et al [55], used deep learning technology to deliver a smart network capable of recognizing emotions from multiple facial expressions in real time based on a convolutional neural network (CNN) on a Raspberry Pi3 B+ (Execution time 200-350 milliseconds) circuit board. The performance of the proposed model in the study was high, achieving an average accuracy of 95.8% in recognizing emotions (surprise: 98%, joy: 96%, anger: 95%, neutrality: 98%, disgust: 92%) on CMU Multi PIE dataset. The system failed to recognize the face when dealing with complex images. In addition to mixing emotions, for example, neutral was recognized as disgust and vice versa.

Gowri et al [56], introduced a system based on a 4-layer Convolutional Neural Network (CNN) with Rectified Linear Unit (ReLU) that detects seven emotions (happiness, sadness, anger, fear, surprise, disgust and hate) by means of facial expressions in real time. The processed dataset (FER2013) was pre-processed by a number of filters (Haar cascade, Harris Corner Detector) and the problem of overfitting was reduced by increasing the dataset during the training phase. The model achieved an accuracy of 73%. The researchers suggest using their model after developing it in a number of applications, such as business promotion and others.

The real world was simulated in **Pandey et al** [57] in order to identify human emotions in a complex environment in real time through a CNN model. A webcam was relied on for the purpose of identifying emotions by detecting the face and then classifying the emotions. Also, the transfer learning technique was used for the purpose of verifying the accuracy, which was 76.62% after the proposed model (derived from VGG-16) was trained on the FER2013 dataset. For the purpose of improving performance, (ReLU) has been replaced by ELU as an activation function to avoid speed up training and gradient dispersion, while working to bypass overfitting through the Dropout layer. The study deals with the four emotions (happiness, surprise, anger, and sadness) without using two emotions (fear, disgust), as only 24,282 images from FER2013 were used.

Hussain et al [58]**,** aimed at discovering the face, then recognizing the faces, and finally categorizing the emotions of facial expressions, which are divided into six emotions (happy, neutral, angry, sad, disgusted, and surprised). The researchers used Haar Cascade with the Open CV library to detect a face through a real-time camera and store facial features in a database for later identification. Then in the second step, a CNN model (VGG 16) trained by the KDEF dataset was used for facial recognition and by matching the detected faces in the database. In the final stage, the recognizable facial expressions were rated to one of the six emotions mentioned above in real time, and the performance measures were 88% correct. The recognition of the emotion of fear was not considered in this study, in addition to the lack of recognition of multiple faces, although it was discovered because the faces were not trained on the used database, with writing the emotions classified on top of each other when working with more than one face in real time.

Badhe and Chaudhari [59]**,** present a CNN model based on four convolution layers and two FC layers with training the model using the FER2013 dataset for the purpose of dynamically recognizing human emotions through facial expressions. The model achieved 94% as accuracy in identifying human emotions (happiness, sadness, anger, neutrality, surprise) and failed to perform well when trying to identify the emotions of fear and disgust.

Rathour et al [60]**,** introduced a facial recognition and emotion detection system based on the Internet of Medical Things (IoMT) using deep convolutional neural networks in real-time. The system was implemented through a Raspberry-Pi board and a Pi camera with modem and Wi-Fi for the purpose of storing datasets in the cloud. Mini_Xception model was used for training purposes (FER 2013 dataset) which is lighter than traditional models as it does not require convolution across all channels. An accuracy of 69% was achieved after training the network via Google Co-Lab in batches using the Adam optimizer. Recognizing emotions in real-time through facial expressions achieved the best performance with an accuracy of 73%.

Lu et al [61]**,** introduced an Emotion, Age, and Gender Recognition system called EAGR. In the domain of emotion, they considered real-time recognition of the seven emotions (the six baseline plus neutral) through a webcam using a CNN deep learning model. Initially, for the purpose of face detection, NFC pre-processing was applied, in addition, by rotating the images, the data was increased. As a result of the application of NFC pre-processing technology, the performance is more efficient and the model execution time is reduced, which leads to a reduction in the training time. In order to test the system in real time, five participants (Three men, two women) with an execution time of five minutes for each participant were used to simulate the seven consecutive facial expressions (neutrality, disgust, happiness, sadness, surprise, anger, fear) in front of the camera. The average accuracy of real-time emotion recognition is 72.16%.

Shruti and Nandi [49]**,** built a model that aims to classify human emotions (happiness, surprise, sadness, fear, anger, neutrality, disgust) in real time while reducing the computational complexity of the convolutional neural network by reducing parameters (146,000 parameters were used), and then test the power of the model on Eight different emotion datasets for images of faces from different backgrounds and age groups. The presented convolutional neural network consisted of 20 layers of convolution and assembly with ReLU and was trained for 8 hours to get a training accuracy of 65%, then the model was validated and achieved an accuracy of 74%. This improved resolution is achieved through a number of measures (filter scaling, changing the number of convolutional layers, reducing the image size in the late network) to enable the network to extract the best features across the deep layer.

Pathar et al [62]**,** built a model capable of detecting the human face and then classifying emotions based on facial expressions into seven categories in real-time (basic plus neutral) by using convolutional neural networks trained on the gray facial image dataset FER2013. Several measures were adopted to improve the performance of the model by experimenting with different depths and max pooling layer, in addition to using dropout technique to bypass overfitting. The presented model consists of eight convolutional layers, four pooling layers, and three fully connected layers. After completing the training of the model, it was implemented on a webcam and tested on a number of faces, where it succeeded in detecting multiple faces and then classifying the facial expressions into the appropriate emotions, where the model achieved a performance with an accuracy of 89.98% using the swish activation function (ReLU) in the fully connected layer. The model failed to classify some emotions, as there was confusion that led to the classification of sadness as neutral or fearful. In addition to the negative impact of the CPU, which led to limited detection of deeper layers in the convolutional neural network.

Ozdemir et al [63]**,** based on the Convolutional Neural Network (CNN), a low-cost LeNet architecture was introduced for the purpose of recognizing the seven human emotions through real-time facial expressions. For training purposes, the JAFFE datasets, KDEF, were combined with a dataset created by the researchers. This combination between the datasets led to obtaining a higher accuracy in training 96.43% and verification 91.81%. Then by using Keras and TensorFlow libraries the model was trained. The structure of the proposed model consists of two layers for each of the convolutional and max pooling layers with one fully connected layer, where the model limits the learning to the pixel values of face detection in the rectangular region (using Haar Cascade technique) to ensure the speed of queries with the deep artificial neural network model, which reducing the training time and number of networks. When the model is tested in real time with a camera, human faces are detected at a rate of 30 images per second, and then by the query function that executes for each image per second, the type of emotion is displayed above the frame. The performance of the model was less accurate when predicting the emotion of sadness and more accurate with emotions (fear, surprise, neutrality).

Cheng Liu et al [64]**,** proposed a model based on a convolutional neural network (CNN) to get rid of the problems arising from changes in image features while the camera captures images in real time at high speed by taking these changes into account while building the model. The FER2013 dataset was used to train the model, which consists of 1 input layer, 3 convolution and pooling layers, and fully connected layers. After a sufficient number of training sessions, the parameters generated from the training process were kept for the purpose of being used in the real-time emotion recognition step. In the first step of the implementation and in order to detect the face, a webcam and parameters obtained from the training were used, then the captured images were converted to grayscale with resizing, and then the captured images were fed into the trained framework, which leads to a detection rate of 10 images per second. After applying the proposed method, the errors in real-time emotion recognition were minimized due to the weights obtained from CNN and used in calculating the average of the estimation results for the current frame and the previous frame. The researchers did not address the performance of the model in terms of accuracy.

6. DISCUSSION

There are a number of factors that affect the recognition of the basic emotions of humans through facial expressions in terms of the accuracy of the results, and these factors include:

- Training algorithm.
- The type of dataset.
- Number of categories and photos.
- The form used structure.

For example, the type of dataset according to the dataset table (TABLE I) in Section K is divided into a dataset obtained in a controlled environment, such as a laboratory, and another dataset that represents realistic images, such as those taken from the web. The first set represents a dataset with good accuracy in terms of facial clarity and simulating the expressions of different emotions well and with focus, while the second set is considered unclear, and facial expressions do not reflect emotions well. On the other hand, when both groups are used in a particular training algorithm, the results will be more accurate because the algorithm has learned how to distinguish faces from different images and distinguish emotions based on different facial expressions. In addition to ensuring that the problem of overfitting is overcome, which reflects good results in the training phase but fails in the testing and verification phase, In addition, the imbalance of the used datasets can affect the accuracy of the results negatively, as the number of images for one of the expressions is less than for other expressions, and this can be seen in the difficulty of predicting the emotion of disgust in the CK+ dataset.

As for the techniques used to detect emotions, the results from CNN have shown superiority over traditional methods. Where the traditional methods represented an obstacle to increasing the performance of the system because the process of extracting features and classifying each of them is done independently, deep learning networks overcome this problem through the possibility of learning from start to finish. On the other hand, there is a correlation between the techniques for teaching deep learning networks and the size of the dataset. The larger the dataset, the better the performance of the model, at the expense of the computational cost. Transfer learning techniques were used to solve this problem. This saved time and improved performance by using an existing model instead of starting from scratch or with a lot of data.

One of the most important steps in recognizing emotions through facial expressions is to detect the face in the images of the used dataset. Most of the studies relied on the Viola-Jones technique based on the Haar cascade to identify faces.

Related studies that aimed to detect emotions in real-time based on deep learning techniques achieved different results depending on the used dataset or the circumstances surrounding the implementation of the systems. [55] presented a system that achieved an average accuracy of 95.8% in emotion recognition with an execution time of 350 milliseconds, while the performance accuracy of the system presented by[32] was 93% with an execution time of 656 milliseconds. Although a different dataset was used in both studies, they both used the convolutional neural network technique, and their models achieved similar results in terms of performance, with the first model outperforming the second in terms of execution time. As for the studies[56] [60] [61] that used the FER2013 dataset, they achieved an average performance accuracy of close to 73%, based on deep learning technology. The rest of the studies did not explicitly report the average accuracy achieved when trying the system in real-time. From the above, the best performance was achieved in the system presented by[55] , taking into consideration the confusion of the system when dealing with complex images, which can be addressed by training the model on a complex dataset or merging more than one dataset together.

7. CHALLENGES

In this study, some challenges related to the methods of detecting human emotions based on deep learning techniques were noted, including:

- There is a diversity of face shapes among people with different biometric shapes, which leads to a difference in the accuracy of facial expressions for the same emotion between different people.
- Some expressions of emotions are similar, such as difficulty distinguishing between fear and pain.

8. CONCLUSIONS

Despite the development in the field of computer vision, especially deep learning techniques, detecting emotions through facial expressions remains a challenge for the pioneers in this field. Human emotions that reflect the psychological state of a person can be predicted and known automatically in humans based on a number of expressions that reflect the emotion of sadness, happiness, etc. Pioneers in artificial intelligence (AI) are trying to make a deep learning system called a convolutional neural network (CNN) that can automatically recognize how people feel.

The study aimed to limit all the information and terms related to the subject of the research in this systematic review in order to be a useful reference for researchers in the field of identifying emotions through facial expressions. The literature review from 2019 to the present time related to techniques for detecting emotions using deep learning techniques based on facial expressions has been systematically reviewed. In addition, the data sets used in training, testing, or verifying the models were collected from the reviewed literature, and a comparison was made between them in terms of the number and type of elements and the environment of data collection, whether it was in the wild or laboratory. Concerning the methods for recognizing emotions, they compared and studied how to use their systems, with a focus on how to recognize emotions in real time.

In conclusion, the study concluded that the dataset used mainly affects the performance of the model, as the use of the dataset obtained in the laboratory, which is customized in a way that makes face detection and emotion prediction easy and direct, can lead to incomplete learning of the algorithm. When this kind of dataset is used to train a model, it has trouble recognizing faces, predicting emotions in complex images, and recognizing emotions in real time.

Author's Note

Abstract version of this paper was presented at 10th International Conference on Advanced Technologies (ICAT'22), 25-27 November 2022, Van, Turkey.

References

[1] B. G. K. Reddy, P. Yashwanthsaai, A. R. Raja, A. Jagarlamudi, N. Leeladhar, and T. T. Kumar, "Emotion Recognition Based on Convolutional Neural Network (CNN)," in *International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation*, 2021, pp. 1–5, doi: 10.1109/ICAECA52838.2021.9675688.

[2] A. Sharma, K. Sharma, and A. Kumar, "Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion," *Neural Comput. Appl.*, vol. 2, 2022.

[3] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput. Appl.*, vol. 8, 2021, doi: 10.1007/s00521-021-06012-8.

[4] K. Wang, Y. Ho, Y. Huang, and W. Fang, "Design of Intelligent EEG System for Human Emotion Recognition with Convolutional Neural Network," in *IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2019, pp. 142–145.

[5] W. Liu, J. Qiu, W. Zheng, and B. Lu, "Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 2, pp. 715–729, 2022.

[6] L. Sandra, Y. Heryadi, Lukas, W. Suparta, and A. Wibowo, "Deep Learning Based Facial Emotion Recognition using Multiple Layers Model," in *International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation*, 2021, pp. 137–142, doi: 10.1109/ICAMIMIA54022.2021.9809908.

[7] A. Landowska *et al.*, "Automatic Emotion Recognition in Children with Autism: A Systematic Literature Review," *Sensors*, vol. 22, no. 4, pp. 1–30, 2022, doi: 10.3390/s22041649.

[8] S. Al-asbaily and K. Bozed, "Facial Emotion Recognition Based on Deep Learning," in IEEE 2nd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA), 2022, vol. 22, no. 16, pp. 534–538, doi: 10.3390/s22166105.

[9] S. Kaur and N. Kulkarni, "A Deep Learning Technique for Emotion Recognition Using Face and Voice Features," in *IEEE Pune Section International Conference, PuneCon 2021*, 2021, pp. 1–6, doi: 10.1109/PuneCon52575.2021.9686510.

[10] M. U. Khan, M. A. Abbasi, Z. Saeed, M. Asif, A. Raza, and U. Urooj, "Deep learning based Intelligent Emotion Recognition and Classification System," in *Proceedings - International Conference on Frontiers of Information Technology, FIT*, 2021, pp. 25–30, doi: 10.1109/FIT53504.2021.00015.

[11] A. I. Siam, N. F. Soliman, A. D. Algarni, F. E. Abd El-Samie, and A. Sedik, "Deploying Machine Learning Techniques for Human Emotion Detection," *Comput. Intell. Neurosci.*, 2022, doi: 10.1155/2022/8032673.

[12] S. Palaniswamy and Suchitra, "A Robust Pose Illumination Invariant Emotion Recognition from Facial Images using Deep Learning for Human-Machine Interface," 2019, doi: 10.1109/CSITSS47250.2019.9031055.

[13] S. Yuvaraj, J. V. Franklin, V. S. Prakash, and A. Anandaraj, "An Adaptive Deep Belief Feature Learning Model for Cognitive Emotion Recognition," *8th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2022*, vol. 1, pp. 1844–1848, 2022, doi: 10.1109/ICACCS54159.2022.9785267.

[14] G. Chartrand *et al.*, "Deep learning: A primer for radiologists," *Radiographics*, vol. 37, no. 7, pp. 2113–2131, 2017, doi: 10.1148/rg.2017170077.

[15] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00444-8.

[16] L. Arnold, S. Rebecchi, S. Chevallier, and H. Paugam-Moisy, "An introduction to deep learning," *Eur. Symp. Artif. Neural Networks*, pp. 477–488, 2021, doi: 10.1201/9780429096280-14.

[17] D. Canedo and A. J. R. Neves, "Facial expression recognition using computer vision: A systematic review," *Appl. Sci.*, vol. 9, no. 21, pp. 1–31, 2019, doi: 10.3390/app9214678.

[18] A. Saravanan, G. Perichetla, and D. K.S.Gayathri, "Facial Emotion Recognition using Convolutional Neural Networks," *arXiv*, vol. 1, 2019.

[19] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *International Conference on Engineering and Technology, ICET*, 2018, pp. 1–6, doi: 10.1109/ICEngTechnol.2017.8308186.

[20] X. Sun, P. Xia, L. Zhang, and L. Shao, "A ROI-guided deep architecture for robust facial expressions recognition," *Inf. Sci. (Ny).*, vol. 522, pp. 35–48, 2020, doi: 10.1016/j.ins.2020.02.047.

[21] R. Chauhan, K. K. Ghanshala, and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," in *International Conference on Secure Cyber Computing and Communications*, 2018, pp. 278–282, doi: 10.1109/ICSCCC.2018.8703316.

[22] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: A case study on FER-2013," *Smart Innov. Syst. Technol.*, vol. 85, pp. 1–16, 2018, doi: 10.1007/978-3-319-66790-4_1.

[23] S. H. Lee, "Facial data visualization for improved deep learning based emotion recognition," *J. Inf. Sci. Theory Pract.*, vol. 7, no. 2, pp. 32–39, 2019, doi: 10.1633/JISTaP.2019.7.2.3.

[24] A. Y. Nawaf and W. M. Jasim, "Human emotion identification based on features extracted using CNN Human Emotion Identification Based on Features Extracted Using CNN," in *AIP Conference Proceedings*, 2022, vol. 020010.

[25] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," *arXiv*, pp. 1–11, 2015.

[26] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018, doi:

10.1016/j.patcog.2017.10.013.

[27] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning," *arXiv*, pp. 1–20, 2018.

[28] J. T. Barron, "A general and adaptive robust loss function," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 4326–4334, 2019, doi: 10.1109/CVPR.2019.00446.

[29] C. Wan, D. Chen, Z. Huang, and X. Luo, "A wearable head mounted display bio-signals pad system for emotion recognition," *Sensors*, vol. 22, no. 1, pp. 1–20, 2022, doi: 10.3390/s22010142.

[30] H. D. Nguyen, S. H. Kim, G. S. Lee, H. J. Yang, I. S. Na, and S. H. Kim, "Facial Expression Recognition Using a Temporal Ensemble of Multi-Level Convolutional Neural Networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 226–237, 2022, doi: 10.1109/TAFFC.2019.2946540.

[31] M. M. L. Joshi and M. S. Agarwal, "FACIAL EMOTION RECOGNITION USING DEEP LEARNING: A SURVEY," *OORJA*, vol. 19, 2021.

[32] R. Gill and J. Singh, "A Deep Learning Approach for Real Time Facial Emotion Recognition," in *10th International Conference on System Modeling and Advancement in Research Trends, SMART*, 2021, pp. 497–501, doi: 10.1109/SMART52563.2021.9676202.

[33] J. C. Hung, K. C. Lin, and N. X. Lai, "Recognizing learning emotion based on convolutional neural networks and transfer learning," *Appl. Soft Comput. J.*, vol. 84, 2019, doi: 10.1016/j.asoc.2019.105724.

[34] S. Norouzi and M. Ebrahimi, "A Survey on Proposed Methods to Address Adam Optimizer Deficiencies," 2019.

[35] A. Gupta, R. Ramanath, J. Shi, and S. S. Keerthi, "Adam vs. SGD: Closing the generalization gap on image classification," *OPT2021 13th Annu. Work. Optim. Mach. Learn.*, pp. 1–7, 2021.

[36] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," *Sensors (Switzerland)*, vol. 20, no. 2, 2020, doi: 10.3390/s20020342.

[37] R. Zhi, M. Liu, and D. Zhang, "A comprehensive survey on automatic facial action unit analysis," *Vis. Comput.*, vol. 36, no. 5, pp. 1067–1093, 2020, doi: 10.1007/s00371-019-01707-5.

[38] J. Lamoureux and W. Luk, "An overview of low-power techniques for field-programmable gate arrays," in *Proceedings of the NASA/ESA Conference on Adaptive Hardware and Systems, AHS*, 2008, pp. 338–345, doi: 10.1109/AHS.2008.71.

[39] Raspberry pi Foundation, "Raspberry Pi 3 model B+," 2018.

[40] A. B. Rohan and S. Surve, "Deep learning framework for facial emotion recognition using CNN architecture," *J. Phys. Conf. Ser.*, vol. 2236, no. 1, pp. 1777–1784, 2022, doi: 10.1088/1742-6596/2236/1/012004.

[41] G. K. Sahoo, S. K. Das, and P. Singh, "Deep Learning-Based Facial Emotion Recognition for Driver Healthcare," *Natl. Conf. Commun. NCC*, pp. 154–159, 2022, doi: 10.1109/NCC55593.2022.9806751.

[42] M. Jeong and B. C. Ko, "Driver's facial expression recognition in real-time for safe driving," *Sensors (Switzerland)*, vol. 18, no. 12, 2018, doi: 10.3390/s18124270.

[43] W. Liu, J. L. Qiu, W. L. Zheng, and B. L. Lu, "Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition," *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 2, pp. 715–729, 2022, doi: 10.1109/TCDS.2021.3071170.

[44] J. Kwon, J. Ha, D. H. Kim, J. W. Choi, and L. Kim, "Emotion Recognition Using a Glasses-Type Wearable Device via Multi-Channel Facial Responses," *IEEE Access*, vol. 9, pp. 146392–146403, 2021, doi: 10.1109/ACCESS.2021.3121543.

[45] A. Khattak, M. Z. Asghar, M. Ali, and U. Batool, "An efficient deep learning technique for facial emotion recognition," *Multimed. Tools Appl.*, vol. 81, no. 2, pp. 1649–1683, 2022, doi: 10.1007/s11042-021-11298-w.

[46] I. Hina, A. Shaukat, and M. U. Akram, "Multimodal Emotion Recognition using Deep Learning Architectures," in *International Conference on Digital Futures and Transformative Technologies, ICoDT2*, 2022, pp. 2–7, doi: 10.1109/ICoDT255437.2022.9787437.

[47] A. Atanassov and D. Pilev, "Pre-trained Deep Learning Models for Facial Emotions Recognition," 2020, doi: 10.1109/ICAI50593.2020.9311334.

[48] T. Tumakuru, T. Tumakuru, T. Tumakuru, and T. Tumakuru, "Real Time-Employee Emotion Detection system (RtEED) using Machine Learning," in *Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks*, 2021, no. Icicv, pp. 759–763.

[49] S. Jaiswal and G. C. Nandi, "Robust real-time emotion detection system using CNN architecture," *Neural Comput. Appl.*, vol. 32, no. 15, pp. 11253–11262, 2020, doi: 10.1007/s00521-019-04564-4.

[50] Y. Said and M. Barr, "Human emotion recognition based on facial expressions via deep learning on high-resolution images," *Multimed. Tools Appl.*, vol. 80, no. 16, pp. 25241–25253, 2021, doi: 10.1007/s11042-021-10918-9.

[51] D. Liu, Z. Wang, L. Wang, and L. Chen, "Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning," *Front. Neurorobot.*, vol. 15, no. July, pp. 1–13, 2021, doi: 10.3389/fnbot.2021.697634.

[52] S. Miao, H. Xu, Z. Han, and Y. Zhu, "Recognizing facial expressions using a shallow convolutional neural network," *IEEE Access*, vol. 7, pp. 78000–78011, 2019, doi: 10.1109/ACCESS.2019.2921220.

[53] M. I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64827–64836, 2019, doi: 10.1109/ACCESS.2019.2917266.

[54] B. S. Ajay and M. Rao, "Binary neural network based real time emotion detection on an edge computing device to detect passenger anomaly," in *Proceedings of the IEEE International Conference on VLSI Design*, 2021, vol. 2021-Febru, pp. 175–180, doi: 10.1109/VLSID51830.2021.00035.

[55] S. Saxena, S. Tripathi, and T. S. B. Sudarshan, "Deep Dive into Faces: Pose Illumination Invariant Multi-Face Emotion Recognition System," in *IEEE International Conference on Intelligent Robots and Systems*, 2019, pp. 1088–1093, doi: 10.1109/IROS40897.2019.8967874.

[56] S. M. Gowri, A. Rafeeq, and S. Devipriya, "Detection of real-time facial emotions via deep convolution neural network," in *Proceedings - 5th International Conference on Intelligent Computing and Control Systems*, 2021, no. Iciccs, pp. 1033–1037, doi: 10.1109/ICICCS51141.2021.9432242.

[57] S. Pandey, S. Handoo, and Yogesh, "Facial Emotion Recognition using Deep Learning," in *International Mobile and Embedded Technology Conference, MECON*, 2022, pp. 348–352, doi: 10.1109/MECON53876.2022.9752189.

[58] S. A. Hussain and A. Salim Abdallah Al Balushi, "A real time face emotion classification and recognition using deep learning model," *J. Phys. Conf. Ser.*, vol. 1432, no. 1, 2020, doi: 10.1088/1742-6596/1432/1/012087.

[59] S. Badhe and S. Chaudhari, "Deep Learning Based Facial Emotion Recognition System," in *ITM Web of Conferences, ICACC*, 2022, vol. 03058, pp. 1–5, doi: 10.1109/TIPTEKNO50054.2020.9299256.

[60] N. Rathour *et al.*, "Iomt based facial emotion recognition system using deep convolution neural networks," *Electron.*, vol. 10, no. 11, 2021, doi: 10.3390/electronics10111289.

[61] T. Lu, S. Yeh, C. Wang, and M. Wei, "Cost-effective real-time recognition for human emotion-age-gender using deep learning with normalized facial cropping preprocess," *Multimed. Tools Appl.*, pp. 19845–19866, 2021.

[62] R. Pathar, A. Adivarekar, A. Mishra, and A. Deshmukh, "Human Emotion Recognition using Convolutional Neural Network in Real Time," 2019, doi: 10.1109/ICIICT1.2019.8741491.

[63] M. A. Ozdemir, B. Elagoz, A. Alaybeyoglu, R. Sadighzadeh, and A. Akan, "Real time emotion recognition from facial expressions using CNN architecture," *TIPTEKNO - Tip Teknol. Kongresi*, pp. 2–5, 2019, doi: 10.1109/TIPTEKNO.2019.8895215.

[64] K.-C. Liu, C.-C. Hsu, W.-Y. Wang, and H.-H. Chiang, “Real-Time Facial Expression Recognition Based on CNN,” in *International Conference on System Science and Engineering (ICSSE)*, 2019, no. 1, pp. 1–16.

[65] W. Liu, J. Qiu, W. Zheng, and B. Lu, “Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition,” *IEEE Trans. Cogn. Dev. Syst.*, vol. 14, no. 2, pp. 715–729, 2022, doi: 10.1109/TCDS.2021.3071170.

[66] R. Gill and J. Singh, “A Deep Learning Model for Human Emotion Recognition on Small Dataset,” in *International Conference on Emerging Smart Computing and Informatics, ESCI 2022*, 2022, pp. 1–5, doi: 10.1109/ESCI53509.2022.9758261.

[67] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, “Facial emotion recognition using transfer learning in the deep CNN,” *Electron.*, vol. 10, no. 9, 2021, doi: 10.3390/electronics10091036.

[68] M. Bentoumi, M. Daoud, M. Benaouali, and A. Taleb Ahmed, “Improvement of emotion recognition from facial images using deep learning and early stopping cross validation,” *Multimed. Tools Appl.*, vol. 81, no. 21, pp. 29887–29917, 2022, doi: 10.1007/s11042-022-12058-0.

[69] D. Acharya, A. Billimoria, N. Srivastava, S. Goel, and A. Bhardwaj, “Emotion recognition using fourier transform and genetic programming,” *Appl. Acoust.*, vol. 164, p. 107260, 2020, doi: 10.1016/j.apacoust.2020.107260.

[70] A. Poulose, C. S. Reddy, J. H. Kim, and D. S. Han, “Foreground Extraction Based Facial Emotion Recognition Using Deep Learning Xception Model,” in *International Conference on Ubiquitous and Future Networks, ICUFN*, 2021, vol. 2021-Augus, pp. 356–360, doi: 10.1109/ICUFN49451.2021.9528706.

[71] N. Kumari and R. Bhatia, “Efficient facial emotion recognition model using deep convolutional neural network and modified joint trilateral filter,” *Soft Comput.*, pp. 7817–7830, 2022.

[72] A. Vijayvergia and K. Kumar, “Selective shallow models strength integration for emotion detection using GloVe and LSTM,” *Multimed. Tools Appl.*, pp. 28349–28363, 2021.

[73] D. K. Jain, P. Shamsolmoali, and P. Sehdev, “Extended deep neural network for facial emotion recognition,” *Pattern Recognit. Lett.*, vol. 120, pp. 69–74, 2019, doi: 10.1016/j.patrec.2019.01.008.

[74] A. Agrawal and N. Mittal, “Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy,” *Vis. Comput.*, vol. 36, no. 2, pp. 405–412, 2019, doi: 10.1007/s00371-019-01630-9.

[75] E. Ivanova and G. Borzunov, “ScienceDirect ScienceDirect Optimization of machine learning algorithm of emotion recognition Optimization of in machine learning algorithm of emotion recognition terms of human facial expressions in terms of human facial expressions,” *Procedia Comput. Sci.*, vol. 169, no. 2019, pp. 244–248, 2020, doi: 10.1016/j.procs.2020.02.143.

[76] A. Gupta, S. Arunachalam, and R. Balakrishnan, “ScienceDirect ScienceDirect ScienceDirect Deep self-attention network for facial emotion recognition Deep network for facial emotion recognition Arpita self-attention,” *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1527–1534, 2020, doi: 10.1016/j.procs.2020.04.163.

Table II. Related Works

(#CL.EM.: No of classified emotions, R. Time De.?: real-time detection? -: Not mentioned)

Ref.	Year	Dataset	#CL Em*	Face Detection Technique	Model	Platform	Acc.	R.* Tim
[1]	2021	FER2013	7	OpenCV and Haar-cascade	CNN	Python	65.2%	No
[2]	2022	MELD, IEMOCAP	7	MTCNN	ResNet50	Python	(65.88% IEMOCAP), (61.27% MELD)	Yes
[3]	2021	CK+	7	-	Transfer learning approaches	Google Colab	96% for MobileNet	No
[4]	2019	DEAP	-	No Face detection	CNN	-	83.88%	Yes
[6]	2021	FER2013	7	-	ResNet50, DCNN	-	60%	No
[8]	2022	FER2013	7	-	VGG16	Google Colab	89%	No
[9]	2021	RAVDESS, FER2013	7	-	CNN	-	73%	No
[10]	2021	FER2013	7	-	MobileNetV2	Keras with tensor flow on Python	98.7%	No
[11]	2022	CK+, JAFFE, RAF-DB	6	-	MediaPipe face mesh algorithm based on real-time deep learning	-	CK+: 97%	No
[12]	2019	CMU Multi-PIE, JAFFE, CK+ and KDEF	5	-	DPIIER	CUDA	96.55%	No
[13]	2022	RAVDESS	8	-	DBFL	-	98%	No
[20]	2020	CK+, JAFFE, FER2013	7	Viola-Jones	CNN	-	CK+: 94.67% JAFFE: 53.77% FER2013: 40.13%	No
[23]	2019	CK+ and MMI	CK+:6, MMI:7	Landmark points	CNN	Python	CK+:92.63%, MMI:74.11%	No
[32]	2021	LFW, Yale Face B, Google FEC	6	Dlib Facial landmarks	CNN	-	93%	Yes
[33]	2019	FER2013, JAFFE, and KDEF	7	AdaBoost-based	Dens_ FaceLiveNet	-	JAFFE: 90.97 % KDEF: 95.89 % FER2013: 84.59 %	No

[40]	2022	KDEF, RAFD, RAF-DB, SFEW, and AMFED+	-	Viola-Jones	CNN, InceptionResnetV2 and VGG	-	RAF-DB (75%), RAFD (69%), SFEW (41%), AMFED+ (54%), KDEF (100%)	No
[41]	2022	CK+, KDEF, FER2013, and KMU-FED	7	-	SqueezeNet transfer learning	Google Colab	KMU-FED (95.83%), CK+ (91%), KDEF (86%), FER13 (61%)	No
[44]	2021	FDFB	4	No Face detection	SVM with RBF kernel	MATLAB	84.55%	No
[45]	2022	JAFFE, CK+, UTKFace	7	-	CNN	-	95.65%	No
[49]	2020	Fer2013, CK, CK+, Chicago Face, JAFFE, FEI, IMFDB, TFEID, custom dataset builds in laboratory having different angles, faces, backgrounds, and age groups.	7	-	CNN	-	74%	Yes
[50]	2021	UMD, CelebA	7	Linear regression	FS-CNN	Python	95%	No
[51]	2021	MOSI, MELD	6	-	CNN-LSTM	-	MOSI: 87.56, MELD: 90.06%	No
[53]	2020	AffectNet, FER2013, FER+	7	-	CNN models (VGG-face, VGG-f, VGG-13)	-	FER2013: 75.4% FER+: 87.7% AffectNet: 59.5%	No
[54]	2021	FER2013	6	Viola-Jones	LBP, BNN	Python	75%	Yes
[55]	2019	CMU Multi PIE	5	Viola-Jones	DMFERS	Keras with tensor flow on Python	95.8%	Yes
[56]	2021	FER2013	7	Haar cascade	CNN	-	73%	Yes
[57]	2022	FER2013	7	Haar cascade	CNN	Python	76.62%	Yes
[58]	2020	KDEF	7	Haar cascade	CNN (VGG16)	Python	88%	Yes
[59]	2022	FER2013	7	-	CNN	-	94.27%	Yes
[60]	2021	FER2013	7	-	Deep CNN	Google Colab	73%	Yes
[61]	2021	JAFFE, RaFD, CK+, KDEF, MMI,	7	NFC	CNN	-	72.16%	Yes
[62]	2019	FER2013	7	-	CNN	Python	89.98%	Yes
[63]	2019	JAFFE, KDEF, custom dataset	7	Haar cascade	CNN based LeNet	Python	91.81	Yes
[64]	2019	FER2013	7	Haar cascade	CNN	Python	-	Yes
[65]	2022	SEED-V, DREAMER	5	No face detection	DCCA	-	(85% SEED-V), (DREAMER Arousal 89%. Valence 91%. Dominance 91%.)	No
[66]	2022	JAFFE, KDEF	7	Distinct landmarks in the FACE	CNN	TensorFlow	97%	No
[67]	2021	KDEF, JAFFE	7	-	TL on DCNN	Python	KDEF: 96.51%, JAFFE: 99.52%	No
[68]	2022	CK+, JAFFE, KDEF	7	-	CNN by TL, and MLP classifier	Python	CK+:100%, JAFFE: 96%, KDEF: 98%	No
[69]	2020	23 Hindi language film clips	4	No Face detection	Fast Fourier Transform (FFT)	-	89.14%	Yes
[70]	2021	Collected by the researcher	7	-	Deep Learning Model (Xception)	-	97.51%	No
[71]	2022	CK+	7	-	CNN +CLAHE+MJTF	Matlab	98.01%	No
[72]	2020	Twitter	7	-	GloVe and LSTM	Google Colab	86.16%	No
[73]	2019	CK+, JAFFE	7	DNN	Deep CNN	-	CK+: 93.24% JAFFE: 95.23%	No
[74]	2019	FER2013	7	-	CNN	Python	65%	No
[75]	2020	FER2013	7	Viola-Jones	Neural Networks based Google Net	Python	69%	No
[76]	2020	FER2013	7	-	CNN,ResNet and Attention	Python	64.4%	No