IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 28, 2021, accepted November 26, 2021, date of publication November 30, 2021,
date of current version December 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3131733

A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets and Future Directions

CHIRAG DALVI[1], MANISH RATHOD[1], SHRUTI PATIL[2], SHILPA GITE[2], AND KETAN KOTECHA[2]

[1]Symbiosis Centre for Applied Artificial Intelligence (SCAAI), Symbiosis International University (Deemed University), Pune 412115, India
[2]Computer Science and Information Technology Department, Symbiosis Institute of Technology, Symbiosis International University (Deemed University), Pune 412115, India

Corresponding authors: Shruti Patil (shruti.patil@sitpune.edu.in), Shilpa Gite (shilpa.gite@sitpune.edu.in), and Ketan Kotecha (drketankotecha@gmail.com)

This work was supported by the Research Support Fund (RSF) of Symbiosis International (Deemed University), Pune, India.

ABSTRACT Facial expressions are mirrors of human thoughts and feelings. It provides a wealth of social cues to the viewer, including the focus of attention, intention, motivation, and emotion. It is regarded as a potent tool of silent communication. Analysis of these expressions gives a significantly more profound insight into human behavior. AI-based Facial Expression Recognition (FER) has become one of the crucial research topics in recent years, with applications in dynamic analysis, pattern recognition, interpersonal interaction, mental health monitoring, and many more. However, with the global push towards online platforms due to the Covid-19 pandemic, there has been a pressing need to innovate and offer a new FER analysis framework with the increasing visual data generated by videos and photographs.Furthermore, the emotion-wise facial expressions of kids, adults, and senior citizens vary, which must also be considered in the FER research. Lots of research work has been done in this area. However, it lacks a comprehensive overview of the literature that showcases the past work done and provides the aligned future directions. In this paper, the authors have provided a comprehensive evaluation of AI-based FER methodologies, including datasets, feature extraction techniques, algorithms, and the recent breakthroughs with their applications in facial expression identification. To the best of the author's knowledge, this is the only review paper stating all aspects of FER for various age brackets and would significantly impact the research community in the coming years.

INDEX TERMS Facial emotion recognition (FER), feature extraction, machine learning, facial expressions.

I. INTRODUCTION

Human facial expressions that people see visually are all around them. They are natural signals that help them understand emotions from any person in front of them or via images or videos. These emotions are highly complex and challenging to understand for machines but easily understandable by humans. To understand how humans could understand such emotions, Mehrabian, a famous psychologist, found from his research that the emotional data that humans classify as emotions are distributed in sections. He found that only 7% of the emotional data total is passed by language, and 38% is transported by our language auxiliary, which differs from culture to culture, such as the rhythm of speech, tone, pitch, etc. So far, the highest percentage of emotional data shown by facial expression is 55% [1]. This indicates that many sensible emotional data can be obtained by recognizing facial emotions that effectively understand any human's state of mind and actions directly associated with emotions [2]. So, it is essential to explore this research domain in more detail as less accurate systems plague its commercial implementation.

Human facial emotion recognition has been broadly used in numerous human-computer interactions such as smartphones, affective computing, intelligent control systems, psychological, behavioral study, pattern searching, defense, social sites, robotics, and other fields [3]–[5]. By evaluating these emotions, one could deliver maximum user satisfaction

The associate editor coordinating the review of this manuscript and approving it for publication was Rosalia Maglietta.

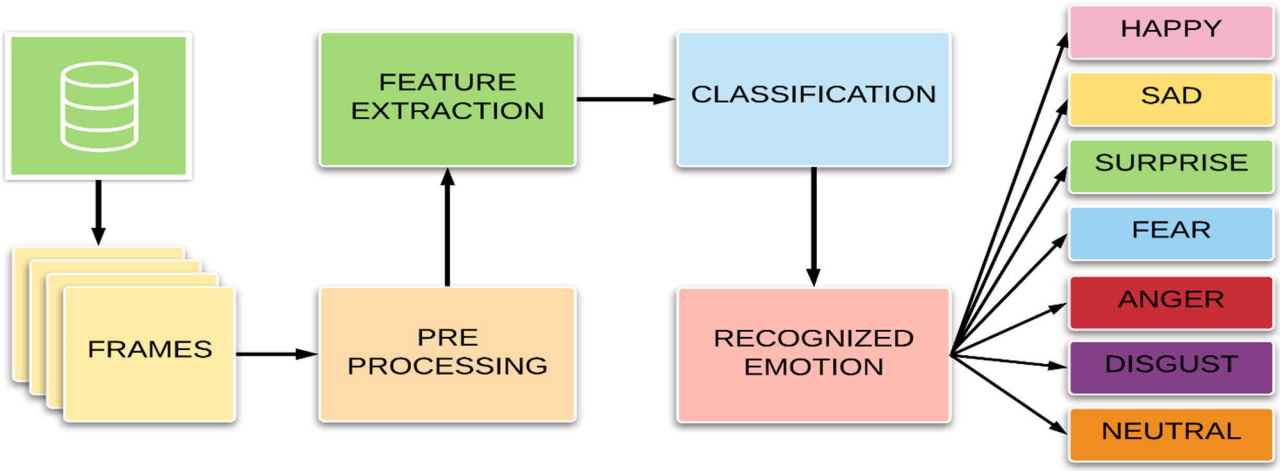

FIGURE 1. Facial emotion classification process.

and feedback to improve current technologies. This can only be done in the domains of computer vision and deep learning. To create several Facial Emotion Recognition (FER) systems that have been evaluated for encoding and transmitting Information from facial representations. In the twentieth century, Ekman and Friesen identified six fundamental emotions based on cross-cultural research that revealed that humans convey these fundamental emotions in the same way regardless of culture [6]. Face expressions include anger, disgust, fear, happiness, sadness, and surprise. Contempt was later added to this list of feelings. To do Facial Emotion Recognition, there are basic initial steps which are divided into three essential stages. The facial features of the face are detected from the entire frame of a video at the first stage, which is a pre-processing stage. The eyebrows, brows, nose, mouth, and chin are among the facial features. More descriptive features from different areas of the Face are removed in the second level. Likewise, more descriptive features from different areas of the Face are removed in the second level. Finally, a classifier is trained using the training data before generating labels for the Emotions, illustrated in Figure 1.

Recent advancement in neuroscience and psychology research has sparked a debate that Newton's three law claimed to be universal is culture-specific and not universal. Because of this, it has raised questions about whether emotions differ based on gender, age, and culture. Today the need for emotion classification has surpassed the barrier of age because of the global shift towards online platforms such as online education to teach or gain knowledge virtually globally to all remote areas, IoT enabled health monitoring systems and temperature setters in cars and households, robotics, psychiatric evaluation based on violent behaviors of criminals or those mentally disturbed, mood swings study on adolescents to help guide them mentally, deepfake detection, gaming, and many such applications are currently being innovated using state of the art technologies.

Also, numerous studies have been conducted on Facial Emotion Recognition by using Computer vision because of its practicality in intelligent robotics, health-related treatment, IoT, Security surveillance, criminal psychological analysis, observation of driver exhaustion, and other human-computer interfaces mechanisms [7]–[9]. With more virtual connectivity through videos and images, the need to adopt the latest technology based on people's emotions is now a critical factor in driving user-friendliness and maximum user satisfaction.

Emotions are nothing but a cognitive state or phase perceived by a human and associated with moods. Usually, these emotions are often twisted with attitude, temper, character, disposition, and motivation. They can also be defined into binary sentiments such as positive(pleasure) or negative (displeasure) under different circumstantial psychological tasks or events. Such emotions bend a person's mind psychologically that the behavior of humans changes over time. Humans handle these emotions by either behavioral response, psychological states triggered by any events or by a person in front of themselves, subjective experience of the situation, and cognitive processes. Humans understand that emotions are not easy to quantify or replicate artificially from this complex set of actions. Many researchers use their version of emotion definitions and assumptions. This makes research in human facial emotions troublesome because all the studies that have been done have significant variance in them and do not draw a generalized conclusion. Although all humans have naturally occurring sets of emotions that can be perceived even cross-culturally, this is also mentioned in the Discrete Emotion Theory, which says that such emotions are distinguishable by an individual's features [10]. Ekman claimed that these emotions are perceived by humans not only culturally but also universally. His proposed model suggested that emotions are categorized into Fear, Happiness, Sad, Surprise, Disgust, and Anger. These categorical emotions are classified using facial and vocal data, which allows them

FIGURE 2. Human emotions.

to perform a better human FER efficiently. Alternatively, there is another proposed model by Plutchik [11], who claims that there are more basic emotions (i.e., joy, fear, anger, sadness, trust, disgust, surprise, and anticipation). To understand it further, figure 2 represents such emotions which are grouped into positive and negative boundaries. Facial Emotions Classification and its study can be done using both unsupervised and supervised methodologies such that it can be multi classified as per Plutchik's model, which is illustrated in figure 3, i.e., wheel of emotion, which shows different ways that they respond to each other along with those that are opposite and can be converted into another Emotion.

Ekman agreed that these illustrated emotions are not unique and cannot be recognized universally. The list of these emotions is then broadened and classified into both facial and vocal expressions.

Different datasets use different combinations of emotions for research. For example, very few kids' datasets have 'angry' emotions. Those using it have recorded by posing. Recording the angry emotion from spontaneous expressions is difficult. But for Adult datasets, it is straightforward to pose for an angry emotion. Datasets like 'RML' have recorded the emotions in a controlled environment with good lighting conditions. Most datasets cut movie clips or tv shows and use them for classification. The category of emotions differs from dataset to dataset.

Along with the category, the number of samples for each emotion also varies from each other. Hence, it is necessary to use a balanced dataset for a good result. Figure 4 and Figure 5, respectively, below show the difference between adult (RML) and kids' (LIRIS) datasets. There is a difference between the category of emotions as well as the recording conditions. RML dataset consists of 8 posed emotions recorded in a controlled environment whereas the LIRIS dataset has 6 spontaneous emotions recorded from a webcam. Recording emotions in a controlled environment gives RML and edge over LIRIS dataset in terms of quality. Apart from dataset quality and emotion category, the facial features also differ

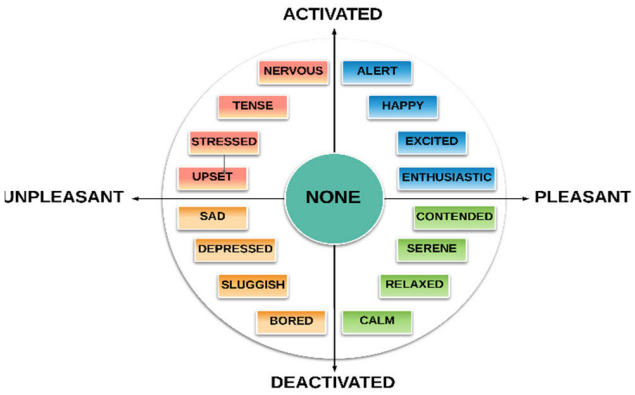

FIGURE 3. Wheel of emotions.

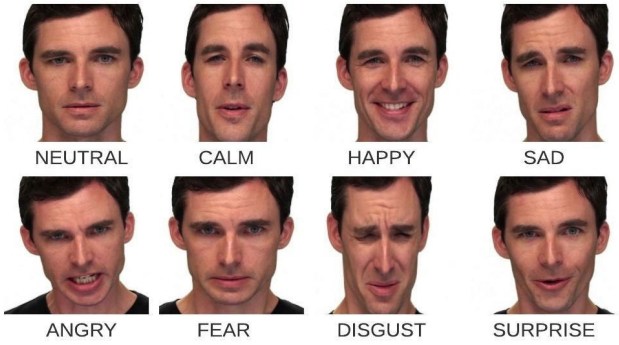

FIGURE 4. Adult emotions (RML dataset)*.

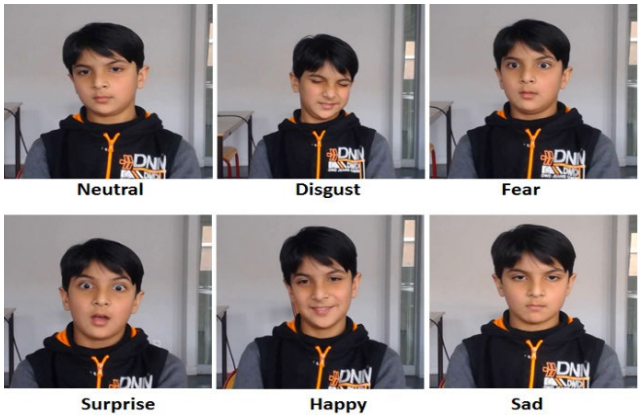

FIGURE 5. Kids' emotions (LIRIS dataset) * (*the emotion labels font vary based on the datasets).

from each other. Apart from dataset quality and emotion category, the facial features also differ from each other.

Apart from Plutchik's model, which depicts the well-known wheel of emotions that classifies emotions, the well-known Circumplex Model of Affects is also illustrated in Figure 3, proposed in a study [12] comparable to Plutchik's model. It is divided into four portions: arousal (activation/deactivation) and valence (pleasant/unpleasant) axes. Every emotion depicted directly results from linear combinations of these two parts of varying degrees of valence and arousal. Four quadrants are created by combining

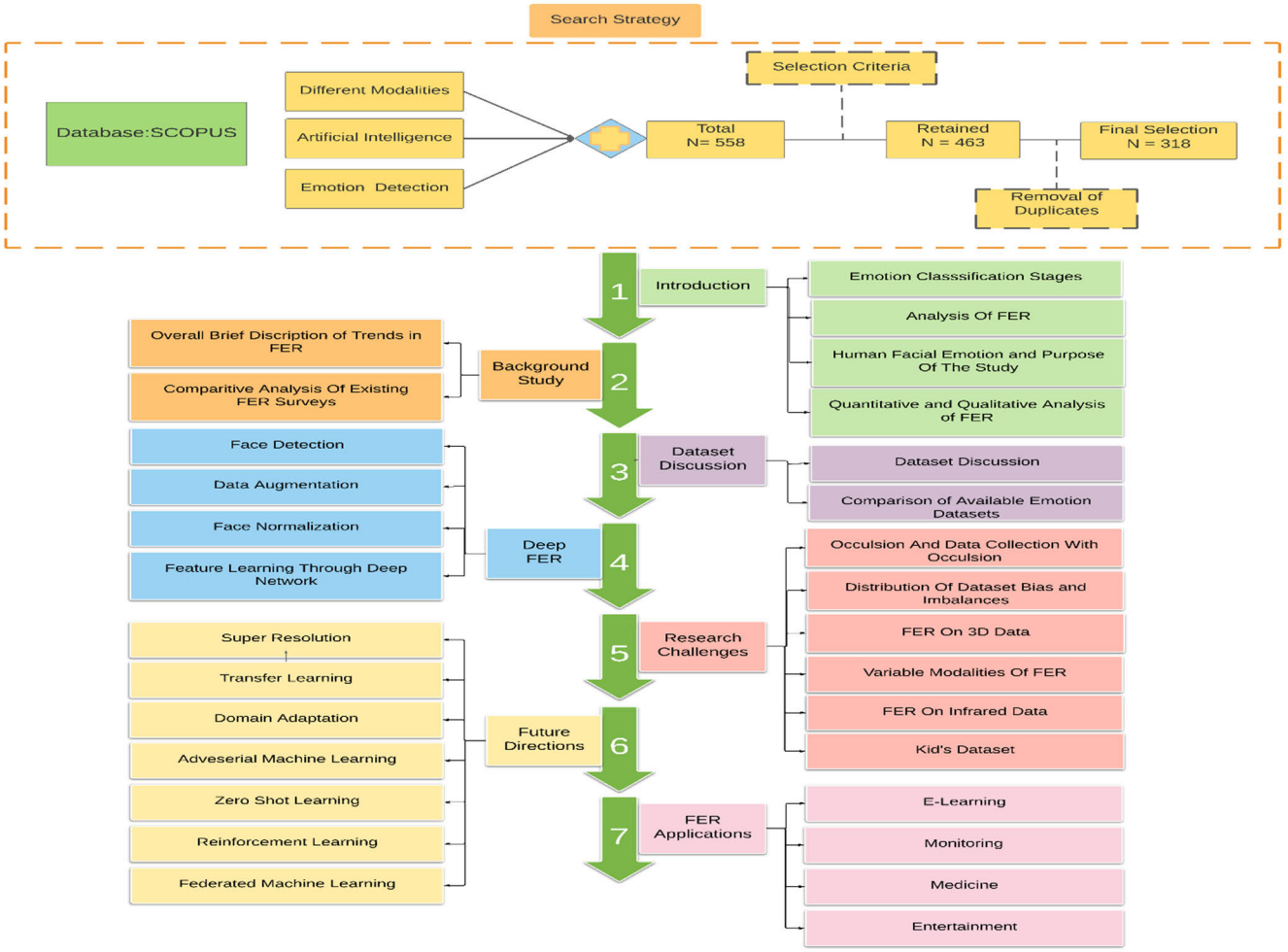

FIGURE 6. **Literature review process along with paper organisation.**

high/low and positive/negative for arousal and valence, respectively.

The purpose of this paper is:

- To review all the research and review done on FER
- To show a comparative analysis of all the research on every category of datasets such as Adults, Kids, and Senior Citizens
- Discuss challenges in FER and plausible suggestions to deal with it
- Discuss Future trends that will impact the field of FER
- To provide a brief idea of the potential real-life applications that can be applied using FER

The paper organization is as follows: Studies reviewed commenced from a brief dataset discussion until Aug 2021. Next, the authors show brief techniques and approaches used on those Datasets. Then there is a brief detailed discussion of all the existing methods/algorithms used and models used in all the publications until 2020. Then, the authors give a critical summary and suggest some future study pathways that will likely enrich the body of knowledge in this study endeavor. This organization is also illustrated in Figure 6 which also shows the Literature Review Process whose deatil is given in Ouantitative Analysis section below which is associated with creation of such paper organisation.

II. QUANTITATIVE ANALYSIS OF FER RESEARCH AND ITS PUBLICATION

Facial Emotion Recognition (FER) is a big part of computer vision case studies. Authors have provided many studies that need to be using a systematic review process to understand research questions. In this section, authors have analyzed FER-based research papers based on crucial keywords which is given in Table 1, which provide quantitative data, geographical parameters, articles, citations, and published papers available on the SCOPUS database. All of the searches were restricted to journal articles and reviews that were written between 2005 and 2021. The English language was implemented in the search. This search approach retrieved a total of 558 documents. After extracting selection criteria was applied in which lecture notes,conferences,Workshops were excluded and filtered down to 463 documents which were retained and then the duplicates were removed further.At the end 318 research papers were chosen and included in the

TABLE 1. Quantitative analysis keywords.

Primary Keywords	Secondary Keywords Using (OR)	Secondary Keywords Using (AND)
"Facial Based Emotion Recognition"	"Artificial intelligence", "video","images", "machine learning,"deep learning"	"Facial based emotion recognition," "facial based emotion analysis"

paper. Each document's publication title, publication year, Journal/Source title, the number of citations are considered for analysis. Thus, the abstract, the title, the keywords, and cited references were retrieved.

The objective of this survey in FER are as follows:

A. EXAMINE THE SUBJECT WISE RESEARCH TREND IN THE AREA RELATED TO FER

As per data, this keyword is used most in Computer Science, followed by Engineering and Medicine. Since FER is related to a computer vision problem, it's evident that research trends related to Keywords are highest in the Computer Science subject. This is clearly illustrated below in Figure 7.

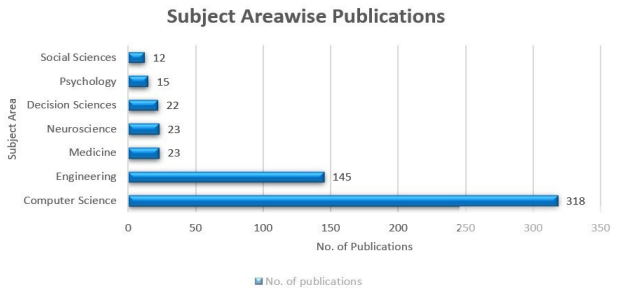

FIGURE 7. Number of publications subject wise source: http://www.scopus.com (assessed on 10th June 2021).

Although there is a huge trend of such publications in Computer Science, it is also observed that Engineering subjects also have a significant observable trend followed by Medicine, Neuroscience, etc.

B. DISTRIBUTION OF THE CONTRIBUTIONS DONE ACROSS VARIOUS TYPES OF PUBLICATIONS

As evident in Figure 7, the authors show the distribution of the publications across the various subject fields. However, their distribution has deferred over the years in different types of publications. For example, based on the SCOPUS dataset, it is observed that in 2005 there were very few publications, but then it started to rise till 2013, and then it had a slight dip in the year 2014. Still, it rose from the following year onwards, and the most publications ever were recorded in 2020. Still, due to Covid 19 pandemic, the pace of the publications has reduced significantly in the current year, i.e., 2021. This is evident from Figure 8 given below.

Although, according to the above Figure 8, there is a dip in publications in the year 2021, which might be due to the Covid-19 pandemic second wave across the world,

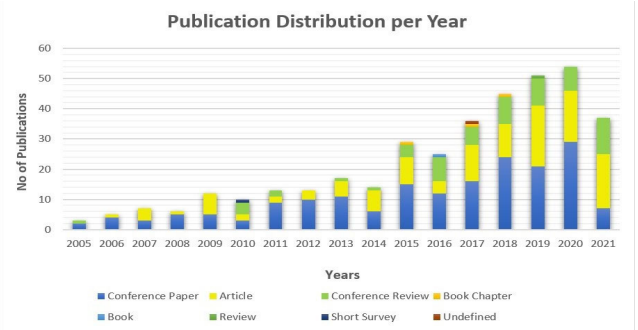

FIGURE 8. Year-wise publication distribution source: http://www.scopus.com (assessed on 10th June 2021).

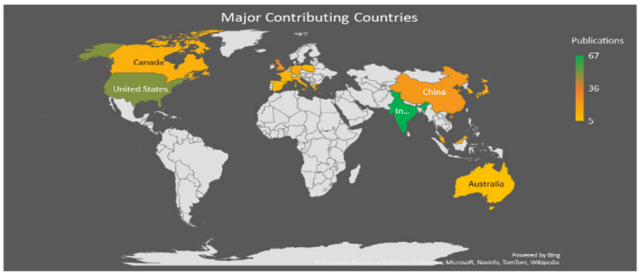

FIGURE 9. Publication distribution around the world source: http://www.scopus.com (assessed on 10th June 2021).

with forthcoming there might be a boost in the number of publications.

C. BASED ON THE GEOGRAPHICAL DATA, DISTINGUISH PUBLICATION DISTRIBUTION RELATED TO KEYWORDS

As per the analysis of these publications, it is observed that India is on the top to publish most of the papers related to Facial Based Emotion Recognition as per results obtained by Keywords. This is evident in the following Figure 9.

Following India, next is the United States, then United Kingdom, China, and Germany, which leads the world related to these publications.

D. DISTRIBUTION OF THE PUBLICATIONS BASED ON QUANTITATIVE DATA

From the analysis of the SCOPUS database, the authors found that the papers related to FER are distributed among various publications. This is illustrated in Figure 10 below.

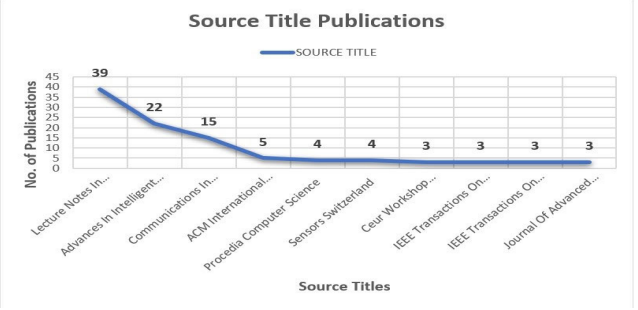

FIGURE 10. Source publications source: http://www.scopus.com (assessed on 10th June 2021).

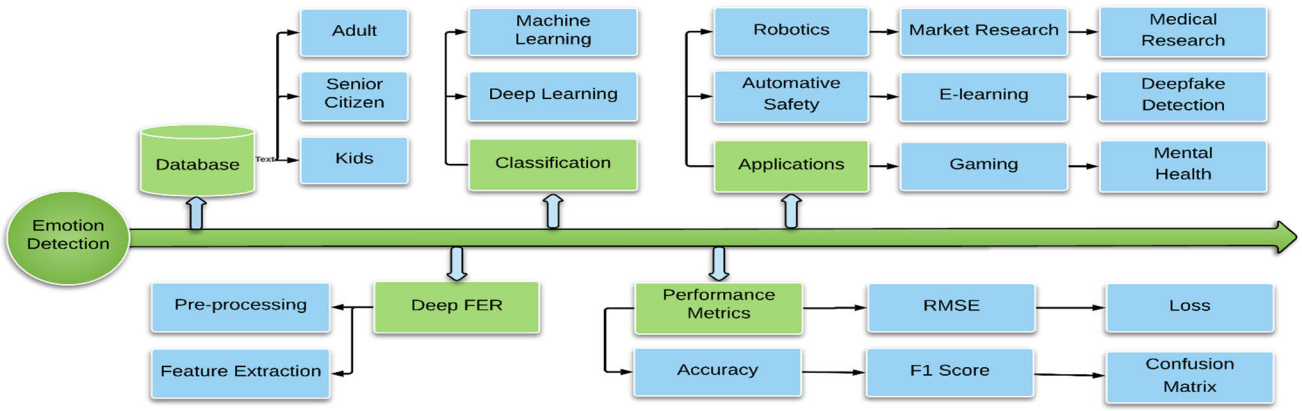

FIGURE 11. Overall brief description of trends in facial emotion recognition.

From the above Figure 10, it can be seen that the maximum number of publications in Lecture Notes of Computer Science, Artificial Intelligence and Bioinformatics, followed by Advances in Intelligent Systems, Communications in Computer Science and Information Science, ACM International, Coeur Workshops, IEEE Transactions on Affective Computing, etc.

III. QUALITATIVE ANALYSIS OF FER RESEARCH AND ITS PUBLICATION

There were only unimodal systems in the early years of the Artificial Intelligence(AI) era. Machine Learning(ML) models are used to predict emotion by only facial expressions. These facial expressions are gathered either by static images or converting a video into a series of static images to train the model and predict. In recent years all the models currently follow the same basic approach to train for facial expressions.

Such an example is shown in [13], where they used OpenCV's Haar feature Detection algorithm to create an image pre-processing program and proposed Deep Belief Network (DBN) took advantage of DBN's algorithm's ability to identify complex patterns in the input, which would yield high accuracy in their classification task. Their accuracy was around 20% on a limited set of 4 emotions. Although DBN is the older version, recent research demonstrated that Convolution Neural Network (CNN) is the most used and efficient method compared to DBN used in facial Emotion detection. Authors in [14] proposed the standard CNN model and two versions with additional customization of activation functions such as ReLu and defining Max Pooling Filters accordingly to achieve results. Such versions of CNN have consistently outperformed the original model; this example is shown in [15] the research. They proposed Ensemble of Multilevel CNN, where they used three CNN models with different filters and layers and fused them to classify emotions. These types of CNN can increase the accuracy level only up to a specific limit, so to handle this problem, research done in [16] proposed the use of Autoencoders, which is a form of neural network that can recreate its input in a lower-dimensional space, were used in conjunction with CNN to improve its Emotion Recognition accuracy. Now with the latest models apart from CNN, a lot of research has been done on hybrid models such as the CNN-RNN model where the Deformable Part Model is used for face detection, Dlib for facial extraction, as illustrated in [17], which gave the highest mean performance accuracy when compared to other state-of-the-art models. GoogleNet, which Google created, is one of these cutting-edge models used for research in [18] for Emotion detection using video clips. Still, along with the multimodal approach of using Geometric Features of the Face and tracking facial landmarks, which are the key deciding factors of Emotion classification, these facial features classifications are decided using a cluster-based strategy where image frames captured from video clips of similar positions are grouped which is assigned to specified cluster. That closest centroid of all the clusters was considered the ideal framework for training and testing using GoogleNet. These results are fused with audio emotion classification results, which resulted in significant result values and a much higher accuracy level than expected in other audio-visual models. There is much recent research done on the latest CNN models such as InceptionNet VGG, Resnet, SqueezeNet, and many more with different combinations of novel approaches [19]–[23].

With time even though new deep learning methods, algorithms, and new FER datasets are being studied using novel approaches, the gist of FERs basic flow, illustrated in Figure 11, has always been the same for more than a decade. This flow starts from the main requirement for all kinds of FER approaches a high-quality, diverse, balanced dataset. From a good dataset, only one can use any novel approach to gain the best-desired results. These results are only achieved when the required data is computed, and pre-processing data avoids unnecessary data/noise. Then after pre-processing, use a deep learning approach to train the dataset, and then the trained model is tested on performance metrics. The results achieved from these metrics would determine whether their approach is good enough to get desired outputs. Based on

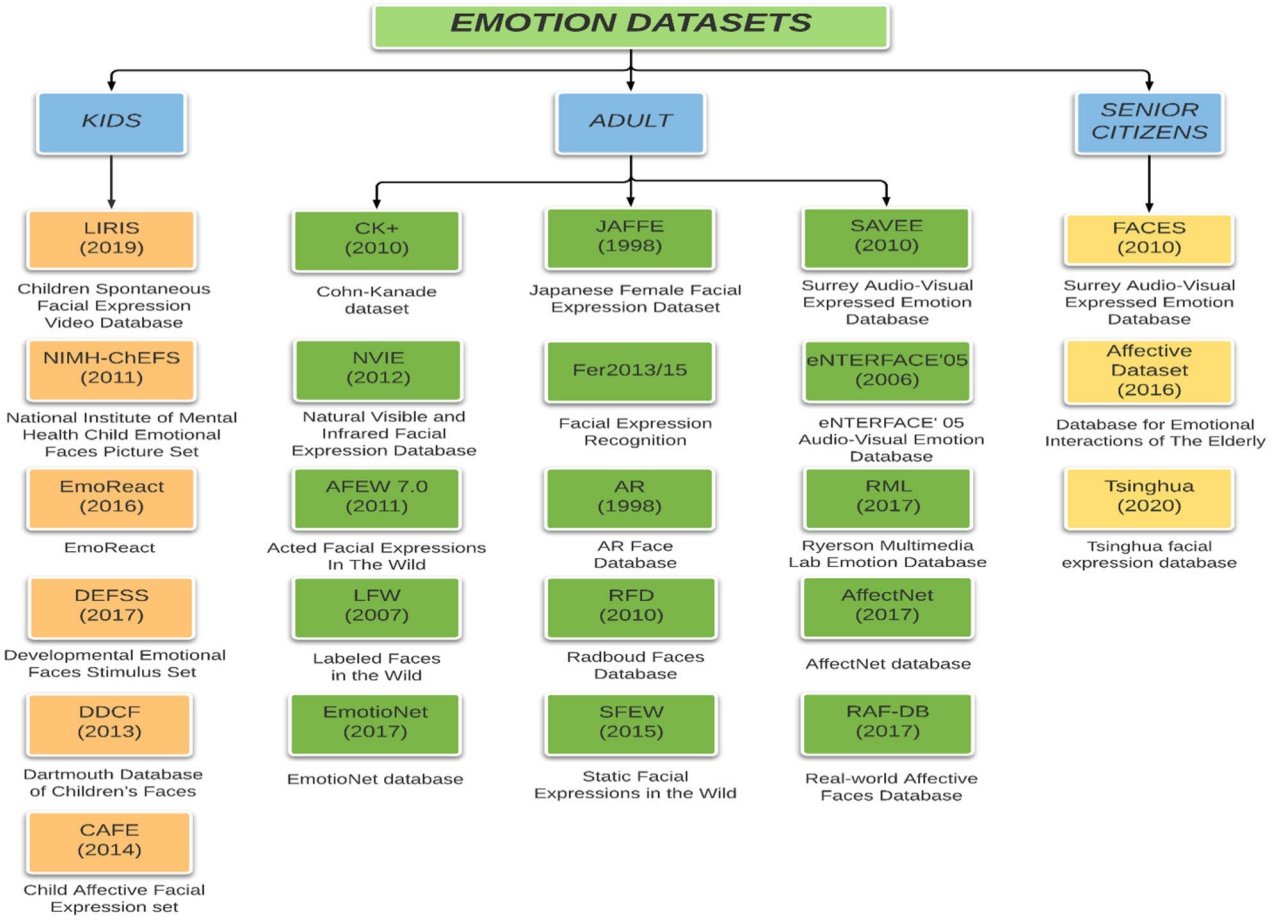

FIGURE 12. Description of different dataset available.

these outputs, the authors give a general flow of deploying a trained model into different applications.

A. DISCUSSION OF INSIGHT GAINED FROM EXISTING SURVEYS

Much work was done on FER based on different novel approaches, modalities, SOTA models, and a combination of different features to increase accuracy, which shows many insights for potential growth [24], [25]. These immense amounts of work must be categorized into and reviewed thoroughly so that all future research will have immense amounts of detailed information, to begin with. So, to understand all the work done, authors have gone through many surveys and made comparisons between them, as shown in Table 2. It was observed some research gaps after taking a look at these survey papers, and hence, authors have proposed this survey that is different from other proposed studies that will deal in line with all the work done, challenges faced, and plausible solutions along with new trends in the field of AI as well as potential applications.

From these comparisons, as shown in Table 2, it is observed that there are no Surveys on Adult and Child Facial Emotion Category Together. In addition, there is a lack of detailed discussions on State-of-the-art models, techniques, and comparative studies on different categories in each category. Therefore, authors have proposed a new survey on both children and adults based on comparative analysis of techniques and state art of the art models used and upcoming trends that will contribute more to FER on and Multimodal Emotion Recognition Research.

IV. DATASET DISCUSSION

Multiple datasets that include different populations and recording environment variations help the researchers design a more robust deep learning system for emotion detection. In this section, the authors have discussed the datasets available for emotion detection, used by researchers worldwide for FER system evaluation. They have divided this section into two parts - Kids and Adults. This is also illustrated in Figure 12. provides a visual of different kids and adult datasets used for emotion detection using video and audio.

However, when these different categories of a dataset are compared, authors comment that there is a scarcity of kids' datasets as compared to FER, so they suggest creating a new

TABLE 2. A comparative analysis of the planned survey and existing FER surveys.

Reference	Year	Objective	Merits	Demerits
[11]	2014	FER facial detection, face feature extraction, and classification approaches were investigated	Detailed review on few FER approaches	Comparison of methods was not appropriately discussed
[12]	2015	Given a thorough examination of several FER techniques based on computational paradigms	Different strategies used in day-to-day life were investigated	Did not explain how different algorithms like CNN, SVM, and others are used to classify data
[13]	2017	A survey of several recent advancements in the FER sector was presented	Examined several related works and made a comparison between them.	There was no detailed information regarding various DL approaches, benefits, drawbacks, or related concerns and obstacles.
[14]	2018	To examine databases used as data sets for algorithms that will detect emotions through facial expressions.	Covers Real-world applications scenario	Lack of discussion of future direction or suggestions as per their study on challenges faced
[15]	2018	Review and present an ensemble classifier based on traditional feature extraction methods.	Reviewing feature extraction approaches such as LBP and LDP, the ensemble classifier was introduced to improve results.	There isn't enough emphasis on new DL approaches for FER
[16]	2018	To investigate and survey a variety of FER approaches	Different techniques for detection, extraction, and classification were discussed.	There is not enough discussion of FER methodologies, future problems, present concerns, and datasets.
[17]	2018	To explain a variety of FER approaches and their critical contributions and the performance of FER strategies based on the number of recognized expressions and algorithm complexity.	Different Techniques and their performance on datasets were discussed	Lack of proper discussion of CNN in FER
[18]	2019	To investigate alternate detection, extraction, and machine learning classification methods for FER through a detailed survey.	A thorough examination of numerous methodologies was carried out, focusing on automated AU analysis from RGB photography	DL techniques were not discussed.
[26]	2019	To examine and evaluate several FER problems related to position, lighting, and age invariance	In FER, investigated several issues related to position, illumination, and age invariance	For DL approaches only CNN were discussed
[27]	2019	To investigate various CNN-based WLR approaches	A detailed examination of the various CNN architectures used for FER on various datasets	CNN's problems and challenges were not highlighted
[28]	2019	To highlight the essential elements of all possible emotion identification modes, show the accessible platform, and detail existing multimodal emotion analysis projects.	Survey different modules used for Emotion Recognition	No DL approaches were discussed thoroughly, and there is a lack of comparative study of techniques and models
[29]	2020	To give FER researchers a detailed overview of numerous AI techniques	Reviewed numerous deep neural networks for FER and presented data on datasets and potential difficulties and difficulties.	For the FER system, no taxonomy has been provided
[30]	2020	To provide a literature review of the various machine learning techniques utilized in FER	For FER, a comparison of different ML pre-processing, feature extraction, and classification algorithms was provided	Future issues for the FER system due to a lack of taxonomy. Furthermore, very few FER datasets were discussed
[31]	2020	To investigate current state-of-the-art AI approaches for FER	An in-depth survey of several AI techniques, their benefits and drawbacks, FER datasets, current concerns, and challenges	A flowchart outlining the process of implementing CNN on FER and illumination normalization and Data Augmentation research was not supplied
[32]	2020	To provide a review on trends of various CNN on FER	Surveyed various CNN models on FER	Did not provide a proper classification of datasets and work done on them
[33]	2020	To identify faces using different deep learning and machine learning methods	A detailed review of 2D and 3D face recognition methods using four different approaches	The deep learning algorithms presented are a bit outdated
[34]	2021	To provide different multimodal deep learning approaches for computer vision and the applications	Various multimodal approaches were discussed along with their applications in detail	The authors didn't specify any algorithms for the proposed methods.
Proposed Study		To provide a brief review of techniques, state-of-the-art models, and methods on different Kids, Adults, and Senior Citizen Category datasets.	Surveyed various approaches based on techniques and methods used on FER datasets of 2 categories, i.e., Kids and Adults, Future Trends, current issues, and challenges in detail	---------

novel dataset that is balanced and has a high quality of data in the kids' category and set up a new benchmark accuracy on it.

A. KIDS VIDEO DATASET

1) LIRIS CHILDREN SPONTANEOUS FACIAL EXPRESSION VIDEO DATABASE

The database (LIRIS-CSE) contains 208 movie clips /dynamic images of 12 ethnically diverse children showing spontaneous expressions. This database contains spontaneous/natural facial expressions of children in different settings showing six universal or prototypic emotional expressions, ''happiness,'' ''sadness,'' ''anger,'' ''surprise,'' ''disgust,'' and ''fear.'' The dataset [35] contains 26,000 frames of emotional data in total. 12 (five males and seven females) ethnically diverse children between the ages of 6 and 12 years with a mean age of 7.3 years participated in the database recording session.

2) DEVELOPMENTAL EMOTIONAL FACES STIMULUS SET (DEFSS)

The Developmental Emotional Faces Stimulus Set (DEFSS) is designed to provide a standardized set of emotional stimuli, including a child, teen, and adult faces, validated by participants across a wide range of ages. The dataset [36] includes 404 validated facial photographs of people ages 8 and 30, displaying five different emotional expressions: happy, angry, fearful, sad, and neutral. The DEFSS also includes a neutral emotion, which compares the positive and negative emotions among various ages.

3) NIMH CHILD EMOTIONAL FACES PICTURE SET (NIMH-CHEFS)

The NIMH-ChEFS was created through a collaborative endeavor between a neuroscience research group at the NIMH and a local children's theater group- Imagination Stage, based in Bethesda, Maryland, Washington DC. The dataset [37] consists of 482 photographs of 5 emotions- fear, angry, happy, sad, and neutral with two gaze conditions: direct and averted gaze. This dataset was recorded in a controlled environment using child actors from the children's theatre. The age of the child actors ranged from 10 to 17 years old, with a mean age of 13.6 years old. There are 39 girls and 20 boys in the picture set, a total of 59 participants. The stimuli were evaluated by 20 volunteers, all faculty and staff working in the CDE at Duke University Medical Center. Duke IRB approved the methodologies in the study for which these images were to be used.

4) DARTMOUTH DATABASE OF CHILDREN'S FACES

The Dartmouth Database of Children's Faces [38] consists of photographs of 80 children-40 male and 40 female Caucasian children between 6 and 16 years of age. Child actors were used to recording the dataset. The actors posed for eight facial expressions and were photographed from five camera angles under two lighting conditions. In addition, the actors wore specific outfit-black hats and black gowns to minimize extra-facial variables. Independent raters were used to validate the images. The raters identified facial expressions, rated their intensity, and provided an age estimate for each model. The Dartmouth Database of Children's Faces is freely available for academic and research purposes.

5) CHILD AFFECTIVE FACIAL EXPRESSION SET (CAFE)

The Child Affective Facial Expression set (CAFE) [39] features photographs of 2 to 8-year-old children posing the six basic emotions defined by Ekman—sadness, happiness, surprise, anger, disgust, and fear—plus a seventh neutral expression. It is also racially and ethnically diverse, featuring European American, African American, Asian, Latino (Hispanic), and South Asian (Indian/Bangladeshi/Pakistani) children. There are 1192 photographs in the entire CAFE set, which includes one subset of faces (Subset 1) that contains only highly stereotypical exemplars of the various facial expressions, consistent with other existing face sets, and a second subset (Subset 2) that in contrast only includes faces that emphasize variation around emotion targets in research participants while minimizing potential ceiling and floor effects.

6) EMOREACT

EmoReact [40] is a multimodal emotion dataset containing 1102 videos of children between 4 and 14. This dataset is annotated for 17 affective states, including eight basic/universal emotions - happiness, sadness, surprise, fear, disgust, anger, neutral, valence, and nine complex emotions - curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment, and frustration. Crowd workers from the online crowdsourcing platform Amazon's Mechanical Turk (MTurk) were recruited to obtain the labels in EmoReact. Three independent workers annotated each video for a total of 17 labels. The interface for annotations contained the definitions of each label for consistency. As a test of the rater's vigilance and rational decision-making, a question about the gender of the child in the video was included. The length of these videos ranges between 3 seconds to 21 seconds, with an average length of about 5 seconds. Sixty-three different children, 32 females and 31 males, expressed the emotions, with some diversity in ethnicity.

B. ADULT VIDEO DATASET

1) RADBOUD FACE DATABASE

The Radboud Faces Database (RFD) [41] is laboratory-controlled and has 1,608 images from 67 subjects with three different gaze directions, i.e., front, left, and right. Each sample is labeled with one of eight expressions: anger, four contempt, disgust, fear, happiness, sadness, surprise, and neutral.

2) EXTENDED COHN-KANNADE (CK+)

The Extended Cohn-Kanade (CK+) [42] dataset consists of 593 video sequences from a total of 123 different subjects. The age of the subjects ranges from 18 to 50 years of age with various genders and heritage. Each video shows a facial shift from the neutral expression to a targeted peak emotion, recorded at 30 frames per second (FPS) with a resolution between 640 × 490 or 640 × 480 pixels. A total of 327 videos are labeled with one of 7 universal emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise.

3) JAPANESE FEMALE FACIAL EXPRESSION (JAFFE)

The JAFFE dataset [43] includes 213 images of different facial expressions from 10 different Japanese female subjects. Each subject was asked to do seven universal/basic facial expressions (6 basic facial expressions plus neutral). The images were annotated with average semantic ratings on each facial expression by 60 annotators.

4) NVIE

A total of 215 healthy students (157 males and 58 females), ranging in age from 17 to 31, appear in the dataset. There are 105 subjects under front illumination for the spontaneous database, 111 subjects under left illumination, 112 subjects under right illumination, and 108 subjects contributed to the posed database [44].

5) FER2013

The FER2013 database [45] was introduced during the ICML 2013 Challenges in Representation Learning. The dataset contains 35,887 grayscale images of faces with 48 * 48 pixels. The dataset consists of 7 basics/universal expressions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The images are stored in a CSV format. Each of the 35,887 rows contains emotion indexes: 0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, and 6 = Neutral. The images are stored as 2304 integers which is the grayscale intensity of associated pixel to 48 × 48 image (2304 = 48 * 48) and are separated by space. Whether it is for training or public test, or private tests, the usage is also defined.

6) AR FACE DATABASE

Aleix Martinez and Robert Benavente created the AR facial expressions database [46] in the Computer Vision Center (CVC) at the UAB. It contains over 4,000 color images corresponding to 126 people's faces consisting of 70 men and 56 women. The images feature a frontal view of faces with different facial expressions, illumination conditions, and occlusions.

7) ACTED FACIAL EXPRESSIONS IN THE WILD (AFEW)

This database [47] has been used as an evaluation platform for the annual Emotion Recognition in The Wild Challenge (EmotiW) since 2013. AFEW dataset contains video clips from different movies with spontaneous expressions, multiple head poses, occlusions, and illuminations. AFEW is a temporal and multimodal database that provides vastly different environmental conditions in both audio and video. The samples are labeled with seven basic/universal expressions: anger, disgust, fear, happy, sad, surprise, and neutral. The annotation of expressions has been continuously updated, and reality TV show data have been continuously added. The AFEW 7.0 is independently divided into three data partitions in terms of subject and movie/TV source: Train (773 samples), Val (383 samples), and test (653 samples), which ensures data in the three sets belong to mutually exclusive movies and actors.

8) AFFECTNET

AffectNet [48] is a database of facial expressions in the wild created by collecting and annotating facial images. Affect is a psychological term used to describe the outward expression of emotion and feelings. AffectNet contains more than 1M facial images collected from the Internet by querying three major search engines using 1250 emotion-related keywords in six different languages. About half of the retrieved images (∼440K) were manually annotated for the presence of seven discrete facial expressions (categorical model) and the intensity of valence and arousal (dimensional model).

C. SENIOR CITIZENS DATASET (MIXED WITH YOUNG AND ADULTS CATEGORY)

1) TSINGHUA FACIAL EXPRESSION DATABASE

Tsinghua Facial Expression Dataset [49] consist of face were 67 native-Chinese adults, aged 19–35 years (M = 23.82 models that years, SD = 4.18 years; age range, 19–35 years; 34 women) and 70 native-Chinese people, aged 60–76 years (M = 64.40 years, SD = 3.51 years; age range, 60–76 years; 35 women). M denotes for mean, and SD denotes Standard Deviation. It is reported that the identification rate of this dataset is between 70.19% to 88.87%. However, the average identification rate is 79.08%. In this dataset, the Chinese subjects, such as young and older female and male faces, portray eight basic facial expressions (Neutral, Sadness, Disgust, Fear, Anger, Happiness, Content, and Surprise).

2) DATABASE FOR EMOTIONAL INTERACTIONS WITH ELDERLY

The database [50] was created using audio and video from sixteen actors (8 female and 8 male) who participated in daily TV series discussions and covered seven different emotions: anger, boredom, pleasure, sorrow, surprise, neutrality, and disgust. There are 810 speech-video snippets in the collection from 118 talks. In this dataset, Each voice and video segment lasts 3-5 seconds. This dataset was recorded from the "Empty Nest grandpa," which reflects the elderly life. Anger, boredom, pleasure, sadness, anxiety, neutrality, and disgust are among the seven types of emotions covered in this dataset.

3) FACES

FACES [51] is a database consisting of 171 naturalistic faces of young, middle-aged, and older women and men. Each face is represented with two sets of six facial expressions (neutrality, sadness, disgust, fear, anger, and happiness), resulting in 2,052 individual images. N = 154 young, middle-aged, and older women and men rated the faces in terms of facial expression and perceived age. With its large age range of faces displaying different expressions, FACES is well suited for investigating developmental and other research questions on emotion, motivation, and cognition, as well as their interaction.

From Figure 12, it's clear that there has been a lot of Adult FER dataset compared to Kids and Senior citizen dataset. The oldest and most commonly used dataset is the JAFEE dataset which is among the Adult category, and among these datasets, the most famous datasets are FER2013/15, Cohn Kanade(CK+), and Survey Audio-Visual Expressed Emotion Database(SAVEE). The latest dataset in the adult category was created in 2017: EmotioNet and Real-world Affective Face Dataset(RAF-DB). In the Kids category, the earliest dataset is the Dartmouth Database of Children's Faces(DDCF), used heavily in FER, and the latest one is the LIRIS dataset, in which very little work is done. Also, very little work is done using the Senior Citizen dataset, even on the older dataset which is FACES. However, it is observed that there is a low amount of work done using the Kids and Senior Citizen dataset, as shown in the Table 3. This might be because of the unbalanced dataset or scarcity of diversity, making the models less efficient in real-world applications. So, there are many scopes to create new datasets in Kids and senior citizens that can be diverse, balanced, and with different high-quality modalities.

V. DEEP FACIAL EMOTION RECOGNITION

In this section, the authors presented describing the in-depth steps required for FER. Every step has multiple techniques which can be implemented depending upon different cases. For example, the authors would be dealing with pre-processing, feature extraction, and different state-of-the-art models in detail illustrated in Figure 13. This section gives a comparison and shows insights that can be useful for literature in FER in recent years or upcoming research.

A. PRE-PROCESSING

In this stage, the authors clean up the dataset by eliminating noise and compressing the data or not having any more data than one should need. Following are the stages in pre-processing of the data in the form of image or video frames:

Face Detection: It is used to find the face in every photograph or picture. Face detection is a subset of object-class detection that checks for the presence of a face in an image.

Normalization: Feature scaling is another name for it. After this stage, the image features are reduced and normalized without altering the distinguishable spectrum of feature values. To carry out normalization, one can use Z Normalization, Min-Max Normalization, and Unit Vector Normalization; some of the commonly used normalization methods increase numerical consistency and enhance model preparation.

Data Augmentation: To cope with less data, it is augmented, which is used to generate new data by using various transformations of an image with face data intact.

1) FACE ALIGNMENT/FACE DETECTION

In several facial recognition tasks, face orientation is a standard pre-processing stage. The most commonly used and open-source implementations for deep FER are shown in this section. Provided a traditional set of training data, the first initial step would be to detect the face, followed by removing non-facial components, including backgrounds. As presented below, there are a variety of techniques for detecting faces.

a: THE VIOLA-JONES (V & J) FACE DETECTOR

It is one of the most extensively used face detection implementations [51]. To detect frontal faces is both reliable and computationally inexpensive. Because detection of a face is the only technique the authors need to enable feature learning and alignment of a face using local landmark coordinates to achieve high accuracy in FER, this step is crucial because the variation of face positions better it will work. Based on the research done by authors named Viola and Jones, there are three types of Haar-like features [52] which is illustrated in Figure 14, and these are the following:

b: HAAR CLASSIFIER

Usually, by reducing the pixel size group, Haar features are measured. Haar Classifier has used Haar-like features to detect an image. This method allows objects to be detected in multiple sizes [53]. Haar classifiers identify features contributing the most to solve face detection problems in the training phase. It may indicate high detection accuracy, and the computation complexity is small. In Figure 15, it is evident how these classifiers detect faces.

c: ADAPTIVE SKIN COLOR

The adaptive skin-color model is used as a face detection method based on a skin-color model to detect the face region [54]. This algorithm shows a high accuracy since skin color is used for image segmentation. Hence it can be easy to differentiate the face region and non-face region. The only drawback is that this algorithm does not work with different levels of illumination. An adaptive gamma corrective method can avoid this problem, but it cannot be used in real-time due to its extremely high computational power.

d: ADABOOST CONTOUR POINTS

Due to the low computational power required, Adaboost is most suitable for real-time scenarios [55]. In this method, several classifiers can be cascaded. First, it trained the faces and built a robust classifier that gives high accuracy in

TABLE 3. Comparison of available datasets.

Category	Dataset	Sample	Subjects	Recording Condition	Elicitation Method	Expressions	Dataset link
KIDS	LIRIS	208 Video Samples	12	Lab & Home	Spontaneous	Disgust, Happy, Sad, Surprise, Neutral, Fear	https://childrenfacialexpression.projet.liris.cnrs.fr/
KIDS	DEFSS	404 images	116	Lab	Posed	Happy, Angry, Fear, Sad, Neutral	https://reflectionsciences.com/researchers-and-clinicians/
KIDS	NIMH-Chefs	480 images	59	Lab	Posed	Fear, Happy, Sad, Neutral	http://www.scanlab.org/downloads.html
KIDS	DDCF	3200 images	80	Lab	Posed	Angry, Sad, Disgust, Afraid, Happy, Surprised, Contempt, Neutral	https://lab.faceblind.org/k_dalrymple/ddcf
KIDS	CAFE	1192 images	90 females 64 males	Lab	Posed (exception: surprise)	Sad, Happy, Surprise, Anger, Disgust, Fear, Neutral	https://nyu.databrary.org/volume/30
KIDS	EmoReact	1102 Videos	32 females 31 males	Web	Spontaneous	Happy, Sad, Surprise, Fear, Disgust, Anger, Neutral, Valence, Curiosity, Uncertainty, Excitement, Attentiveness, Exploration, Confusion, Anxiety, Embarrassment, Frustration	https://www.behnaznojavan.com/emoreact
ADULT	RFD	1608 images	67	Lab	Posed	Sad, Happy, Surprise, Anger, Disgust, Fear, Neutral	http://www.socsci.ru.nl:8180/RaFD2/RaFD
ADULT	CK+	593 video samples	123	Lab	Spontaneous + Posed	Angry, Sad, Disgust, Afraid, Happy, Surprised, Contempt, Neutral	http://www.consortium.ri.cmu.edu/ckagree/
ADULT	JAFFE	213 images	10	Lab	Posed	Sad, Happy, Surprise, Anger, Disgust, Fear, Neutral	https://zenodo.org/record/3451524#.YHwzpegzaUk
ADULT	NVIE	236 images	157 males 58 females	Lab	Spontaneous + Posed	Happy, Sad, Surprised	https://nvie.ustc.edu.cn/
ADULT	FER2013	35887 images	N/A	Web	Spontaneous + Posed	Sad, Happy, Surprise, Anger, Disgust, Fear, Neutral	https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge
ADULT	AR	4000 images	70 males 56 females	Lab	Posed	Happy, Angry, Neutral, Scream	http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html
ADULT	AFEW 7.0	1809 videos	N/A	Movies	Posed	Sad, Happy, Surprise, Anger, Disgust, Fear, Neutral	https://sites.google.com/site/emotiwchallenge/
SENIOR CITIZEN	Tsinghua	110 images	32 males 31 females 21 old males 26 old females	Lab	Posed	Anger, Contempt, Disgust, Fear, Happy, Neutral, Sad and Surprised	N/A
SENIOR CITIZEN	Database for emotion interactions of the elderly	810 videos 810 audios	8 males 8 females	Movies	Posed	Anger, Disgust, Anxiety, Happiness, Neutrality, Sadness, and Boredom	N/A
SENIOR CITIZEN	FACES	171 images	58 young males 56 young females 28 old males 29 old females	Lab	Posed	Happy, Anger, Fear, Sad, Disgust, Neutral	http://faces.mpib-berlin.mpg.de.

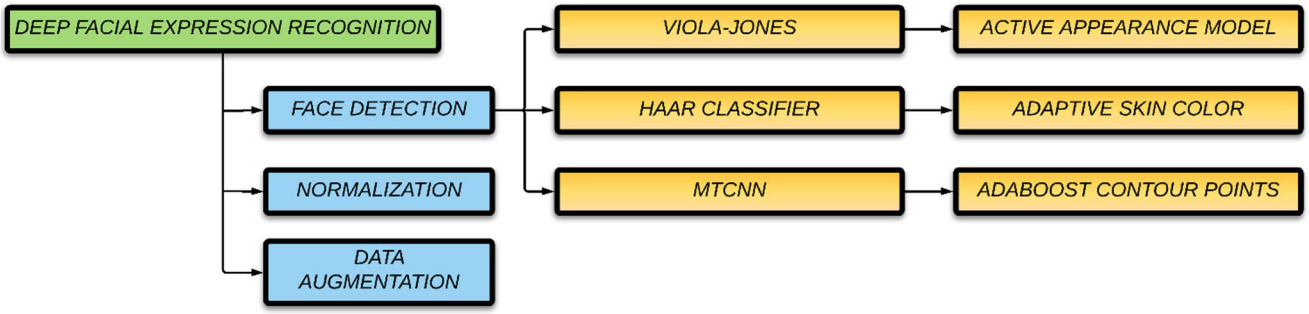

FIGURE 13. Overview of pre-processing.

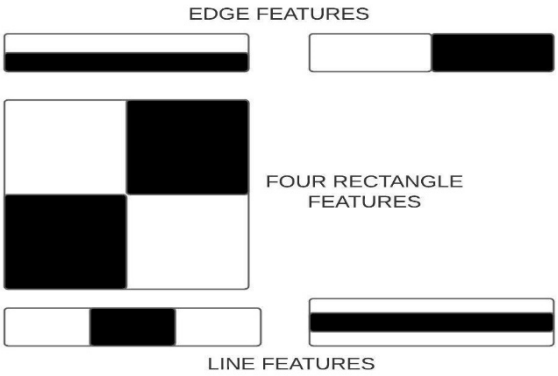

FIGURE 14. V & J face detection.

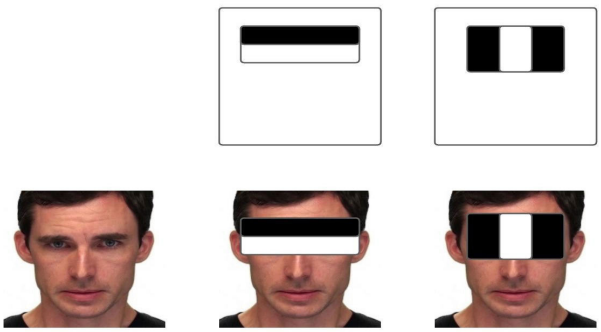

FIGURE 15. Haar classifier illustration.

detecting faces. Then the new Face is compared with the model built by the classifier. Along with that, there are also usage contour points to detect faces. The contour points may give good accuracy and performance because very low features are extracted at the end, making it less complex.

e: ACTIVE APPEARANCE MODEL

It is one of the basic computer vision algorithms for transferring the statistical model of object shape and appearance to a whole new image. First, the model is built during a training phase [56]. Then, a series of images, which are set together with coordinates of landmarks on the faces that appear in all the images, is provided to the training supervisor.

f: MTCNN

Multi-task Cascaded Convolutional Networks (MTCNN) is a well-known framework developed as a solution for both face detection and face alignment in solving many computer vision-related problems [57]. The process comprehends three crucial stages. First, a convolutional network can recognize faces and landmarks such as eyes, nose, and mouth. There are three stages of MTCNN.

- *Proposal Network (P-Net)*
- *Refine Network (R-Net)*
- *Output Network (O-Net)*

In the first stage, a shallow CNN is used to produce candidate windows. Then it is refined using a complex CNN in the second stage, and at the end in the last stage, which is the third stage, a more complex CNN is used to refine the result and output further and plot facial landmark positions accurately. The authors have also given an illustration in Figure 16 where how faces are detected in MTCNN.

A wide variety of facial detection algorithms make it difficult to select a proper algorithm that will detect Faces based on different application cases. In the real-time scenario, the authors comment that one must select the best algorithm, to begin with, not hindering the application's computation and quality data gathering. So authors have illustrated a comparative study of Face Detection Methods, as shown in Table 4, based on a Real-time environment.

2) FACE NORMALIZATION

Usually, the data is not consistent in illumination and head poses in any facial datasets, significantly reducing the Facial Emotion Recognition model's performance. To overcome this, the authors suggest using either of two normalization methods for FER, which are the following:

a: ILLUMINATION NORMALIZATION

For any Facial image, its illumination and contrast can be different even though they consist of the same expressive emotions, especially in non-isolated environments, which

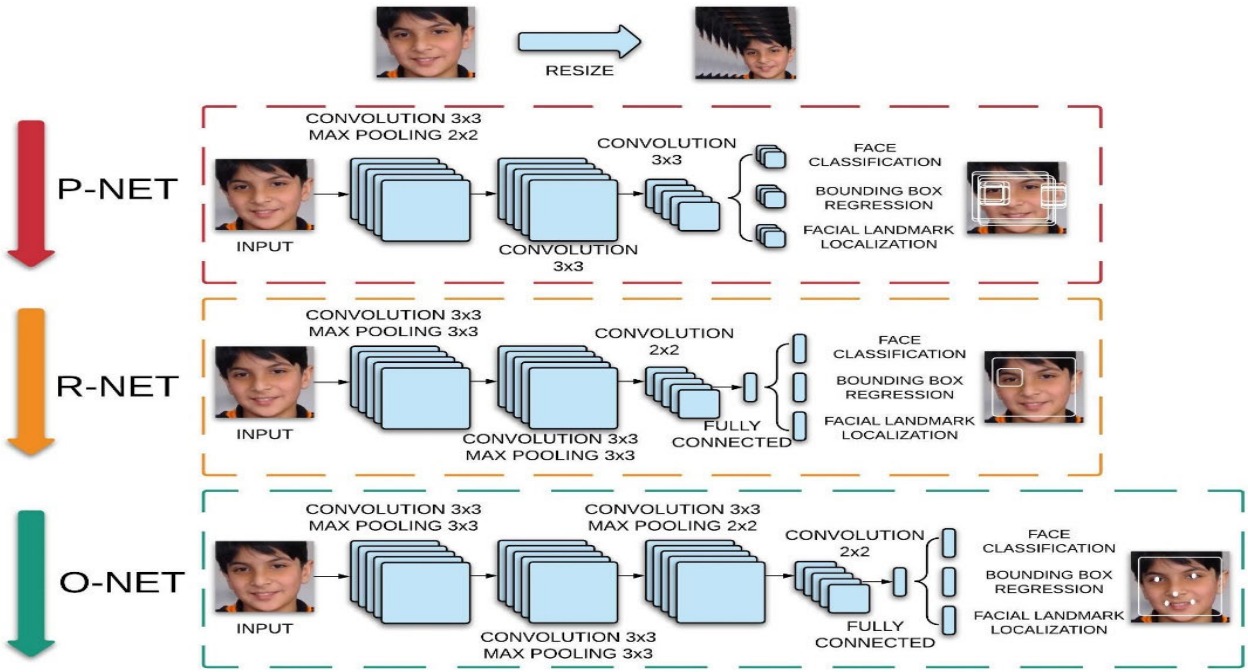

FIGURE 16. MTCNN.

TABLE 4. Comparison of different face detection methods on real-time face detection.

Algorithm	Reference and Year	Reason for Good Accuracy	Observations
Viola-Jones	[58] 2018	As it was the first real-time face detection algorithm, the accuracy may vary with the number of detected faces in the image database.	Using various Image Processing Techniques performed on images, faces have been detected through skin segmentation in greyscale. The face regions were found by analyzing the greyscale and binary variation in different face regions. Face recognition implementation was successfully enabled through a global classification technique, which classifies the faces from the standard deviation difference between input and average faces.
Haar Classifier	[59] 2019	The accuracy is high because it is good at detecting edges and lines	The computational complexity is less since features contribute the maximum during the training time in the face detection problem.
Adaptive Skin Colour	[60] 2011	The accuracy is good since it is easy to identify skin color, but it fails in different illumination levels.	The illumination problem necessitates a high computational complexity, making it unsuitable for use in a real-time setting. The gamma correction process is used to deal with this.
Adaboost Contour Points	[55] 2011	A robust classifier is used to detect a single face using contour points which result in good accuracy.	A smaller number of features require low computational complexity, which the training model uses, which leads to a low computational cost
Active Appearance Model	[61] 2006	It achieves high accuracy because it is widely used to extract features from human faces under various physical and environmental conditions.	Fitting the model to the original image, on the other hand, is a difficult task in such an active appearance model.
MTCNN	[62] 2019	It consists of a custom CNN which simultaneously works on face detection and alignment in real-time, which leads to good accuracy.	It performs better than most other methods and has the speed advantage, but it does not convince a lower resolution image.

would result in significant internal differences in their respective features. In [63], there has been immense usage of many algorithms such as isotropic diffusion(IS), discrete cosine transform (DCT) [64], and their difference of Gaussian (DoG), which were later studied and analyzed for a thorough evaluation of illumination normalization. Also in [65] used normalization based on homomorphic filtering, which consistently showed the best results to eliminate illumination

normalization. There have also been recent studies that show that histogram equalization in combination with illumination normalization has proved to be performing far better than the results from normalization alone. Furthermore, there has been much research (e.g [66]–[69]) which have used histogram equalization for pre-processing by enhancing the global contrast of facial images. This technique is very constructive since the foreground and background brightness is indistinguishable. However, this may exaggerate local contrast by using it straightforward. To overcome this, [70] proposed approach where the weighted summation is used to fuse linear mapping and histogram equalization. There has also been a comparison as shown in [67] (i.e., global contrast normalization (GCN), local normalization (LN), and histogram equalization(HE)). However, HE and GCN have shown the best accuracy for testing steps and training steps, respectively.

b: POSE NORMALIZATION

In a moving face in videos or series of non-frontal images of faces, pose variation is common. Some studies used pose normalization methods to get frontal faces depicting categorical emotions for Facial Emotion Recognition (e.g., [71], [72]), but the most famous one was one proposed in the study [73]. After localized facial landmarks, a model is generated, a generic 3D texture effectively predicts facial components. Then, back-projecting face images, synthesized from the initial frontal Face, are used to predict the Face's visual components. However, [74] proposed a model that stores landmarks locally and uses frontal faces in view only into facial poses, which can act as an alternative.

3) DATA AUGMENTATION

Generally, Image Data Augmentation is used to ensure generalizability to some specific detection tasks. This method [75] is often used to get accurate results when training deep neural networks. Since training datasets associated with FER do not have enough images, data augmentation becomes necessary for training the data and getting the highest accuracy level. This is also illustrated in, which discussed the same concept of enriching the available training dataset. Also, authors in [76] presented different augmentation methods, including poses synthesis, glasses synthesis, illumination synthesis, hairstyle synthesis, and landmark perturbation. Augmentation transforms the image into three categories which are illustrated in Figure 17.

FIGURE 17. Data augmentation.

B. FEATURE EXTRACTION

1) TEXTUAL FEATURES

The following are the descriptors that carry out feature extraction using texture-based feature techniques. The Gabor filter, which combines phase and magnitude information, is one of the most used texture descriptors for feature extraction. The Gabor filter restricts the information about the organization of the facial image using the magnitude feature [90]–[94]. LBP features are typically created using binary code and can be achieved by thresholding between the center pixel and its neighboring pixels [95], [96]. LBP with Three Orthogonal Planes (TOP) features is retrieved for multi-resolution techniques, as illustrated in [97]. It's also used to extract non-dynamic appearances from a group of static face photos using features [98]. LBP features are usually formed with binary code produced by thresholding between the center pixel and its neighbors. Based on this study, texture-based feature descriptors are more effective for feature extraction than other methods because they extract texture characteristics connected with the look, resulting in crucial feature vectors for FER.

Weber Local Descriptor (WLD) is a texture-based feature extraction methodology that derives high discriminant texture-based features from segmented face images [96]. The Supervised Descent Method is used to extract features in three phases (SDM). The primary facial positions are correctly retrieved initially, and then the corresponding locations are selected. Finally, it calculates the distance between distinct facial components [99]. Another descriptor, Weighted Projection-based LBP (WPLBP), is a feature extraction method that extracts LBP features based on instructional regions and then weights these features depending on the relevance of the instructional areas [100]. The Discrete Contourlet Transform (DCT) recovers texture-based characteristics by dividing the image into two essential steps. The Laplacian Pyramid (LP) and Directional Filter Bank stages are used in the modified domain (DFB). The image is partitioned into a low pass, bandpass, and positional discontinuities in the LP stage. The DFB stage processes the bandpass and generates the linear composition by associating the positional discontinuities, just like the LP stage. Many texture feature-based descriptors, such as the Local Directional Number (LDN) pattern, the Local Directional Ternary Pattern (LDTP) [101], the KL-transform Extended LBP (KELBP) [102], and the Discrete Wavelet Transform (DWT) [103], are frequently employed as feature descriptors in recent years in the field of FER.

2) EDGE-BASED FEATURES

The following are the descriptors that are used to extract features using edge-based approaches. The Line Edge

Map (LEM) descriptor is a face expression descriptor that uses the dynamic two-strip technique to improve geometrical structural information (Dyn2S) [104]. Two facial features are often extracted based on motion analysis: discriminative and non-discriminative face features [105]. Based on a graphics processing unit, Edge feature extraction can be done with edge detection, tone mapping, enhancement, and local appearance model matching. The image ratio of features is retrieved from the expressed face images using the Active Shape Model (GASM). Edge feature extraction may be done using edge detection, tone mapping, enhancement, and local appearance model matching. The image ratio of features can be extracted from the expressed face images using the Active Shape Model based on a graphics processing unit (GASM). Also, there has been the usage of Histogram of Oriented Gradients (HOG), a feature extractor that uses gradient filters for edge-based featured data.

3) GLOBAL AND LOCAL FEATURES

The following are the descriptors for extracting features using global and local feature-based approaches. Principal Component Analysis (PCA) is a feature extraction approach that extracts global and low-dimensional features. It is one of the most used methods in FER. Independent Component Analysis (ICA) is another feature extraction method that uses multichannel observations [106] to extract local characteristics. Stepwise Linear Discriminant Analysis (SWLDA) is a feature extraction methodology that extracts localized features using both backward and forward regression models based on the class labels of F-test values, which are predicted for both regression models [107]. Discrete Fourier Transform (DFT) is more of a conventional way of extracting global features. Along with this, the authors suggest using Gabor wavelet transform(GWT) to extract local features as per a recent study [108].

4) GEOMETRIC FEATURES

Methods for extracting discrete geometric characteristics from photos are known as geometric feature learning methods. Geometric aspects are simple objects of geometric elements such as lines, points, curves, or surfaces. These characteristics include corners, which are a fundamental but essential property of objects. The corner features of complex things are frequently different from one another. The technique known as Corner detection can extract the corners of an object [109]. The distance and angle between two straight line segments were utilized to define a corner uniquely. Edges are one-dimensional structure features of an image, whereas features are defined as a parameterized mixture of many components. They demarcate the boundaries of several image regions. The outline of an object can be easily determined by employing edge detection to locate the item's edge. Also, blobs that represent sections of images are recognized using the blob detection method [110]. A ridge can be thought of as a one-dimensional curve that indicates an axis of symmetry by ridges from a practical standpoint. Local

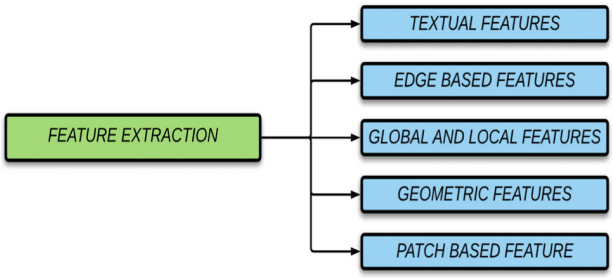

FIGURE 18. An overview of different types of feature extraction.

Curvelet Transform (LCT), a feature descriptor that extracts geometric features based on the wrapping mechanism, is one of the descriptors that extract features based on geometric feature-based approaches [17]. These geometrical features are generally mean, entropy, and standard deviation as per [111], and kurtosis is extracted by using a three-stage steerable pyramid representation [112] with addition to the energy of these geometrical features.

5) PATCH-BASED FEATURE

Face movement is recovered as patches based on distance characteristics, which are achieved using patch extraction and patch matching, commonly achieved by converting extracted patches into distance characteristics. This method is also applicable to videos which are illustrated in a recent study [113]. Also, based on a recent study [114] using this method, faces are divided into patches, and then their features are extracted and then used KNN for classification. Gabor features have also been combined with a patch-based extractor to overcome the lack of accuracy on the linear representation of the small sample size [115]. Another study [116] used this combination where the 3D Gabor features and patch method were used. In some cases, the extracted patches of images are converted into an image matrix in a PCA framework using the patch method. Then by calculating the correlation of these distinguishable patches and using that, a projection matrix is generated, which is later used by KNN for the classification of faces [117].

C. FEATURE LEARNING THROUGH DEEP NETWORKS

Deep learning has risen to prominence as a hot research issue in computer vision, with state-of-the-art performance in several applications such as image categorization using classification methods [118]. Deep learning uses hierarchical designs of many nonlinear transformations and representations to capture high-level abstractions. In this section, the authors briefly introduce such methods used for emotion recognitions using images/videos. Among these, there are four standard methods which are used in FER in recent years. These methods are Deep Believe Network, CNN, Deep Autoencoder, and Recurrent Neural Network. This is also illustrated in following Figure 19.

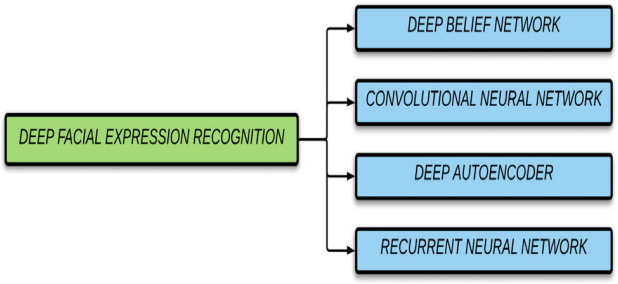

FIGURE 19. Overview of feature learning through deep learning.

1) DEEP BELIEF NETWORK (DBN)

Based on Restricted Boltzmann Machine (RBM) [119] and its unsupervised and abstract input signals from feature extraction, the DBN is introduced [120], which can learn abstract facial image information by itself and is susceptible to activity factors because of the generation of the probability distribution over observed data and their labels. This type of network is generated by putting RBMs on top of one another and trained them using the greedy approach proposed in a recent study [121]. Generally, a layer-by-layer approach is taken when applying a greedy strategy for initializing the network and later fine-tuning the weights to get the desired output. DBNs are also called graphical models because while training; it learns hierarchical representation in which joint dissemination between noticed vector x and l has hidden layer h k as shown in equation 1:

$$P(x, h1, \ldots, hl) = (\prod l - 2k = 0P(hk \mid hk + 1))P(hl - 1, hl) \quad (1)$$

where x = h 0, P(h k|h k + 1) is a conditional distribution for the visible units at level k conditioned on the hidden units of the RBM at level k + 1, and P(h l − 1|h l) is the visible-hidden joint distribution in the top-level RBM.

Fused with other modules, it has been assured to be a better emotion recognition approach from facial images. An example would be Boosted Deep Belief Network [122], which used feature selection and classifier construction combined by executing an enumeration of three training stages. In such a framework, features are fine-tuned before selecting them to construct a powerful classifier. Additionally, the discriminative qualities of selected features are repeatedly strengthened depending on their relative relevance to the robust classifier, and then highly complex features from the face pictures are learned. Many combinations have been done on DBN in many studies, such as fusing of unsupervised feature learning module of DBN with Multi-Layer Perceptron(MLP), which acts as a classification module where DBN is used to extract abstract facial features such as primary pixels of images of a face expressing emotions. MLP is used as a classifier by using learning results obtained from DBN [123]. Another proposed combination is of Local Binary Patterns(LBP) features robust to rotation and light, and DBN is used to extract another feature and emotion classification [123]. Alternately a triple combination of Local Directional Position Pattern(LDPP), Principal Component Analysis(PCA), Generalised Discriminant Analysis(GDA) features are fused with DBN for recognition and emotion modeling, which not only has tolerance against variation in illumination factors which also extracts salient features which gave far better accuracy than the traditional ones [124]. Authors have also illustrated, Deep Believe Network (DBN) architecture for emotion classification illustrated in Figure 20.

2) CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN is an improvement from Artificial Neural Network (ANN) [125]. There are multiple applications of CNN. For example, this CNN has been used in a study [126], showed that if neurons with similar parameters are applied on patches of the previous layer at various areas, a type of translational invariance is gained, is one of the primary computational models based on these nearby networks among neurons and progressively coordinated changes of the picture. Generally, a CNN includes three kinds of essential layers with an additional layer which are as follows:

a: CONVOLUTIONAL LAYERS

Like the transitional component maps, a CNN utilizes distinct parts to convolve the complete picture in the convolutional layers, resulting in different element maps. However, because of the benefits of the convolution activity, research [127] has advocated that it should not replace related layers to achieve faster learning times.

b: POOLING LAYERS

The spatial measurements (width and tallness) of the info volume for the next convolutional layer are reduced by pooling layers. The depth of the volume measurement is unaffected by the pooling layer. This layer's activity is also known as subsampling or downsampling because reducing size causes data loss. Such a tragedy is beneficial to the organization since the size reduction reduces computational overhead for the organization's subsequent layers and eliminates overfitting. The most often used systems are normal pooling and maximum pooling. The paper [128] provides a detailed hypothetical comparison of max pooling and normal pooling exhibitions, while [129] demonstrates that maximum pooling can speed up assembly, pick prominent invariant highlights, and improve speculation. However, other distinct types of pooling layer in the literature, each inspired by different ideas and fulfilling certain needs, such as stochastic pooling [130], spatial pyramid pooling [131], and def-pooling [132].

c: FULLY CONNECTED LAYERS

The high-level thinking in the neural organization is done by totally associated layers after a few convolutional and pooling layers. As the term implies, neurons in a related layer have

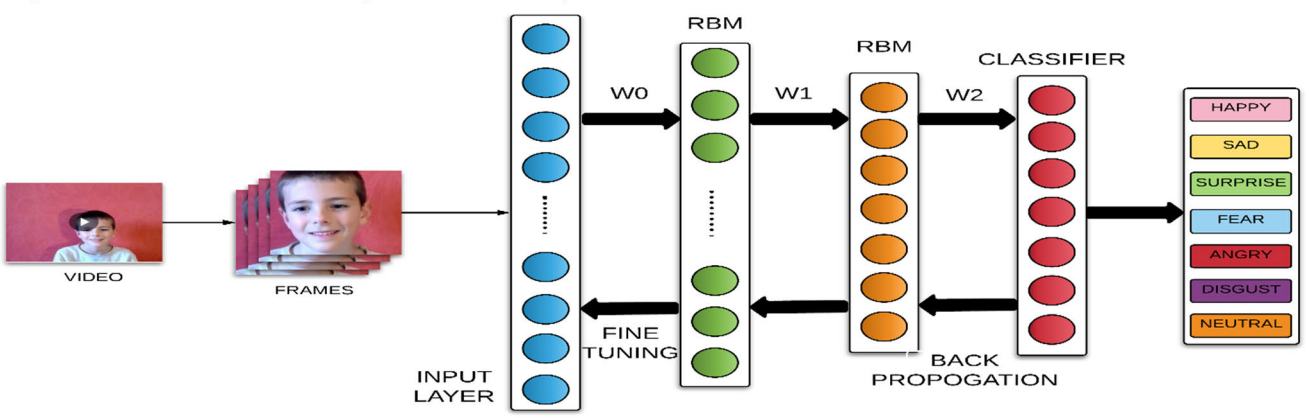

FIGURE 20. Deep belief network architecture.

connections with those in the previous layer. Following that, a network augmentation and an inclination counterbalance can be used to record their implementation. At the end of the process, fully associated layers transform the 2D component mappings into a 1D element vector. The determined vector might be divided into a predetermined number of categories for grouping [133] or treated as a component vector for additional handling.

CNN's are built using three crucial ideas: (a) adjacent responsive fields, (b) linked loads, and (c) spatial subsampling. Every unit in a convolutional layer receives contributions from neighboring units with a place with the previous layer in light of the adjacent open field. In this way, neurons are well-suited to distinguishing primitive visual highlights such as edges and corners. The next convolutional layers connect these highlights to detect higher request highlights. Furthermore, the concept of linked loads implements the potential that simple component indications, which are helpful on a section of an image, will most likely be helpful across the entire picture. The concept of linked loads demands a group of units with indistinguishable burdens. The concept of linked loads mandates a group of units with indistinguishable burdens. A convolutional layer's units are solidly coordinated in planes. A plane's units all have a similar load arrangement. Along these lines, each plane is responsible for constructing a specific component, and these plane outputs are called Include maps. Because each convolutional layer comprises a few planes, several element guides may be created in each region.

A unit whose states are stored at comparing areas in the feature map examines the whole picture throughout constructing a feature map. This progression is identical to that of a convolution activity, with an additive bias term and sigmoid function shown in equation 2:

$$y(d) = \sigma(Wy(d-1) + b) \tag{2}$$

where d is the convolutional layer's depth, W denotes the weight matrix, and b denotes the bias term. The weight matrix is complete for fully connected neural networks, meaning it connects every input to every unit with distinct weights. However, due to linked weights, the weight matrix W for CNNs is relatively sparse. Therefore, W looks like shown in equation 3:

$$\begin{matrix} w & \cdots & 0 \\ \vdots & w & \vdots \\ 0 & \cdots & w \end{matrix} \tag{3}$$

where w is networked with comparable measures to the responsive fields of the units, using a sparse weight framework reduces the number of adjustable limits in an organization's tunable parameters and increases its generalization capacity. Convoluting the contribution with w, which can be seen as a trainable filter, is like duplicating W with layer by multiplying the inputs given in equation 4 and 5.

$$y_{(i,j)}^{(d)} = \sigma(x_{(i,j)}^{(d)} + b) \tag{4}$$

With

$$x_{(i,j)}^{(d)} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} w_{ab} y_{(i+\alpha)(j+b)}^{(d-1)} \tag{5}$$

The bias term is scalar in this case. The feature map for the corresponding plane is generated by successively applying (4) and (5) to all (i, j) input locations. CNN is the most used method, and it is highly effective based on the application applied to it.

In various computer vision applications, including image detection and scene segmentation, and Facial Expression Recognition. Considerable research in the Emotion Recognition literature finds that CNN is a good tool for Facial Expression Recognition after using various methods for FER. When faced with position shifts and scale variations, CNN outperforms multilayer perceptron (MLP), RNN, Deep Autoencoders, and DBN [134]. A standard CNN is made up of different layers: Convolutional, Pooling, and Fully Connected. Local connection and weight sharing are characteristics of CNN, which result in fewer network parameters, quicker training speed, and a regularization impact. Figure 21 is an example of a CNN-based FER technique.

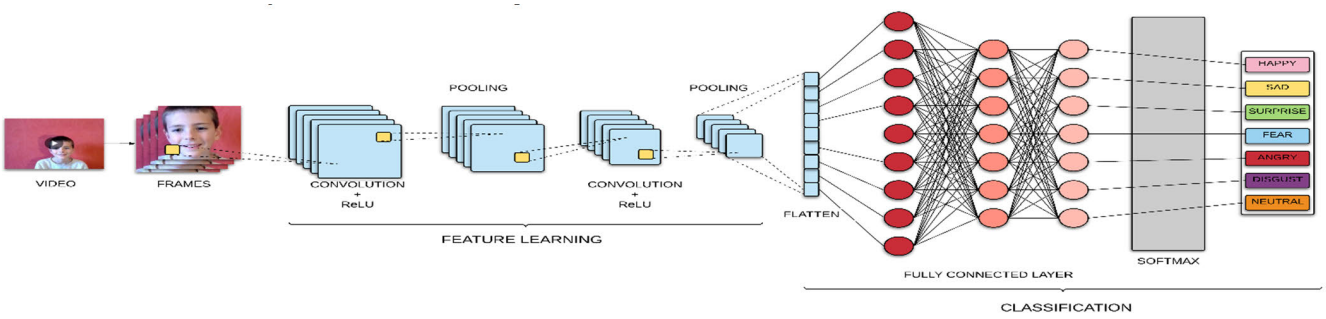

FIGURE 21. Convolution neural network architecture.

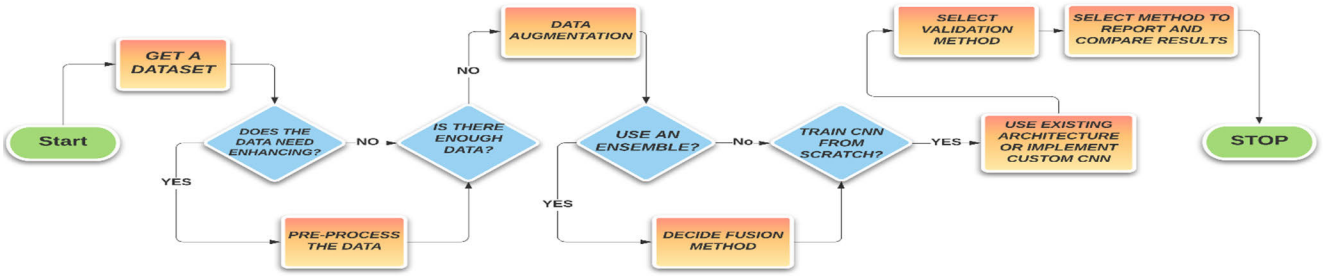

FIGURE 22. Flowchart illustrating the primary steps involved when implementing a CNN.

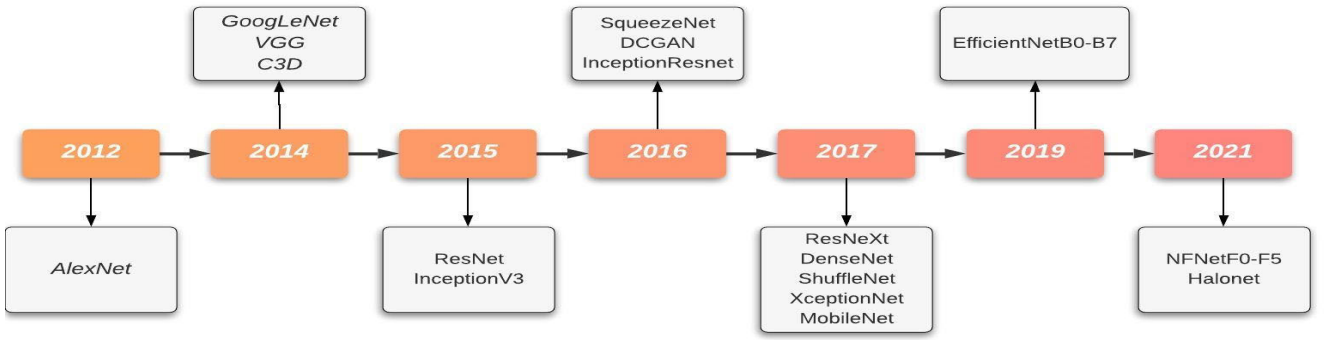

FIGURE 23. Timeline diagram of different CNN pre-trained model.

For dynamic emotion analysis, a study [135] suggested a deformable facial action components model. As a result, the 3D CNN includes a deformable facial parts learning module that can identify a specific facial action part under defined spatial constraints while also obtaining a representation based on the discriminating part. However, many standard methods make use of pre-trained models to achieve higher results. These pre-trained CNN models are easy to use and are efficiently deployable on a large-scale platform. To understand how CNN is applied, authors have illustrated an algorithm or a flowchart to apply CNN for emotion classification. This illustration is shown in Figure 22.Several pre-trained models are experimented with using novel approaches for FER, which significantly reduce computational power and increase accuracy, such as Transfer Learning.

With time new models are emerging which are more complex and yet easy to use deep learning in FER. To understand it further, the authors have shown a timeline of models that are being used in the field of FER, which is illustrated in Figure 23 in which authors have shown a timeline of different pre-trained models proposed from 2012 to 2021. AlexNet has eight layers, and it was one of the first models to win the ImageNet challenge [136]. VGG Nets are dense models with 3×3 convolutions stacked together throughout the whole model [137]. It makes the model very slow to train and hard to deploy for real-world conditions. GoogleNets and Inception models use 'Inception modules' to solve the problem of a deeper network [138]. They use filters of multiple sizes on the same layer, which makes the models wider than deeper. Even with inception modules, the networks are still deep. To solve deep layered networks, the 'residual module' was introduced

TABLE 5. Comparative analysis of CNN against different datasets.

Model	Layers	ImageNet Accuracy	Dataset	Accuracy	Features
AlexNet	5+3	84.7%	CK+ JAFFE FER2013	76.58% 54.46% 60.33%	The first model to win the ImageNet challenge.
GoogleNet	21+1	93.3%	RML SAVEE	81.4% 95.99%	Uses inception module to handle objects at multiple scales
VGG	13/16+3	92.7%	FER2013 CK+	69.4% 89.9%	VGG has multiple weight layers because of Small-size convolution filters. More layers lead to improved performance.
C3D	8+2	85.5%	CK+ EmotiW	74.38% 59.02%	C3D is commonly used as a video segment descriptor that can embed both motion and temporal characteristics. In addition, C3D features are used as the first step before further processing for various applications.
InceptionV3	48	94.2%	FER2013 CK+	74.04% 93.2%	Label Smoothing and factorized 7 x 7 convolutions are used in Inception V3. Label information is propagated lower down the network using an auxiliary classifier.
SqueezeNet	11+2	82.5%	FER2013 CK+ Oulu-CASIA	72.53% 98.7% 89.7%	Three times faster and 500 times smaller than AlexNet. Contains a fire module that consists of a squeeze convolution layer, feeding into an expanded layer.
ResNet	151+1	84.7%	FER2013	72.4%	There is a residual block that employs a technique known as skip connections. The residual block bypasses a few stages of training and links directly to the output.
DenseNet	200+1	93.7%	FER2013 VGAF AFEW	52.06% 64.75% 51.44%	Each layer's feature map is concatenated with the next layer's input within a dense block. This enables higher levels of the network to use the features of lower levels directly. In addition, it allows for network-wide feature reuse.
EfficientNet B0-B7	237-813	93.2% - 97.1%	AffectNet AFEW	65.75% 59.26%	To scale all depth, breadth, and resolution dimensions consistently, a compound coefficient is utilized. Using a set of predefined scaling coefficients, the scaling approach equally scales network width, depth, and resolution

in ResNets [139]. The residual module uses skip connections to skip layers in between, which solves the vanishing gradient problem. With the growing trend of 'inception modules' and 'residual modules,' both were combined in Inception Resnets [140]. It makes the network very computationally efficient. Another CNN called SqueezeNets is 50× smaller than AlexNets while achieving the same if not better accuracy [141]. It consists of 'Fire Modules' containing squeeze (1 × 1 filters) and expand (1 × 1 & 3 × 3) layers. Another CNN called DenseNets uses concatenation, i.e., the next layers take inputs from previous layers to pass on their feature maps [142]. XceptionNets adapts from InceptionNets, but the inception modules are replaced with depthwise separable convolutions and take the inception hypothesis to an extreme [143]. The ResNeXts are like ResNets but with the addition and scaling of parallel towers (cardinality) within each module [144] but, XceptionNets and ResNeXts cannot take 1 × 1 convolutions (pointwise convolutions) without hindering the accuracy. The next two models solve this problem. ShuffleNets and MobileNets are designed for mobile devices [145], [146]. The shuffle unit consists of pointwise convolutions with channel shuffle, which makes ShuffleNet computationally efficient. The same pointwise convolutions technique is used in MobileNets with a slight change [146]. The pointwise convolutions are applied in depthwise separable convolutions, which drastically reduced the computation and model size. The EfficientNets use the compound scaling method to increase accuracy and efficiency [147]. Instead of scaling individual dimensions, it balances all network dimensions like width, depth, and image resolution. NFNets are 8.7× faster than EfficientNets, with the base model (F0) achieving the accuracy of the top-of-the-line B7 [147]. NFNets use modified residual branches and convolutions and adaptive gradient clipping to achieve state-of-the-art accuracy [148]. On the other hand, C3D capturing motion information is incorporated in several adjacent video frames and is usually preferred when analyzing videos for classification [149]. To understand how these CNN have performed with publicly available datasets, authors have shown Table 5, representing a comparative analysis of the highest performance of CNN against different datasets.

3) DEEP AUTOENCODER (DAE)

Deep autoencoder is similar to deep neural networks, which was first introduced in [121]. It is used to reproduce the input dataset at the output. This means that the number of neurons at the input is the same as the output. It encodes the information x into a representation r(x), allowing information to be regenerated from r(x) [150]. In this way, the autoencoder's goal yield equals the autoencoder's input. As a result, the yield vectors have a dimensionality similar to the information vector. The remaking blunder is limited throughout this cycle, and the associated code is the learned feature component.

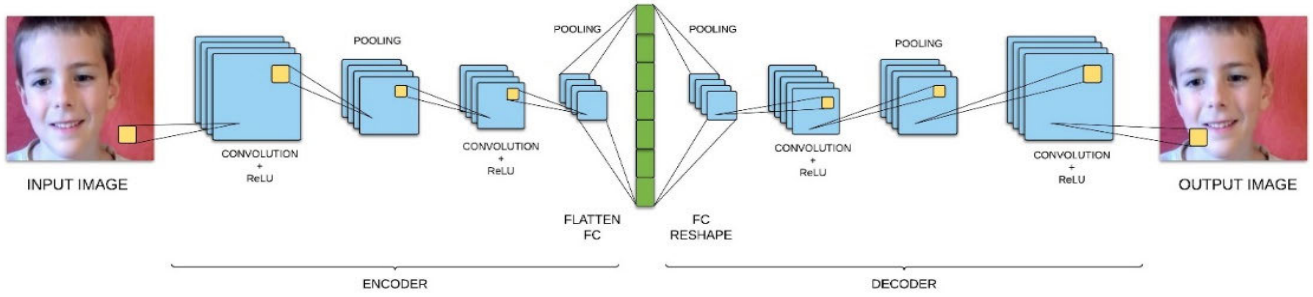

FIGURE 24. Deep auto encoder.

Suppose the network is prepared using the mean squared blunder model, and there is a hidden layer. In that case, the k hidden units determine how to expand the contribution to the range of the major k head portions of the information [151]. If the hidden layer is nonlinear, the autoencoder behaves differently from PCA, allowing it to capture multimodal portions of the information transmission [152]. The model's parameters are being improved to reduce the likelihood of recreating errors. There are other ways to assess the remaking error, including the conventional squared blunder:

$$L = \|x - f(r(x))\|2 \tag{6}$$

where f(r(x)) is the reconstruction produced by the model and f is the decoder. The loss function of the reconstruction might be expressed as cross-entropy if the input is represented as bit vectors or vectors of bit probabilities which is illustrated in the following formula:

$$L- = \sum_i x_i \log f_i\,(r\,(x)) + (1 - x_i)\,log(1 - f_i r(x))) \tag{7}$$

R (x) can't successfully compress all input x since it is not lossless. The optimization approach yields low reconstruction error on test instances from the same distribution that can collect the locations along the data's significant fluctuations as the training examples. Still, high reconstruction error on samples picked randomly from the input space.

In short, one can summarize that the job of the autoencoder is to produce the compressed version of the input image with low data loss. On the other hand, the encoder's job is to break down the input image into a compressed version.

The job of the encoder is to break down the input image into a compressed version. As a result, the overall size of the data is reduced, excluding the important parts with minimal data loss. This is called Dimensionality reduction. The structure of an autoencoder is as follows:

- Encoder: A feed-forward, fully connected neural network is referred to as an encoder. It is used to compress the input image and reduce the size. The altered form of the original image is the compressed picture.
- Decoder: It is also a feed-forward network. This network is in charge of reassembling the input from the code to its original dimensions.

It is optimized to rebuild by reducing the rebuilding error of its inputs instead of the previously discussed networks, which are taught to anticipate goal values. The denoising autoencoder [153] recovers the original undistorted input from partially corrupted data; the sparse autoencoder network (DSAE) [154] imposes sparsity on the learned feature representation; and the contractive autoencoder [155], convolutional autoencoder [156], which uses CNNs convolutional layer (pooling is optional) layers for the hidden layers in the network; and the variational auto-encoder [157], which is a directed graphical model with certain types of latent variables to design complex generative models of data. There is also given a Figure 24 which is an illustration of DAE.

4) RECURRENT NEURAL NETWORK (RNN)

RNN is a model that incorporates temporal information and is better suited to predicting sequential data of arbitrary durations. RNNs have recurrent edges that span neighboring time steps and share the same parameters across all steps, in addition to training the deep neural network in a single feed-forward. The RNN is built using the standard back proliferation through time (BPTT) method [158], and its modules have a chain-like structure consisting of four repeating modules, as illustrated in Figure 25. Long-short term memory (LSTM), illustrated in [159], is a type of conventional RNN that is used to identify the inclination fading and blast problems that often occur while creating RNNs wherein its cell state is regulated and controlled by three gates in LSTM:

1. Input gate - It permits or prevents the input signal from changing the cell state.
2. Output gate - It allows or inhibits the state of a cell from affecting the state of other neurons.
3. Forget gate alters the cell's self-recurrent connection, allowing it to remember or accumulate its prior state.

LSTM can simulate long-term dependencies in a sequence by combining these three gates, and it has been widely used for video-based expression recognition applications. In FER, there have been many usages of RNN/LSTM; in a recent study [160], LSTMs were used. A recurrent network was used to consider the temporal dependencies in the image sequences during classification. Furthermore, experimental results involving two types of LSTMs (bidirectional and unidirectional) were also used. This study found that bidirectional networks perform significantly better

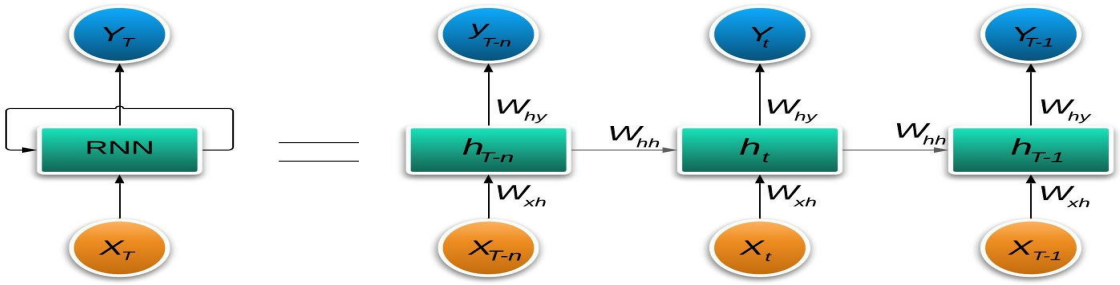

FIGURE 25. Recurrent neural network architecture.

TABLE 6. Comparison of different techniques used on various FER datasets of adults.

Dataset	Reference	Preprocessing	Feature Extraction	Classification model/method	Recognition Rate
FER2013	[32]	--	Gabor + LP	Ensemble classifier	88.9%
	[33]	Haar like + Adaboost	PCA + LBP	SVM	89.64%
	[34]	AAM	Gabor transform	Adaboost + Dynamic Bayesian Network	94.05%
	[35]	✓	Structured Streaming Skeleton (SSS)	K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM)	90.33% and 67%, respectively
	[36]	--	Constrained Local Model (CLM)	SVM	82.03%
	[37]	Viola-Jones	Curvelet transform	Online sequential ML learning with radius bias function RBF	95.17%
CK+	[38]	Viola-Jones	HOG	3 stage SVM	93.29%
	[39]	ViolaJones	HOG	SVM	88.70%
	[40]	AAM + Intraface	--	DNN	93.20%
	[41]	Interface	pretrained CNN used as an extractor	CNN	98.47%
	[42]	AAM	HOG + PCA	DAE (DSAE)	89.84% - 95.79%
	[43]	OpenCV Haar	--	Deep Belief Nets	20%
JAFFEE	[38]	Viola-Jones	Boosted LBP	3 Stage SVM	93.29%
	[44]	--	Patch Based + Gabor	SVM	92.30%
	[44]	--	HOG	SVM	94.30%
	[39]	ViolaJones	LDA	SVM	95.71%
	[45]	--	patch-based	Boosted DBN	91.8%
	[46]	--	CAE + Gabor	CNN, CAE (Network Ensemble)	95.8%
	[47]	--	--	RAU CNN	59.25% 86.38%
RFD	[38]	Viola-Jones	HOG	3 Stage SVM	99.72%
	[48]	--	HOG	SVM	98.50%
	[49]	Haar Feature-Cascaded Classifier	--	CNN	71%
NVIE	[50]	✓	recursive nonparametric discriminant analysis (RNDA)	KNN +Adaboost	76.7%
	[51]	AAM	IR onset	SVM	76.82%
eNTERFACE'05	[52]	Fera2015	PCA	SVM CNN FUSION	79% 77% 99%

than unidirectional LSTMs. Alternately by using multi-angle-based optimal configurations, a study [161] proposed a multi-angle optimal pattern-based deep learning (MAOP-DL) method to correct the problem of sudden changes in illumination and find the right alignment of the feature set. In this approach, the background is initially removed, and the foreground is focused on, and then the texture patterns and the relevant facial features are extracted. Finally, the

TABLE 7. Comparison of different techniques used on various FER datasets of kids.

Dataset	Reference	Preprocessing	Feature Extraction	Classification model/method	Recognition Rate
LIRIS	[18]	--	CNN	VGG	75%
DDCF	[53]	Viola-Jones	pyramid local binary pattern	SVM MLP Random Forest	68.4% 63.3% 61.5%
CAFE	[54]	EMGUCV	--	CNN-AFFDEX	44.88%
NIMH-Chefs	[55]	✓	Anthropometric Model	SVM	47%
EmoReact	[54]	EMGUCV	--	CNN-AFFDEX	55.76%

TABLE 8. Comparison of different techniques used on various FER datasets of senior citizens.

Dataset	Reference	Preprocessing	Feature Extraction	Classification model/method	Recognition Rate
Tsinghua facial expression database	Note: NO RESEARCH WORK WAS DONE USING THESE DATASETS	N/A	N/A	N/A	N/A
database for emotional interactions of the elderly		N/A	N/A	N/A	N/A
FACES					

relevant features are selected to predict the correct facial expression label, and an LSTM-CNN analysis is performed. However, in the case of videos, there has been the usage of 3D Inception-ResNet architecture and an LSTM unit to extract the spatial relations and timing relations within the facial images from different frames of a video sequence, which was proposed in a study [162] where they also studied the effects it has on viewing the facial images with different frames.

Table 6, Table 7, and Table 8 bring forth a brief comparison of techniques used in different categories of a dataset (i.e., Kids, Adults, and Senior Citizen) divided into three stages of FER (i.e., processing, Feature extraction, and classification) along with their recognition rate.

VI. RESEARCH CHALLENGES AND OPEN ISSUES

FER has been a competitive subject in recent years. Numerous studies have yielded excellent results and correctly identified emotions during facial expression recognition analysis. However, many problems and concerns must be tackled. In this section, the authors go over some of the issues and challenges that FER has faced. The authors studied various survey papers to identify the challenges effectively and suggested plausible solutions.

A. OCCLUSION AND DATA COLLECTION WITH OCCLUSION

On FER, the most common stumbling block is occlusion. The authors found that current research is already publicly available, such as JAFFE, CK+ datasets without occlusion. There is a scarcity of natural facial occlusion in many datasets. There is a need to create datasets that have occlusion. Although it is usually time-consuming and a difficult task to do but it is a necessary evil. FER datasets should be created by decisive manual occlusion. There has been no worthy training, and testing in many occluded datasets remains a significant obstacle [182]. Also, on the other hand, the collection of spontaneous datasets of emotion under occlusion is a hectic process. The selection of the impeded region, the occlusion level, its type, and preparatory materials pose a significant challenge to create such datasets in the first place. Happiness, surprise, and sadness are all easily elicited, but attentiveness and curiosity are two emotions that are particularly difficult to elicit, especially under occlusion. There is a need to consider strategies that instigate accuracy and are provisionally dependent [167], [183].

Plausible Solution: The raw pixel values of the occluded region may be used to overcome the dataset construction problem, but adequate data on some facial area features in the image might not be captured. Detecting essential factors such as materials, locations, and components is essential for Facial Emotion Recognition. To detect occlusion, one can use a pre-processing layer to improve accuracy [159].

B. DISTRIBUTION OF DATASET BIAS AND IMBALANCES

Another challenge that one can face on FER is the scarcity of an excellent illustrative dataset for training in good quality and quantity. There is a huge imbalance in datasets in FER such as gender, age, face color, and cross-cultural imbalance. Also, most datasets have images/videos of a specific range of age groups, but not of all age groups, including children and senior citizens [29]. Because of the inconsistent Facial Emotional datasets imbalance, FER performance cannot improve consistently over time, even by directly increasing the dataset for training by joining multiple datasets [184].

Plausible Solution: To create a dataset that has good quality and quantity of FER that has no imbalance of data and that has sufficient data on all parameters of age, gender,

face color, and cross-cultural imbalance. Developing such a dataset would help in developing research on diverse FER. Alternatively, the authors suggest that one could balance the class with the training dataset class distribution. At the same time, the pre-processing phase uses techniques such as data augmentation, splitting, and synthesis of data from these components.

C. FER ON 3D DATA

Today's existing works on FER on 2D data are usually the main focus which poses certain obstacles to parameters such as variable pose [63] and illumination. But in 3D facial shape models are robust to these factors. This 3D dataset contains depth images and videos, recorded with the relative intensity of face pixels as per the distance of the depth camera from the face, containing important facial - geometric relations information. A Kinect depth sensor is a great example [163] that obtains gradient direction data and uses CNN on an unregistered depth image for Emotion Recognition from the face. Many works have recently proposed merging the two-dimensional and three-dimensional data to increase the model's performance.

Plausible Solution: To encourage more 3D FER datasets to create new innovative research papers on it. Additionally, one could begin exploring the 4D FER by examining the existing dynamic deformation patterns commonly seen on datasets of Emotions to increase the existing dynamic deformation patterns that are typically seen on datasets of Emotions.

D. VARIABLE MODALITIES IN FER

Humans can only recognize Facial Emotion modality that can be used to understand another human's behavior. But there are many combinations of other patterns that are usually unimportant for a human's naked eyes, although it is still an essential factor for FER. These combinations are infrared images, data captured by 3D models, and physiological data is now an emerging hotspot research area that further enhances the robustness.

Plausible Solution: To encourage the creation of new multimodal databases which include not only audio modalities but also infrared and 3D data so that future research will show more robustness in results and can be applied in real-life applications, which will become a potential direction for upcoming research because of the immense appreciation for facial emotions by AI.

E. FER ON INFRARED DATA

There is an immense trend of using grayscale and RGB data in deep FER, but this poses a more sensitive challenge to light. But, on the other hand, infrared images record these expressive facial emotions created by the skin distribution of the face, which are not sensitive to changing illuminations. For example, in [161], a DBM model containing Gaussian-binary RBM and binary RBM was trained using layer-wise pretraining and joint training on long-wavelength thermal infrared images to learn thermal characteristics. In addition, the author presented a three-stream 3D CNN to combine local and global Spatio-temporal characteristics on illumination-invariant near-infrared pictures for FER.

Plausible Solution: There are very few infrared databases available on adults as well as on children. This opens the opportunity to create all age-based infrared datasets showing emotions. Also, to do FER on such datasets in the study [161] where for FER, a three-stream 3D CNN is proposed to that on illumination-invariant near-infrared images, combine local and global Spatio-temporal characteristics.

F. UNAVAILABILITY OF DATASETS

There are multiple datasets available for the emotion detection of Adults. Most of them are pre-processed. For Kid's and Senior Citizen's emotion detection specifically, very few datasets are available, and most of them are not pre-processed. To get decent accuracy, the requirements for the dataset are high. Datasets like LIRIS have high-quality video clips for training, unlike other datasets like NIMH and DEFSS with high-quality images but do not contain very high numbers of videos/images for more generalization. Also, datasets like FACES, a mixture of young and senior citizen, did not have high numbers of images, and there is only one good enough dataset called Database of Interactions of Elderly, which can be used for FER. Still, there is no reasonable alternative as compared to adult datasets.

Plausible Solution: To create a new Kid's and Senior Citizen people facial emotions dataset with big data that covers all parameters of balanced gender, age, face color, and ethnic backgrounds with all modalities such as grayscale, RGB, audio, infrared, and 3D data since there is a big scarcity of these categories of which are readily available for adult FER datasets.

G. OTHER ISSUES

With new advancements in computer vision, many novel issues have caught attention. The prototypical basis of recognizing dominant and complementary emotions is a difficult task illustrated in [162] and the Challenges of Genuine vs. Fake Expression of Emotions [162]. Another challenge is to build a real-time emotion detection system of all the age groups with different cross-culture and ethnic backgrounds.

Plausible Solution: To focus on new novel issues and build an open-source global dynamic emotion recognition system.

VII. FUTURE DIRECTIONS

A. SUPER-RESOLUTION

Image super-resolution has piqued the curiosity of the scholarly community in recent years. Its purpose is to convert a low-resolution image into a high-resolution image with superior visual quality and detail than the original coarse feature. An image's ''reduced resolution'' can result from a lower spatial resolution/smaller size or degradation such

FIGURE 26. Super-resolution on faces.

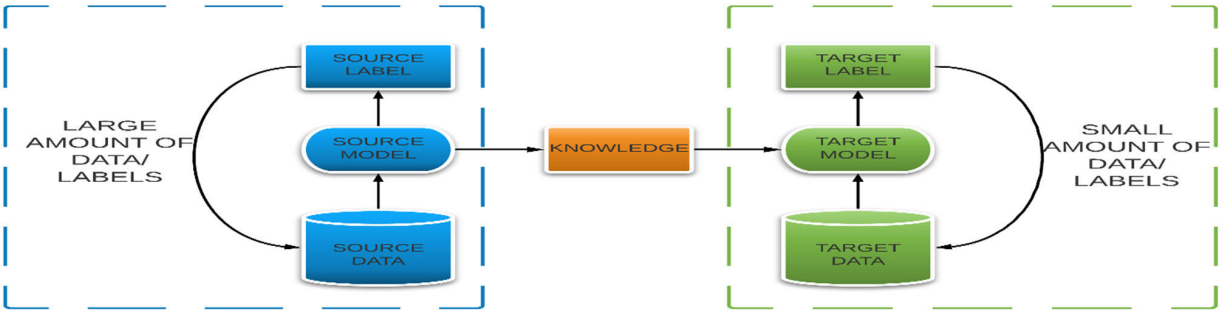

FIGURE 27. Transfer learning in FER.

as blurring. To connect the High Resolution (HR) and Low Resolution(LR) pictures, apply the following equation:

The following formula can be used to model low-resolution images from high-resolution photographs: D stands for degradation function, Iy for high-resolution image, Ix for a low-resolution image, and noise.

$$I_x = D\left(I_y; \alpha\right) \tag{8}$$

This formula can be used to model low-resolution images from high-resolution photographs: D stands for degradation function, I_y for high-resolution image, I_(x) for a low-resolution image, and noise. Usually, only the high-resolution image and its equivalent low-resolution image are provided because the degradation parameters D are unknown. Using only the HR and LR image data, the neural network must learn the inverse deterioration function and show the output, Figure 26.

Super Resolution in Emotion Detection: Super Resolution (SR) approaches usually outperform classical algorithms like nearest-neighbor interpolation, bilinear, and bicubic in tackling the problem of tiny image size or blurriness. While it is straightforward to downscale a high-resolution image to a low-resolution one, the reverse is difficult. Low-resolution pixels that have gone missing must be retrieved. In a recent study [185], the Super-Resolution Convolutional Neural Network (SRCNN) was mentioned in its literature; a deep CNN model acts on low-resolution and high-resolution feature maps and outputs a high-resolution image. In short, it can be summarized as a simple interpolation technique that outperforms bicubic interpolation. Very Deep Super Resolution (VDSR) was also mentioned in this literature, which is built similarly to the SRCNN. However, it is more in-depth as compared to SRCNN. Like SRCNN, different techniques can be used to achieve super-resolution, e.g., ESPCN and EDSR.

B. TRANSFER LEARNING

Transfer learning is known as transfer learning by using the weights of a model trained on another dataset on a new different dataset [186]. Its commonly heard in the field of FER since it can be used on a minimum dataset. This is extremely important because training the model on every dataset, which amounts to millions, would bring many inefficiencies, so transferring learned features onto another would be the most appropriate solution. Instead of starting, the authors suggest using patterns acquired from completing a comparable task and applying them to new data, which is also illustrated in Figure 27, which shows the working of Transfer Learning in FER.

Transfer Learning in Emotion Detection: To approach transfer learning on emotion recognition, there is a requirement to obtain high-level features using CNN, which is trained on huge datasets (e.g., [187]–[190]). The originally trained datasets might not necessarily contain the same labeled classes compared to target classes on the target dataset, different from the initial model trained upon. An occluded dataset is also used, which is common and realistic in daily life, which is utilized to increase the generalization and robustness in the proposed study [191].

C. DOMAIN ADAPTATION

Domain adaptation is a branch of machine learning that deals with situations when a model trained on one distribution is applied to a different (similar) target distribution [192]. Domain adaptation is a technique for solving new issues in

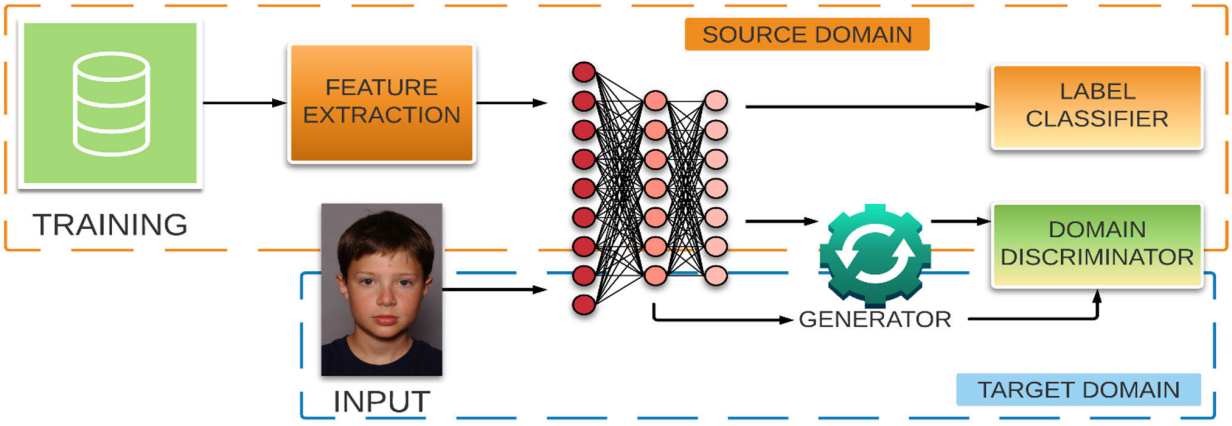

FIGURE 28. Working of domain adaptation in FER.

a target domain using labeled data from one or more source domains. This is also evident in the working of Domain adaptation definition, as shown in Figure 28. In this, the degree of similarity between the source and target domains impacts the success of the adaptation in most cases. When the task space is the same, and the only change is input domain divergence, domain adaptation can be an issue.

Domain Adaptation in Emotion Detection: In a new study [193], a novel approach of domain adaptation methodology has been proposed to recognize facial emotions from the fusion of facial, non-facial, and non-human components. The proposed system is predicted using an intersection score. It also suggested using pre-trained face emotion recognition models using Attentional CNN. The experiments were executed on the Flickr image dataset, categorized into basic emotions (e.g., angry, happy, sad, and neutral), which showed an accuracy level of 63.87% for emotion recognition which outperformed the benchmark results. Alternately another study [194] proposed an approach where only unlabeled target-specific data is only needed. Finally, the recent study [195] also proposed a regression framework to lean parametric of the classifier and user-specific sample distribution.

D. ADVERSARIAL MACHINE LEARNING

In Adversarial Machine Learning(AML), adversaries are malicious inputs that are purposely designed to make sure the model fails to predict the right labels [196]. Adversaries disrupt the way the model usually predicts so that a real-life error-filled scenario can be recreated to prepare for that or find a new way to avoid such things. In recent years, adversarial machine learning is becoming a crucial part of any computer vision-related task, whether FER, activity recognition, or object detection. AML is divided into three types of adversarial attack, illustrated in Figure 29, which shows the working of AML.

Adversarial Machine Learning in Emotion Detection: In a recent study [197], the adversarial approach was proposed claiming to provide anonymity to individual subjects on which are doing emotion recognition which will be a crucial key point in real-life applications by achieving the highest accuracy and security simultaneously by applying convolutional transformation that will try to degrade individual-specific data for any subsequent fully connected layers. Its output is then passed to two classifiers for the detection of emotions and recognition. Such that emotion-related data and computed identity data are preserved in CNN.

E. ZERO-SHOT LEARNING

Zero-shot machine learning is used to recognize unseen target classes at test time [198], even though that test label is not observed even in training times which is evident in Figure 30, where authors illustrated working of zero-shot in FER.

The data in zero-shot learning consist of the following points:

1. Seen classes: During training, label the photos for some classes.
2. Unseen classes: During the training period, there are no tagged photos for these classes
3. Auxiliary information: At train time, this data contains descriptions, semantic characteristics, and word embeddings for both visible and unseen classes. This data serves as a link between visible and invisible classes.

Zero-Shot Learning in Emotion Detection: In a recent study [199], it has been proposed to use generalized zero-shot learning (GZSL) for emotion recognition. It consists of 3 branches: the first is a Prototype-Based Detector (PBD), which predicts unseen gesture categories from learned data; the second is a stacked autoencoder used for classification. The third branch enhances generalization recognition of emotions.

F. REINFORCEMENT LEARNING

Deep Reinforcement Learning (DRL) is a program that can learn on its own to solve complex problems, with deep neural networks representing the information [200]. The learner is an AI agent who solves a specific task by interacting with

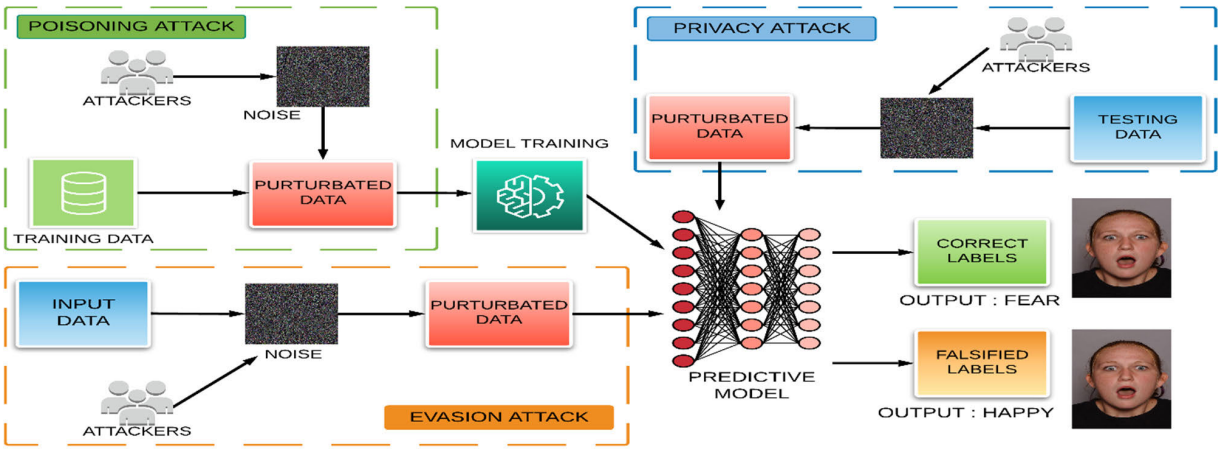

FIGURE 29. Working of adversarial machine learning in FER.

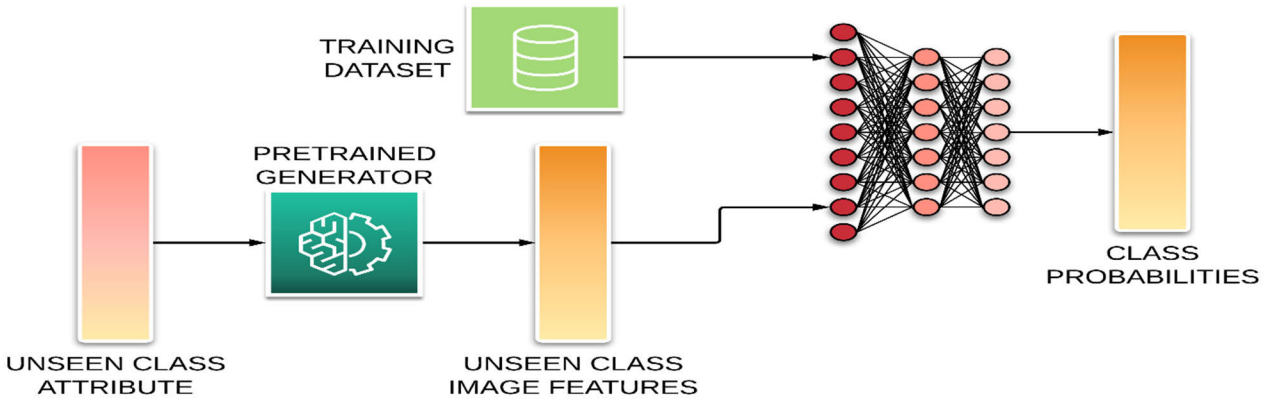

FIGURE 30. Zero-shot working in FER.

its environment in Reinforcement Learning (RL). By doing actions and observing their outcomes, the agent learns how to behave in a given environment (rewards). Reinforcement Learning is based on the premise that the agent learns from the environment and gets rewarded based on interaction. The agent acts in each state and then moves on to the next, earning a reward. There are three stages in all which state, action and reward.

Compared to human perception, face detection and emotion classification are two essential parts of computer vision when creating a vision system. When developing a face recognition system for an AI agent, one should aim to detect and recognize faces and emotions before categorizing them. RL can learn unique emotions that differ from person to person and optimize itself again, which can be an excellent combination in FER. This combination is illustrated in Figure 31, where the authors proposed architecture that can be used in FER.

G. FEDERATED MACHINE LEARNING (FL)

It is a new machine learning method in which the algorithm is dispersed among numerous distributed edge devices or servers that store sample data locally and do not exchange them [201]. This strategy differs from the commonly used centralized machine learning algorithms, which need all local datasets to be uploaded to a single server. Federated learning solves fundamental challenges such as privacy, security, access rights, and access to heterogeneous data by requiring numerous actors to work together to provide a common, robust learning model without sharing data. FL enables the model to gain more experiences from a broad range of data sets located at different geographical locations without any security concerns. This learning model enables multiple organizations such as pharmaceutical, defense, space, heavy machinery manufacturers, and healthcare to develop a faster, distributed, and reliable model without worrying about computation or security concerns. It can also be used in the field of FER, which is illustrated in Figure 32.

Federated Learning in Emotions Detection: A recent study demonstrated [201] feature extraction approaches for extracting features from both images and audio. Using collected face and speech information, the proposed approach detects human emotions. The output is generated by both classifiers on an individual's categorical emotions. The accuracy of

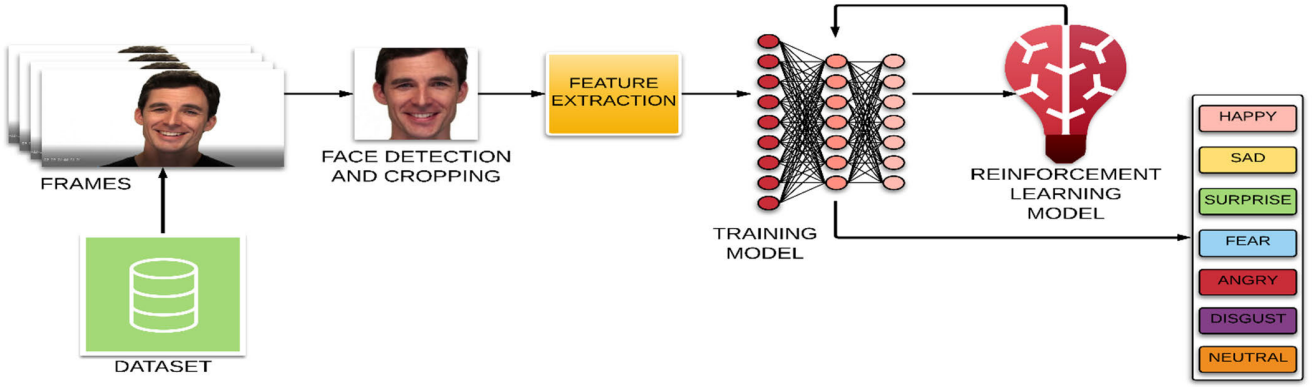

FIGURE 31. Proposed architecture for reinforcement learning in emotion detection.

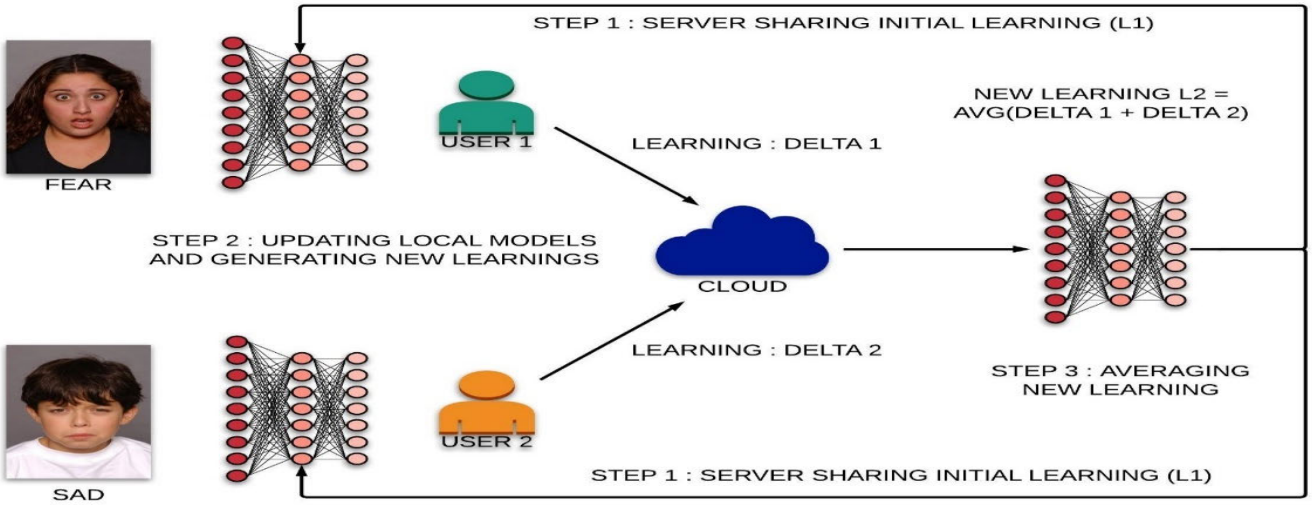

FIGURE 32. Federated learning for emotion detection in FER.

the suggested face and speech emotion detection classifiers is 71.64% and 85.04%, respectively. The result suggests whether a person needs to get counseled by an expert such as a psychologist.

H. EXPLAINABLE AI

Explainable AI is an advanced artificial intelligence concept followed by easily comprehensible reasoning for how it arrived at a given conclusion [202]. Whether via pre-emptive layout or retrospective analysis. These strategies are currently being hired to make the black field of AI less opaque and make models more reasonable and trustworthy for a satisfying reason. Humans are the best judge to classify any human emotion and explain every emotion. Still, in the case of AI, it shows the output from what it has learned without explaining such output, e.g., in the field of medical image analysis, AI can predict whether a person has pneumonia or not just by looking Xray. However, in the end, it will still not be trusted because it doesn't give any explanation, which is crucial, suggesting to take the opinion of a doctor to announce the final results. In such Cases, Explainable AI, will give output and explain its result, far more reliable than previous AI models. This can also be applied in FER, shown in Figure 33, which shows the working of Explainable AI in FER.

VIII. FACIAL EMOTION RECOGNITION POTENTIAL APPLICATIONS

Facial emotions are the result of the movement of muscles beneath the skin. They are predominant channels of conveying social information between individuals. Facial expression analysis provides objective and real-time information about how people's faces intimate emotional content. FER has a wide range of applications spread across medicine, e-learning, monitoring, entertainment, law, etc.

The use of FER in each of the mentioned fields is as discussed below:

A. E-LEARNING

In e-learning, instructors assess students' capacity to comprehend topics by watching emotions and adapting the teaching approach and presentation to the learner's preferred style. This contributes to developing a more robust educational

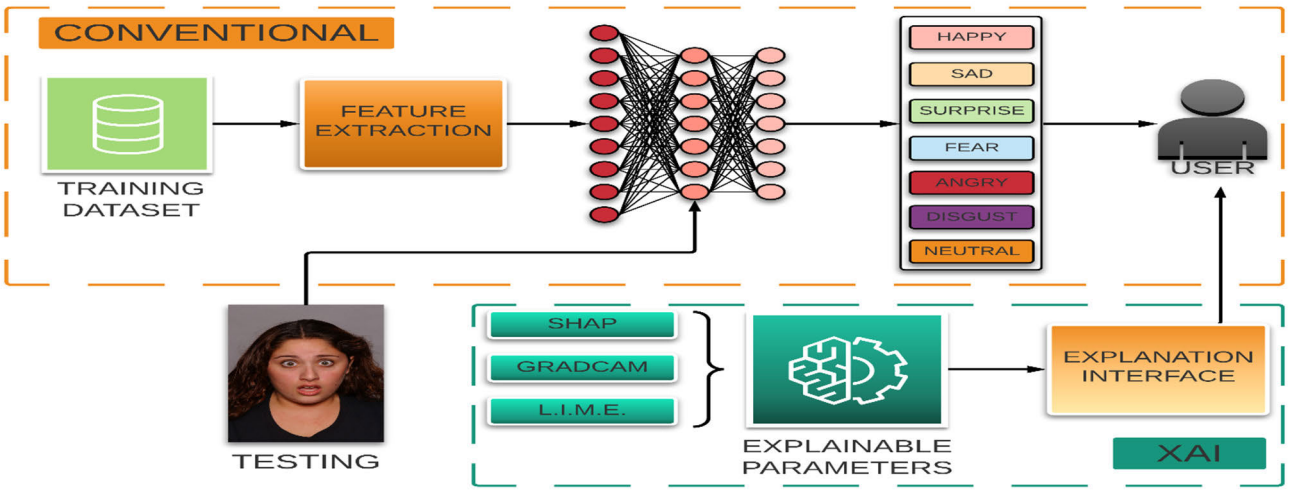

FIGURE 33. Working of explainable AI in FER.

system, from which students benefit greatly, whether through remote learning or otherwise.

B. MONITORING

Emotions have an essential role in safe driving, according to psychological studies. The emotional state of the driver influences the driver's comfort and safety when operating a vehicle. According to the psychological study, anger, despair, and fear lead to reckless and fast driving. Anger, aggressiveness, tiredness, and stress can all raise the likelihood of an accident. Nervousness and melancholy, when present at the same time, may have an impact on driving. As a result, it is self-evident that if there is a FER system that continually monitors driving expressions and identifies them and if they fall into one of the categories mentioned, the system may warn the driver and prevent accidents. FER plays a critical part in police operations by analyzing an individual's facial expressions to determine whether or not that person is scared when withdrawing cash from an ATM.

It then devises a plan to halt cash distribution. Customer preferences and satisfaction may be tracked and evaluated using a FER tool installed in retail stores, which offers data that can be examined to improve the user's shopping experience [203].

C. MEDICINE

When a patient lives in a remote area, is too unwell to travel, or is too elderly to travel, the distance might constitute a barrier for patient check-up appointments. To avoid a gap in medicinal therapy, FER systems can be a solution. The decreased capacity to comprehend faces found in autistic children explains their difficulties during social interactions. Building a FER application on a mobile phone and giving it to autistic youngsters will assist them in detecting facial expressions. Labeling them with an emoji will assist such children when they struggle to comprehend the sentiments of other individuals [204].

D. ENTERTAINMENT

In video games, asserting user experience [205] in real-time aids creators in emotionally attaching players to the game. The authors express a need to monitor and analyze facial expressions in real-time to determine whether a game successfully makes the user experience pleasurable. This aids the developer in developing a more effective solution.

IX. CONCLUSION

Since FER has been catching wide attention in the researchers' community, and less research with a 360-degree overview of this domain is found currently, the paper attempted to present all important aspects related to FER. The authors presented a brief review of methods and state-of-the-art models used in FER for different dataset categories. This paper analyzed all of the existing surveys done in FER, gained insights and knowledge about what they lack, and covered all the low points. FER datasets are categorized into three parts: Kids, Adults, and Senior Citizen people to understand the vast outreach in FER. From the dataset comparison analysis, creating a new database of Kids is a current thrust area since there is a scarcity of well-balanced datasets as of today. This paper also discusses different stages of FER such as pre-processing, feature extraction, and classification using various methods and state-of-the-art CNN models. It also compares different CNN models and their benchmark accuracy with some architectural details, which will help model selection based on the application or dataset on which it will be used. It also presents a database category-wise research survey to understand the similarities and differences among them with potential insights into future work that can be done.

Furthermore, this paper also discusses Open issues and challenges and suggests possible solutions to solve them. It also presents the upcoming trends in FER, which are currently being studied or yet to be done. As this is a hot yet challenging research domain, it comes with many more

potential applications which could be explored more with its developments.

REFERENCES

[1] A. Mehrabian and S. R. Ferris, "Inference of attitudes from nonverbal communication in two channels," *J. Consulting Psychol.*, vol. 31, no. 3, pp. 248–252, 1967, doi: 10.1037/h0024648.

[2] M. A. Lumley, J. L. Cohen, G. S. Borszcz, A. Cano, A. M. Radcliffe, L. S. Porter, H. Schubiner, and F. J. Keefe, "Pain and emotion: A biopsychosocial review of recent research," *J. Clin. Psychol.*, vol. 67, no. 9, p. 942, Sep. 2011, doi: 10.1002/JCLP.20816.

[3] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, Jan. 2020, doi: 10.3390/S20030592.

[4] J. Azcarraga and M. T. Suarez, "Recognizing student emotions using brainwaves and mouse behavior data," *Int. J. Distance Educ. Technol.*, vol. 11, no. 2, pp. 1–15, Apr. 2013, doi: 10.4018/jdet.2013040101.

[5] J. A. R. Eliot and A. Hirumi, "Emotion theory in education research practice: An interdisciplinary critical literature review," *Educ. Technol. Res. Develop.*, vol. 67, no. 5, pp. 1065–1084, Feb. 2019, doi: 10.1007/S11423-018-09642-3.

[6] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 6, pp. 2408–2412, Feb. 2010, doi: 10.1073/PNAS.0908239106.

[7] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Frontiers Robot. AI*, vol. 7, p. 145, Dec. 2020, doi: 10.3389/FROBT.2020.532279.

[8] S. Ramis, J. M. Buades, and F. J. Perales, "Using a social robot to evaluate facial expressions in the wild," *Sensors*, vol. 20, no. 23, pp. 1–24, 2020, doi: 10.3390/s20236716.

[9] Y. K. Bhatti, A. Jamil, N. Nida, M. H. Yousaf, S. Viriri, and S. A. Velastin, "Facial expression recognition of instructor using deep features and extreme learning machine," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–17, Apr. 2021, doi: 10.1155/2021/5570870.

[10] E. Hudlicka, "Computational modeling of cognition–emotion interactions: Theoretical and practical relevance for behavioral healthcare," in *Emotions and Affect in Human Factors and Human-Computer Interaction*. New York, NY, USA: Academic, 2017, pp. 383–436, doi: 10.1016/B978-0-12-801851-4.00016-1.

[11] G. Hemalatha and C. P. Sumathi, "A study of techniques for facial detection and expression classification," *Int. J. Comput. Sci. Eng. Surv.*, vol. 5, no. 2, pp. 27–37, Apr. 2014, doi: 10.5121/IJCSES.2014.5203.

[12] D. Deodhare, "Facial expressions to emotions: A study of computational paradigms for facial emotion recognition," in *Understanding Facial Expressions in Communication*. New Delhi, India: Springer, 2015, pp. 173–198, doi: 10.1007/978-81-322-1934-7_9.

[13] U. Asad, N. Kashyap, and S. N. Singh, "Recent advancements in facial expression recognition systems: A survey," in *Proc. Int. Conf. Comput., Commun. Automat. (ICCCA)*, May 2017, pp. 1203–1208, doi: 10.1109/CCAA.2017.8229981.

[14] D. Mehta, M. Siddiqui, and A. Javaid, "Facial emotion recognition: A survey and real-world user experiences in mixed reality," *Sensors*, vol. 18, no. 2, p. 416, Feb. 2018, doi: 10.3390/S18020416.

[15] K. Chengeta and S. Viriri, "A survey on facial recognition based on local directional and local binary patterns," in *Proc. Conf. Inf. Commun. Technol. Soc. (ICTAS)*, Mar. 2018, pp. 1–6, doi: 10.1109/ICTAS.2018.8368757.

[16] G. Rajeswari and P. IthayaRani, "Literature survey on facial expression recognition techniques," in *Proc. 3rd Int. Conf. Commun. Electron. Syst. (ICCES)*, Oct. 2018, pp. 137–142, doi: 10.1109/CESYS.2018.8723953.

[17] I. M. Revina and W. R. S. Emmanuel, "A survey on human face expression recognition techniques," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 33, no. 6, pp. 619–628, Sep. 2018, doi: 10.1016/J.JKSUCI.2018.09.002.

[18] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 325–347, Jul. 2019, doi: 10.1109/TAFFC.2017.2731763.

[19] E. Di Nardo, V. Santopietro, and A. Petrosino, "Emotion recognition at the edge with AI specific low power architectures," *Microprocessors Microsyst.*, vol. 85, Sep. 2021, Art. no. 104299, doi: 10.1016/J.MICPRO.2021.104299.

[20] V. Franzoni, G. Biondi, D. Perri, and O. Gervasi, "Enhancing mouth-based emotion recognition using transfer learning," *Sensors*, vol. 20, no. 18, pp. 1–15, 2020, doi: 10.3390/s20185222.

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," Jan. 2018, *arXiv:1801.04381*. Accessed: Aug. 29, 2021.

[22] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *Int. J. Cognit. Comput. Eng.*, vol. 2, pp. 57–64, Jun. 2021, doi: 10.1016/J.IJCCE.2021.02.002.

[23] M. Al-Shabi, W. P. Cheah, and T. Connie, "Facial expression recognition using a hybrid CNN–SIFT aggregator," in *Proc. Int. Workshop Multi-Disciplinary Trends Artif. Intell.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 10607, Aug. 2016, pp. 139–149, doi: 10.1007/978-3-319-69456-6_12.

[24] P. J. Bota, C. Wang, A. L. N. Fred, and H. P. Da Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140990–141020, 2019, doi: 10.1109/ACCESS.2019.2944001.

[25] T. Kundu and C. Saravanan, "Advancements and recent trends in emotion recognition using facial image analysis and machine learning models," in *Proc. Int. Conf. Electr., Electron., Commun., Comput., Optim. Techn. (ICEECCOT)*, Dec. 2017, pp. 1–6, doi: 10.1109/ICEECCOT.2017.8284512.

[26] S. Bhattacharya and M. Gupta, "A survey on: Facial emotion recognition invariant to pose, illumination and age," in *Proc. 2nd Int. Conf. Adv. Comput. Commun. Paradigms (ICACCP)*, Feb. 2019, pp. 1–6, doi: 10.1109/ICACCP.2019.8883015.

[27] A. S. Vyas, H. B. Prajapati, and V. K. Dabhi, "Survey on face expression recognition using CNN," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2019, pp. 102–106, doi: 10.1109/ICACCS.2019.8728330.

[28] C. Marechal, D. Mikołajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegrzyn-Wolska, "Survey on AI-based multimodal methods for emotion detection," in *High-Performance Modelling and Simulation for Big Data Applications* (Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11400, 2019, pp. 307–324, doi: 10.1007/978-3-030-16272-6_11.

[29] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, early access, Mar. 17, 2020, doi: 10.1109/TAFFC.2020.2981446.

[30] A. Fathima and K. Vaidehi, "Review on facial expression recognition system using machine learning techniques," in *Advances in Decision Sciences, Image Processing, Security and Computer Vision*. Cham, Switzerland: Springer, 2020, pp. 608–618, doi: 10.1007/978-3-030-24318-0_70.

[31] K. Patel, D. Mehta, C. Mistry, R. Gupta, S. Tanwar, N. Kumar, and M. Alazab, "Facial sentiment analysis using AI techniques: State-of-the-art, taxonomies, and challenges," *IEEE Access*, vol. 8, pp. 90495–90519, 2020, doi: 10.1109/ACCESS.2020.2993803.

[32] E. Dufourq, "A survey on factors affecting facial expression recognition based on convolutional neural networks," in *Proc. Conf. South African Inst. Comput. Sci. Inf. Technol.*, Sep. 2020, pp. 168–179, doi: 10.1145/3410886.3410891.

[33] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, present, and future of face recognition: A review," *Electronics*, vol. 9, no. 8, p. 1188, Jul. 2020, doi: 10.3390/ELECTRONICS9081188.

[34] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets," *Vis. Comput.*, vol. 2021, pp. 1–32, Jun. 2021, doi: 10.1007/S00371-021-02166-7.

[35] R. A. Khan, A. Crenn, A. Meyer, and S. Bouakaz, "A novel database of children's spontaneous facial expressions (LIRIS-CSE)," *Image Vis. Comput.*, vols. 83–84, pp. 61–69, Mar. 2019.

[36] A. S. Meuwissen, J. E. Anderson, and P. D. Zelazo, "The creation and validation of the developmental emotional faces stimulus set," *Behav. Res. Methods*, vol. 49, no. 3, p. 960, Jun. 2017, doi: 10.3758/S13428-016-0756-7.

[37] H. L. Egger, D. S. Pine, E. Nelson, E. Leibenluft, M. Ernst, K. E. Towbin, and A. Angold, "The NIMH child emotional faces picture set (NIMH-ChEFS): A new set of children's facial emotion stimuli," *Int. J. Methods Psychiatric Res.*, vol. 20, no. 3, pp. 145–156, Sep. 2011, doi: 10.1002/MPR.343.

[38] K. A. Dalrymple, J. Gomez, and B. Duchaine, "The Dartmouth database of children's faces: Acquisition and validation of a new face stimulus set," *PLoS ONE*, vol. 8, no. 11, Nov. 2013, Art. no. e79131, doi: 10.1371/JOURNAL.PONE.0079131.

[39] V. LoBue and C. Thrasher, "The child affective facial expression (CAFE) set: Validity and reliability from untrained adults," *Frontiers Psychol.*, vol. 5, p. 1532, Jan. 2015, doi: 10.3389/FPSYG.2014.01532.

[40] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "EmoReact: A multimodal approach and dataset for recognizing emotional responses in children," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 137–144, doi: 10.1145/2993148.2993168.

[41] *Radboud Faces Database—Behavioural Science Institute*. Accessed: Aug. 29, 2021. [Online]. Available: https://www.ru.nl/bsi/research/facilities/radboud-faces-database/

[42] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops (CVPRW)*, Jun. 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.

[43] *JAFFE Dataset | Papers With Code*. Accessed: Aug. 29, 2021. [Online]. Available: https://paperswithcode.com/dataset/jaffe

[44] *NVIE Database*. Accessed: Aug. 29, 2021. [Online]. Available: http://nvie.ustc.edu.cn/

[45] *FER2013 Dataset | Papers With Code*. Accessed: Aug. 29, 2021. [Online]. Available: https://paperswithcode.com/dataset/fer2013

[46] *(13) (PDF) the AR Face Database*. Accessed: Aug. 29, 2021. [Online]. Available: https://www.researchgate.net/publication/243651904_The_AR_face_database

[47] *(13) (PDF) Acted Facial Expressions in the Wild Database*. Accessed: Aug. 29, 2021. [Online]. Available: https://www.researchgate.net/publication/229049764_Acted_Facial_Expressions_In_The_Wild_Database

[48] 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.*, vol. 10, no. 1, pp. 18–31, Jan. 2019, doi: 10.1109/taffc.2017.2740923.

[49] T. Yang, Z. Yang, G. Xu, D. Gao, Z. Zhang, H. Wang, S. Liu, L. Han, Z. Zhu, Y. Tian, Y. Huang, L. Zhao, K. Zhong, B. Shi, J. Li, S. Fu, P. Liang, M. J. Banissy, and P. Sun, "Tsinghua facial expression database—A database of facial expressions in Chinese young and older women and men: Development and validation," *PLoS ONE*, vol. 15, no. 4, Apr. 2020, Art. no. e0231304, doi: 10.1371/JOURNAL.PONE.0231304.

[50] K. Wang, Z. Zhu, S. Wang, X. Sun, and L. Li, "A database for emotional interactions of the elderly," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2016, pp. 1–6, doi: 10.1109/ICIS.2016.7550902.

[51] N. C. Ebner, M. Riediger, and U. Lindenberger, "FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behav. Res. Methods*, vol. 42, no. 1, pp. 351–362, Feb. 2010, doi: 10.3758/BRM.42.1.351.

[52] J. Huang, Y. Shang, and H. Chen, "Improved viola-jones face detection algorithm based on HoloLens," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, pp. 1–11, Feb. 2019, doi: 10.1186/S13640-019-0435-6.

[53] L. Cuimei, Q. Zhiliang, J. Nan, and W. Jianhua, "Human face detection algorithm via Haar cascade classifier combined with three additional classifiers," in *Proc. 13th IEEE Int. Conf. Electron. Meas. Instrum. (ICEMI)*, Oct. 2017, pp. 483–487, doi: 10.1109/ICEMI.2017.8265863.

[54] C.-S. Wang, Y.-C. Jeung, L.-B. Luo, J. Wang, and J.-W. Chong, "Real-time face recognition using adaptive skin-color model," in *Proc. Int. Conf. Inf. Sci. Appl. (ICISA)*, Apr. 2011, pp. 1–6, doi: 10.1109/ICISA.2011.5772343.

[55] K. S. Kumar, S. Prasad, V. B. Semwal, and R. C. Tripathi, "Real time face recognition using AdaBoost improved fast PCA algorithm," *Int. J. Artif. Intell. Appl.*, vol. 2, no. 3, pp. 45–58, Jul. 2011, doi: 10.5121/IJAIA.2011.2305.

[56] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 1407, 1998, pp. 484–498, doi: 10.1007/BFB0054760.

[57] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multi-task cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[58] K. C. Kirana, S. Wibawanto, and H. W. Herwanto, "Facial emotion recognition based on Viola–Jones algorithm in the learning environment," in *Proc. Int. Seminar Appl. Technol. Inf. Commun.*, Sep. 2018, pp. 406–410, doi: 10.1109/ISEMANTIC.2018.8549735.

[59] I. Gangopadhyay, A. Chatterjee, and I. Das, "Face detection and expression recognition using Haar cascade classifier and Fisherface algorithm," in *Recent Trends in Signal and Image Processing*. Singapore: Springer, 2019, pp. 1–11, doi: 10.1007/978-981-13-6783-0_1.

[60] G. Yang, L. Zhang, and H. Li, "Face detection based on adaptive skin color model and geometric features," in *Proc. Int. Conf. Electr. Control Eng. (ICECE)*, Sep. 2011, pp. 1566–1568, doi: 10.1109/ICECENG.2011.6058082.

[61] X. Feng, B. Lv, Z. Li, and J. Zhang, "Automatic facial expression recognition with AAM-based feature extraction and SVM classifier," in *Proc. Mex. Int. Conf. Artif. Intell.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 4293, 2006, pp. 726–733, doi: 10.1007/11925231_69.

[62] A. Ghofrani, R. M. Toroghi, and S. Ghanbari, "Realtime face-detection and emotion recognition using MTCNN and miniShuffleNet V2," in *Proc. 5th Conf. Knowl. Based Eng. Innov. (KBEI)*, Feb. 2019, pp. 817–821, doi: 10.1109/KBEI.2019.8734924.

[63] T. H. H. Zavaschi, A. S. Britto, L. E. S. Oliveira, and A. L. Koerich, "Fusion of feature sets and classifiers for facial expression recognition," *Expert Syst. Appl.*, vol. 40, no. 2, pp. 646–655, Feb. 2013, doi: 10.1016/J.ESWA.2012.07.074.

[64] W. Chen, M. J. Er, and S. Wu, "Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 2, pp. 458–466, Apr. 2006, doi: 10.1109/TSMCB.2005.857353.

[65] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Sep. 2015, pp. 1–6, doi: 10.1109/IST.2015.7294547.

[66] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2015, pp. 435–442, doi: 10.1145/2818346.2830595.

[67] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing CNN with preprocessing stage in automatic emotion recognition," *Proc. Comput. Sci.*, vol. 116, pp. 523–529, Jan. 2017, doi: 10.1016/J.PROCS.2017.10.038.

[68] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2015, pp. 467–474, doi: 10.1145/2818346.2830596.

[69] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*, Oct. 2016, pp. 433–436, doi: 10.1145/2993148.2997627.

[70] C. M. Kuo, S. H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2018, pp. 2202–2210, doi: 10.1109/CVPRW.2018.00286.

[71] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "HoloNet: Towards robust emotion recognition in the wild," in *Proc. 18th ACM Int. Conf. Multimodal Interact. (ICMI)*, Oct. 2016, pp. 472–478, doi: 10.1145/2993148.2997639.

[72] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2017, pp. 553–560, doi: 10.1145/3136755.3143009.

[73] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4295–4304, doi: 10.1109/CVPR.2015.7299058.

[74] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3871–3879, doi: 10.1109/ICCV.2015.441.

[75] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/S40537-019-0197-0.

[76] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, "Data augmentation for face recognition," *Neurocomputing*, vol. 230, pp. 184–196, Mar. 2017, doi: 10.1016/J.NEUCOM.2016.12.025.

[77] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 12372, Jun. 2019, pp. 566–583. Accessed: Aug. 30, 2021.

[78] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1857–1865, doi: 10.5555/3305381.

[79] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797, doi: 10.1109/CVPR.2018.00916.

[80] D. Guo and T. Sim, "Digital face makeup by example," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 73–79, doi: 10.1109/CVPR.2009.5206833.

[81] J.-J. Lv, X.-H. Shao, J.-S. Huang, X.-D. Zhou, and X. Zhou, "Data augmentation for face recognition," *Neurocomputing*, vol. 230, pp. 184–196, Mar. 2017, doi: 10.1016/J.NEUCOM.2016.12.025.

[82] Z.-H. Feng, J. Kittler, W. Christmas, P. Huber, and X.-J. Wu, "Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3681–3686, doi: 10.1109/CVPR.2017.392.

[83] I. Masi, A. T. Tran, J. T. Leksut, T. Hassner, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" Mar. 2016, *arXiv:1603.07057*. Accessed: Aug. 30, 2021.

[84] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 146–155, doi: 10.1109/CVPR.2016.23.

[85] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng. (2017). *3DFaceNet: Real-Time Dense Face Reconstruction Via Synthesizing Photo-Realistic Face Images*. Accessed: Aug. 30, 2021. [Online]. Available: https://github.com/Juyong/3DFace

[86] W. Xie, L. Shen, M. Yang, and J. Jiang, "Facial expression synthesis with direction field preservation based mesh deformation and lighting fitting based wrinkle mapping," *Multimedia Tools Appl.*, vol. 77, no. 6, pp. 7565–7593, Apr. 2017, doi: 10.1007/S11042-017-4661-6.

[87] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–14, Nov. 2015, doi: 10.1145/2816795.2818056.

[88] D. Kim, M. Hernandez, J. Choi, and G. Medioni, "Deep 3D face identification," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 133–142, doi: 10.1109/BTAS.2017.8272691.

[89] *Facial Aging and Rejuvenation by Conditional Multi-Adversarial Autoencoder With Ordinal Regression*. Accessed: Aug. 30, 2021. [Online]. Available: https://www.researchgate.net/publication/324387503_Facial_Aging_and_Rejuvenation_by_Conditional_Multi-Adversarial_Autoencoder_with_Ordinal_Regression

[90] S. Bashyal and G. K. Venayagamoorthy, "Recognition of facial expressions using Gabor wavelets and learning vector quantization," *Eng. Appl. Artif. Intell.*, vol. 21, no. 7, pp. 1056–1064, Oct. 2008, doi: 10.1016/J.ENGAPPAI.2007.11.010.

[91] E. Owusu, Y. Zhan, and Q. R. Mao, "A neural-AdaBoost based facial expression recognition system," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3383–3390, Jun. 2014, doi: 10.1016/J.ESWA.2013.11.041.

[92] L. Zhang, D. Tjondronegoro, and V. Chandran, "Random Gabor based templates for facial expression recognition in images with facial occlusion," *Neurocomputing*, vol. 145, pp. 451–464, Dec. 2014, doi: 10.1016/J.NEUCOM.2014.05.008.

[93] A. Hernandez-Matamoros, A. Bonarini, E. Escamilla-Hernandez, M. Nakano-Miyatake, and H. Perez-Meana, "A facial expression recognition with automatic segmentation of face regions," in *Proc. Int. Conf. Intell. Softw. Methodol., Tools, Techn.*, in Communications in Computer and Information Science, vol. 532, 2015, pp. 529–540, doi: 10.1007/978-3-319-22689-7_41.

[94] G. P. Hegde, M. Seetha, and N. Hegde, "Kernel locality preserving symmetrical weighted Fisher discriminant analysis based subspace approach for expression recognition," *Eng. Sci. Technol., Int. J.*, vol. 19, no. 3, pp. 1321–1333, Sep. 2016, doi: 10.1016/J.JESTCH.2016.03.005.

[95] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 1–12, Jan. 2015, doi: 10.1109/TAFFC.2014.2386334.

[96] M. J. Cossetin, J. C. Nievola, and A. L. Koerich, "Facial expression recognition using a pairwise feature selection and classification approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 5149–5155, doi: 10.1109/IJCNN.2016.7727879.

[97] G. Zhao and M. Pietikäinen, "Boosted multi-resolution spatiotemporal descriptors for facial expression recognition," *Pattern Recognit. Lett.*, vol. 30, no. 12, pp. 1117–1127, Sep. 2009, doi: 10.1016/J.PATREC.2009.03.018.

[98] Y. Ji and K. Idrissi, "Automatic facial expression recognition based on spatiotemporal descriptors," *Pattern Recognit. Lett.*, vol. 33, no. 10, pp. 1373–1380, Jul. 2012, doi: 10.1016/J.PATREC.2012.03.006.

[99] F. Z. Salmam, A. Madani, and M. Kissi, "Facial expression recognition using decision trees," in *Proc. 13th Int. Conf. Comput. Graph., Imag. Vis. (CGiV)*, Mar. 2016, pp. 125–130, doi: 10.1109/CGIV.2016.33.

[100] S. Kumar, M. K. Bhuyan, and B. K. Chakraborty, "Extraction of informative regions of a face for facial expression recognition," *IET Comput. Vis.*, vol. 10, no. 6, pp. 567–576, Sep. 2016, doi: 10.1049/IET-CVI.2015.0273.

[101] B. Ryu, A. R. Rivera, J. Kim, and O. Chae, "Local directional ternary pattern for facial expression recognition," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 6006–6018, Dec. 2017, doi: 10.1109/TIP.2017.2726010.

[102] M. Guo, X. Hou, Y. Ma, and X. Wu, "Facial expression recognition using ELBP based on covariance matrix transform in KLT," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2995–3010, Jan. 2017, doi: 10.1007/s11042-016-3282-9.

[103] S. Nigam, R. Singh, and A. K. Misra, "Efficient facial expression recognition using histogram of oriented gradients in wavelet domain," *Multimedia Tools Appl.*, vol. 77, no. 21, pp. 28725–28747, Nov. 2018, doi: 10.1007/S11042-018-6040-3.

[104] Y. Gao, M. K. H. Leung, S. C. Hui, and M. W. Tananda, "Facial expression recognition from line-based caricatures," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 33, no. 3, pp. 407–412, May 2003, doi: 10.1109/TSMCA.2003.817057.

[105] S. Noh, H. Park, Y. Jin, and J.-I. Park, "Feature-adaptive motion energy analysis for facial expression recognition," in *Proc. Int. Symp. Vis. Comput.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 4841, Nov. 2007, pp. 452–463, doi: 10.1007/978-3-540-76858-6_45.

[106] M. H. Siddiqi, R. Ali, A. Sattar, A. M. Khan, and S. Lee, "Depth camera-based facial expression recognition system using multilayer scheme," *IETE Tech. Rev.*, vol. 31, no. 4, pp. 277–286, Jul. 2014, doi: 10.1080/02564602.2014.944588.

[107] M. H. Siddiqi, R. Ali, A. M. Khan, Y. T. Park, and S. Lee, "Human facial expression recognition using stepwise linear discriminant analysis and hidden conditional random fields," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1386–1398, Apr. 2015, doi: 10.1109/TIP.2015.2405346.

[108] Y. Lee, K. Lee, and S. Pan, "Local and global feature extraction for face recognition," in *Proc. Int. Conf. Audio-Video-Based Biometric Person Authentication*, in Lecture Notes in Computer Science, vol. 3546, 2005, pp. 219–228, doi: 10.1007/11527923_23.

[109] J. Wang and W. Zhang, "A survey of corner detection methods," in *Proc. ICEEA*, vol. 139, 2018, pp. 214–219, doi: 10.2991/iceea-18.2018.47.

[110] R. Acevedo-Avila, M. Gonzalez-Mendoza, and A. Garcia-Garcia, "A linked list-based algorithm for blob detection on embedded vision-based sensors," *Sensors*, vol. 16, no. 6, p. 782, May 2016, doi: 10.3390/S16060782.

[111] A. Uçar, Y. Demir, and C. Güzeliş, "A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering," *Neural Comput. Appl.*, vol. 27, no. 1, pp. 131–142, Mar. 2014, doi: 10.1007/S00521-014-1569-1.

[112] H. Mahersia and K. Hamrouni, "Using multiple steerable filters and Bayesian regularization for facial expression recognition," *Eng. Appl. Artif. Intell.*, vol. 38, pp. 190–202, Feb. 2015, doi: 10.1016/J.ENGAPPAI.2014.11.002.

[113] C. Wang, Y. Wang, and Z. Zhang, "Patch-based bag of features for face recognition in videos," in *Proc. Chin. Conf. Biometric Recognit.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 7701, 2012, pp. 1–8, doi: 10.1007/978-3-642-35136-5_1.

[114] R. Khedgaonkar, M. M. Raghuwanshi, and K. R. Singh, "Patch-based face recognition under plastic surgery," in *Proc. 1st Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, Dec. 2018, pp. 364–368, doi: 10.1109/ICSCCC.2018.8703270.
[115] Z. Xu, Y. Liu, M. Ye, L. Huang, H. Yu, and X. Chen, "Patch based collaborative representation with Gabor feature and measurement matrix for face recognition," *Math. Problems Eng.*, vol. 2018, pp. 1–13, Jan. 2018, doi: 10.1155/2018/3025264.
[116] A. Chandran and V. Ansari, "Facial expression recognition using patch based Gabor features," *Int. J. Appl. Inf. Syst.*, vol. 10, no. 7, pp. 23–28, Mar. 2016, doi: 10.5120/IJAIS2016451526.
[117] T.-X. Jiang, T.-Z. Huang, X.-L. Zhao, and T.-H. Ma, "Patch-based principal component analysis for face recognition," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–9, Jul. 2017, doi: 10.1155/2017/5317850.
[118] M. Xin and Y. Wang, "Research on image classification model based on deep convolution neural network," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, pp. 1–11, Feb. 2019, doi: 10.1186/S13640-019-0417-8.
[119] "Learning and relearning in Boltzmann machines," *Graph. Models*, Jan. 2020, doi: 10.7551/MITPRESS/3349.003.0005.
[120] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/NECO.2006.18.7.1527.
[121] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/SCIENCE.1127647.
[122] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1805–1812, doi: 10.1109/CVPR.2014.233.
[123] Y. Lv, Z. Feng, and C. Xu, "Facial expression recognition via deep learning," in *Proc. Int. Conf. Smart Comput. SMARTCOMP*, Nov. 2014, pp. 303–308, doi: 10.1109/SMARTCOMP.2014.7043872.
[124] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial expression recognition: A survey," *Symmetry*, vol. 11, no. 10, p. 1189, Sep. 2019, doi: 10.3390/SYM11101189.
[125] R. Walecki, O. Rudovic, V. Pavlovic, B. Schuller, and M. Pantic, "Deep structured learning for facial action unit intensity estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5709–5718, doi: 10.1109/CVPR.2017.605.
[126] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980, doi: 10.1007/BF00344251.
[127] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?—Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vols. 7–12, Oct. 2015, pp. 685–694, doi: 10.1109/CVPR.2015.7298668.
[128] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 111–118.
[129] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *Proc. Int. Conf. Artif. Neural Netw.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 6354, 2010, pp. 92–101, doi: 10.1007/978-3-642-15825-4_10.
[130] H. Wu and X. Gu, "Max-pooling dropout for regularization of convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 9489, 2015, pp. 46–54, doi: 10.1007/978-3-319-26532-2_6.
[131] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 8691, 2014, pp. 346–361, doi: 10.1007/978-3-319-10578-9_23.
[132] W. Ouyang, X. Zeng, X. Wang, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, H. Li, K. Wang, J. Yan, C.-C. Loy, and X. Tang, "DeepID-Net: Object detection with deformable part based convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1320–1334, Jul. 2017, doi: 10.1109/TPAMI.2016.2587642.
[133] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[134] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5455–5516, Apr. 2020, doi: 10.1007/S10462-020-09825-6.
[135] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Proc. Asian Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 9006, 2014, pp. 143–157, doi: 10.1007/978-3-319-16817-3_10.
[136] *AlexNet—ImageNet Classification With Convolutional Neural Networks*. Accessed: Sep. 3, 2021. [Online]. Available: https://neurohive.io/en/popular-networks/alexnet-imagenet-classification-with-deep-convolutional-neural-networks/
[137] A. S. Tarawneh, D. Chetverikov, and A. B. Hassanat, "Pilot comparative study of different deep features for palmprint identification in low-quality images," in *Proc. 9th Hungarian Conf. Comput. Graph. Geometry*, Apr. 2018, pp. 1–6. Accessed: Sep. 3, 2021.
[138] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vols. 7–12, Sep. 2014, pp. 1–9. Accessed: Sep. 3, 2021.
[139] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 770–778. Accessed: Sep. 3, 2021.
[140] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 4278–4284. Accessed: Sep. 3, 2021.
[141] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," Feb. 2016, *arXiv:1602.07360*. Accessed: Sep. 3, 2021.
[142] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
[143] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807, doi: 10.1109/CVPR.2017.195.
[144] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995, doi: 10.1109/CVPR.2017.634.
[145] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856, doi: 10.1109/CVPR.2018.00716.
[146] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[147] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[148] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," Feb. 2021, *arXiv:2102.06171*. Accessed: Sep. 3, 2021.
[149] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: Generic features for video analysis," *CoRR*, vol. abs/1412.0767, no. 7, p. 8, 2014.
[150] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–27, Jan. 2009, doi: 10.1561/2200000006.
[151] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biol. Cybern.*, vol. 59, nos. 4–5, pp. 291–294, Sep. 1988, doi: 10.1007/BF00332918.
[152] N. Japkowicz, S. J. Hanson, and M. A. Gluck, "Nonlinear autoassociation is not equivalent to PCA," *Neural Comput.*, vol. 12, no. 3, pp. 531–545, 2000, doi: 10.1162/089976600300015691.
[153] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. 12, pp. 1–38, 2010.
[154] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Dec. 2011, pp. 8595–8598. Accessed: Sep. 3, 2021.

[155] S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, and X. Glorot, "Higher order contractive auto-encoder," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 6912, 2011, pp. 645–660, doi: 10.1007/978-3-642-23783-6_41.

[156] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proc. Int. Conf. Artif. Neural Netw.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 6791, 2011, pp. 52–59, doi: 10.1007/978-3-642-21735-7_7.

[157] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Dec. 2013, pp. 1–14. Accessed: Sep. 3, 2021.

[158] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, Oct. 1990, doi: 10.1109/5.58337.

[159] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[160] A. Mostafa, M. I. Khalil, and H. Abbas, "Emotion recognition by facial features using recurrent neural networks," in *Proc. 13th Int. Conf. Comput. Eng. Syst. (ICCES)*, Dec. 2018, pp. 417–422, doi: 10.1109/ICCES.2018.8639182.

[161] D. K. Jain, Z. Zhang, and K. Huang, "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," *Pattern Recognit. Lett.*, vol. 139, pp. 157–165, Nov. 2020, doi: 10.1016/J.PATREC.2017.06.025.

[162] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2278–2288, doi: 10.1109/CVPRW.2017.282.

[163] Y. Luo, C.-M. Wu, and Y. Zhang, "Facial expression recognition based on fusion feature of PCA and LBP with SVM," *Optik*, vol. 124, no. 17, pp. 2767–2770, Sep. 2013, doi: 10.1016/J.IJLEO.2012.08.040.

[164] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2559–2573, Jul. 2013, doi: 10.1109/TIP.2013.2253477.

[165] N. Chanthaphan, K. Uchimura, T. Satonaka, and T. Makioka, "Multiple classifier learning of new facial extraction approach for facial expressions recognition using depth sensor," in *Proc. 13th Int. Joint Conf. e-Bus. Telecommun. (ICETE)*, vol. 5, 2016, pp. 19–27, doi: 10.5220/0005948000190027.

[166] Z. Pan, M. Polceanu, and C. Lisetti, "On constrained local model feature normalization for facial expression recognition," in *Proc. Int. Conf. Intell. Virtual Agents*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 10011, 2016, pp. 369–372, doi: 10.1007/978-3-319-47665-0_35.

[167] I. Dagher, E. Dahdah, and M. Al Shakik, "Facial expression recognition using three-stage support vector machines," *Vis. Comput. Ind., Biomed., Art*, vol. 2, no. 1, pp. 1–9, Dec. 2019, doi: 10.1186/S42492-019-0034-5.

[168] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition based on facial components detection and HOG features," in *Proc. Sci. Cooperations Int. Workshops Elect. Comput. Eng. Subfields*, 2014, pp. 1–6.

[169] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10, doi: 10.1109/WACV.2016.7477450.

[170] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, Jan. 2018, doi: 10.1016/J.NEUCOM.2017.08.043.

[171] F. Y. Shih, C.-F. Chuang, and P. S. P. Wang, "Performance comparisons of facial expression recognition in JAFFE database," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 22, no. 3, pp. 445–459, Nov. 2011, doi: 10.1142/S0218001408006284.

[172] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2015, pp. 1–2, doi: 10.1109/IJCNN.2015.7280539.

[173] V. V. Salunke and C. G. Patil, "A new approach for automatic face emotion recognition and classification based on deep networks," in *Proc. Int. Conf. Comput., Commun., Control Automat. (ICCUBEA)*, Aug. 2017, pp. 1–5, doi: 10.1109/ICCUBEA.2017.8463785.

[174] P. Shen, S. Wang, and Z. Liu, "Facial expression recognition from infrared thermal videos," in *Intelligent Autonomous Systems 12*. Berlin, Germany: Springer, 2013, pp. 323–333, doi: 10.1007/978-3-642-33932-5_31.

[175] S. Wang, S. He, Y. Wu, M. He, and Q. Ji, "Fusion of visible and thermal images for facial expression recognition," *Frontiers Comput. Sci.*, vol. 8, no. 2, pp. 232–242, Apr. 2014, doi: 10.1007/S11704-014-2345-1.

[176] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Fusion of classifier predictions for audio-visual emotion recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 61–66, doi: 10.1109/ICPR.2016.7899608.

[177] R. A. Khan, A. Meyer, and S. Bouakaz, "Automatic affect analysis: From children to adults," in *Proc. Int. Symp. Vis. Comput.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 9475, Dec. 2015, pp. 304–313, doi: 10.1007/978-3-319-27863-6_28.

[178] A. Lopez-Rincon, "Emotion recognition using facial expressions in children using the NAO robot," in *Proc. Int. Conf. Electron., Commun. Comput. (CONIELECOMP)*, Feb. 2019, pp. 146–153, doi: 10.1109/CONIELECOMP.2019.8673111.

[179] S. Berretti, S. M. Thampi, and S. Dasgupta, Eds., *Intelligent Systems Technologies and Applications*, vol. 385, 2016, doi: 10.1007/978-3-319-23258-4.

[180] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *Proc. IEEE Workshop Adv. Robot. its Social Impacts (ARSO)*, Mar. 2017, pp. 1–7, doi: 10.1109/ARSO.2017.8025197.

[181] B. Nagarajan and V. R. M. Oruganti, "Cross-domain transfer learning for complex emotion recognition," in *Proc. IEEE Region Symp. (TENSYMP)*, vol. 7, Jun. 2019, pp. 649–653, doi: 10.1109/TENSYMP46218.2019.8971023.

[182] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion," *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–49, Jun. 2018, doi: 10.1145/3158369.

[183] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009, doi: 10.1109/TPAMI.2008.52.

[184] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11217, Sep. 2018, pp. 227–243, doi: 10.1007/978-3-030-01261-8_14.

[185] T.-H. Vo, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "Pyramid with super resolution for in-the-wild facial expression recognition," *IEEE Access*, vol. 8, pp. 131988–132001, 2020, doi: 10.1109/ACCESS.2020.3010018.

[186] Q. Yang, "An introduction to transfer learning," *Adv. Data Mining Appl.*, p. 1, Sep. 2008, doi: 10.1007/978-3-540-88192-6_1.

[187] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*. Accessed: Sep. 5, 2021.

[188] H. Kaya, F. Gürpinar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image Vis. Comput.*, vol. 65, pp. 66–75, Sep. 2017, doi: 10.1016/J.IMAVIS.2017.01.012.

[189] S. F. Aly and A. L. Abbott, "Facial emotion recognition with varying poses and/or partial occlusion using multi-stage progressive transfer learning," in *Proc. Scand. Conf. Image Anal.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11482, 2019, pp. 101–112, doi: 10.1007/978-3-030-20205-7_9.

[190] T. Q. Ngo and S. Yoon, "Facial expression recognition on static images," in *Proc. Int. Conf. Future Data Secur. Eng.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11814, Nov. 2019, pp. 640–647, doi: 10.1007/978-3-030-35653-8_42.

[191] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in *Proc. 11th Int. Conf. Natural Comput. (ICNC)*, Aug. 2015, pp. 702–708, doi: 10.1109/ICNC.2015.7378076.

[192] P. Shen, S. Wang, and Z. Liu, "Facial expression recognition from infrared thermal videos," in *Intelligent Autonomous Systems 12*. Berlin, Germany: Springer, 2013, pp. 323–333, doi: 10.1007/978-3-030-45529-3_1.
[193] P. Kumar and B. Raman, "Domain adaptation based technique for image emotion recognition using pre-trained facial expression recognition models," Nov. 2020, *arXiv:2011.08388*. Accessed: Sep. 5, 2021.
[194] G. Zen, E. Sangineto, E. Ricci, and N. Sebe, "Unsupervised domain adaptation for personalized facial emotion recognition," in *Proc. 16th Int. Conf. Multimodal Interact. (ICMI)*, Nov. 2014, pp. 128–135, doi: 10.1145/2663204.2663247.
[195] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, "We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 357–366, doi: 10.1145/2647868.2654916.
[196] A. Kumar, S. Mehta, and D. Vijaykeerthy, "An introduction to adversarial machine learning," in *Proc. Int. Conf. Big Data Anal.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 10721, 2017, pp. 293–299, doi: 10.1007/978-3-319-72413-3_20.
[197] V. Narula, Zhangyang, Wang, and T. Chaspari, "An adversarial learning framework for preserving users' anonymity in face-based emotion recognition," Jan. 2020, *arXiv:2001.06103*. Accessed: Sep. 5, 2021.
[198] O. A. Soysal and M. S. Guzel, "An introduction to zero-shot learning: An essential review," in *Proc. Int. Congr. Hum.-Comput. Interact., Optim. Robot. Appl. (HORA)*, Jun. 2020, pp. 1–4, doi: 10.1109/HORA49412.2020.9152859.
[199] J. Wu, Y. Zhang, X. Zhao, and W. Gao, "A generalized zero-shot framework for emotion recognition from body gestures," Oct. 2020, *arXiv:2010.06362*. Accessed: Sep. 5, 2021.
[200] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "An introduction to reinforcement learning," in *The Biology and Technology of Intelligent Autonomous Agents*, 1995, pp. 90–127.
[201] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Artif. Intell. Stat.*, Apr. 2016, pp. 1273–1282. Accessed: Sep. 5, 2021.
[202] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in *Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extraction*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 12279, 2020, pp. 1–16, doi: 10.1007/978-3-030-57321-8_1.
[203] A. Generosi, S. Ceccacci, and M. Mengoni, "A deep learning-based system to track and analyze customer behavior in retail store," in *Proc. IEEE 8th Int. Conf. Consum. Electron.-Berlin (ICCE-Berlin)*, Sep. 2018, pp. 1–6, doi: 10.1109/ICCE-BERLIN.2018.8576169.
[204] M. I. U. Haque and D. Valles, "A facial expression recognition approach using DCNN for autistic children to identify emotions," in *Proc. IEEE 9th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Nov. 2018, pp. 546–551, doi: 10.1109/IEMCON.2018.8614802.
[205] X. Liu and K. Lee, "Optimized facial emotion recognition technique for assessing user experience," in *Proc. IEEE Games, Entertainment, Media Conf. (GEM)*, Aug. 2018, pp. 407–413, doi: 10.1109/GEM.2018.8516518.

CHIRAG DALVI received the B.Tech. degree in information technology from Symbiosis International University. He is currently pursuing the M.S. degree in information systems with the Stevens Institute of Technology, New Jersey, USA. He is also a Research Intern with the Symbiosis Center for Applied Artificial Intelligence (SCAAI). His research interests include artificial intelligence, machine learning domain, computers vision, and multimodal deep learning.

MANISH RATHOD received the B.Tech. degree in information technology from Symbiosis International University. He is currently working professionally at Amazon. He is a tech and business enthusiast and loves research. He has worked as an Intern with the Symbiosis Centre of Applied Artificial Intelligence (SCAAI). He was also the first runner up in Smart India Hackathon 2020. His research interests include business analytics, artificial intelligence, international politics, economics, and research.

SHRUTI PATIL received the M.Tech. degree in computer science and the Ph.D. degree in data privacy from Pune University. She has been an industry professional in the past, currently associated with the Symbiosis Institute of Technology as a Professor and as a Research Associate with SCAAI, Pune Maharashtra. She has three years of industry experience and ten years of academic experience. She has expertise in applying innovative technology solutions to real world problems. She is currently working in the application domains of healthcare, sentiment analysis, emotion detection, and machine simulation. She is also guiding several U.G., P.G., and Ph.D. students as a domain expert. She has published more than 30 research papers in reputed international conferences and scopus/web of science indexed journals and books. Her research interests include applied artificial intelligence, natural language processing, acoustic AI, adversarial machine learning, data privacy, digital twin applications, GANS, and multimodal data analysis.

SHILPA GITE received the Ph.D. degree in deep learning for assistive driving in semi autonomous vehicles from Symbiosis International (Deemed University), Pune, India, in 2019. Currently, she is working as an Associate Professor with the Computer Science Department, Symbiosis Institute of Technology, Pune. She is also working as an Associate Faculty at the Symbiosis Centre of Applied AI (SCAAI). She has around 13 years of teaching experience. She is currently guiding Ph.D. students in biomedical imaging, self-driving cars, and natural language processing areas. She has published more than 30 research papers in scopus-indexed and SCI-indexed international journals and 25 scopus indexed international conferences. Her research interests include deep learning, machine learning medical imaging, and computer vision. She was a recipient of the Best Paper Award at 11th IEMERA Conference held virtually at Imperial College, London, in October 2020.

KETAN KOTECHA has expertise and experience of cutting-edge research and projects in AI and deep learning for last 25 years (more than). He has published widely in several excellent peer-reviewed journals on various topics ranging from education policies, teaching-learning practices, and AI. He is also a team member for the nation-wide initiative on AI and deep learning skilling and research named Leadingindia.ai initiative sponsored by the Royal Academy of Engineering, U.K., under Newton Bhabha Fund. He currently heads the Symbiosis Centre for Applied Artificial Intelligence (SCAAI). He is considered a foremost expert in AI and aligned technologies. Additionally, with his vast and varied experience in administrative roles, he has pioneered education technology. Previously, he has worked as an Administrator at Parul University and Nirma University and has several achievements in these roles to his credit.