Asian Journal of Electrical Sciences
ISSN: 2249-6297 (P); 2583-7923 (E)
Vol.13 No.1, 2024, pp.35-45

www.crijournals.org
DOI: https://doi.org/10.70112/ajes-2024.13.1.4241

Exploring the Impact of Convolutional Neural Networks on Facial Emotion Detection and Recognition

Rexcharles Enyinna Donatus[1*], Ifeyinwa Happiness Donatus[2] and Ubadike Osichinaka Chiedu[3]
[1]Africa Centre of Excellence on Technology Enhanced Learning (ACETEL), National Open University of Nigeria, Nigeria
[1&3]Department of Aerospace Engineering, Air Force Institute of Technology, Nigeria
[2]Department of Computer Science, Kaduna State University, Nigeria
*Corresponding author: charlly4eyims@yahoo.com
(Received 5 March 2024; Revised 13 April 2024; Accepted 20 April 2024; Available online 24 April 2024)

Abstract **- Emotional analytics is a fascinating blend of psychology and technology, with one of the primary methods for recognizing emotions involving facial expression analysis. Facial emotion detection has advanced significantly, utilizing deep learning algorithms to identify common emotions. In recent years, substantial progress has been made in automatic facial emotion recognition (FER). This technology has been applied across various industries to enhance interactions between humans and machines, particularly in human-centered computing and the emerging field of emotional artificial intelligence (EAI). Researchers focus on improving systems' capabilities to recognize and interpret human facial expressions and behaviors in diverse contexts. The impact of convolutional neural networks (CNNs) on this field has been profound, as these networks have undergone significant development, leading to diverse architectures designed to address increasingly complex challenges. This article explores the latest advancements in automated emotion recognition using computational intelligence, emphasizing how contemporary deep learning models contribute to the field. It provides a review of recent developments in CNN architectures for FER over the past decade, demonstrating how deep learning-based methods and specialized databases collaborate to achieve highly accurate outcomes.**

Keywords: **Facial Emotion Recognition (FER), Deep Learning Algorithms, Convolutional Neural Networks (CNNS), Emotional Artificial Intelligence (EAI), Human-Centered Computing**

I. INTRODUCTION

Emotions profoundly impact the human experience, shaping our interactions, decisions, and overall well-being. Emotion plays a crucial role in our daily lives, reflecting our intentions and mental and physical conditions [1], [2]. Human emotional states can be deduced from verbal and non-verbal cues gathered through various sensors, including physiological signals, facial expressions, and voice tone [3]. Understanding and recognizing emotions is essential not only for advancing human-machine interaction but also for its role in mental health support applications and enhancing customer satisfaction across different sectors [4]. There is a growing expectation for interactive machines to recognize, interpret, and express emotions in ways comparable to human behavior [5]. This shift has increased the demand for machine-human interaction systems where users anticipate machines to demonstrate a full spectrum of emotions. Consequently, the development of emotion recognition systems that can accurately interpret and respond to human emotions is essential for fostering more seamless and efficient interactions between humans and machines [6].

Facial expressions often act as key indicators of a person's emotional state, which is why numerous researchers are particularly focused on this modality [7]. Studies have demonstrated that emotional information is conveyed through multiple channels: 55% visual, 38% vocal, and only 7% verbal [8], [9]. Facial expressions are powerful, natural, and universal indicators of emotions and thoughts, transcending gender, ethnicity, and nationality [10]. Facial emotion recognition is gaining importance in the modern world, with applications in security, clinical psychology, evaluating blood pressure and stress levels, neurology, law enforcement, multimodal human-computer interfaces, and human-computer interaction. These applications aim to enhance communication and understanding between humans and machines by detecting emotions such as happiness, sadness, calmness, and neutrality [11], [12], [13].

Automatic facial emotion recognition is a vital research area that bridges psychological understanding of human emotions and artificial intelligence (AI). The techniques and algorithms developed in this domain can reveal the body's internal mechanisms, enabling early disease detection and offering insights into mental states without requiring direct inquiry [11], [14]. Early face recognition research primarily relied on images captured in controlled settings. However, recognizing the limitations of image-based methods, researchers hypothesized that incorporating temporal or spatio-temporal data from videos could improve recognition accuracy. This led to the collection of several video-based facial datasets between 2000 and 2010 [7]. By 2010, deep neural networks had achieved notable success in object recognition, attracting considerable attention and prompting researchers to explore their applications across various fields [15].

Deep learning can be utilized for emotion detection and facial expression analysis. However, its effectiveness is influenced by the size of the dataset, with larger datasets generally yielding better results [16]. Currently, the available datasets

for facial expression analysis remain too limited for optimal deep learning implementation. To address this, some researchers use data augmentation methods in the pre-processing phase, including techniques like scaling, cropping, mirroring, and translation, which enhance data variability and effectively increase dataset size. These pre-processing strategies have been shown to significantly boost deep learning performance [17], [18], [19].

A. Steps Involve in Facial Emotion Classification

Facial emotion recognition (FER) comprises three key stages: preprocessing, feature extraction, and emotion identification. For preprocessing, several methods are used, including face detection, background removal, keyframe extraction, and facial landmark detection [20].

Face detection utilizes various algorithms such as the Multi-Task Cascaded Convolutional Network (MTCNN), the Viola-Jones Detector [21], the light and fast face detector, tiny face detection, FaceNet, the Caffe-based face detector, the Haar feature-based cascade detector, and the Single Stage Headless (SSH) detector. Libraries such as OpenFace, OpenCV, and Dlib are used for landmark extraction and face detection, enabling the isolation of key facial regions while minimizing background noise [2], [22], [23].

Since 2014, deep neural networks (DNNs) have been successfully utilized in face recognition systems, driven by improvements in processing power and the increased availability of large, multilabel datasets [24]-[26]. The DeepFace technique [24], which employs DNNs, achieved a face recognition accuracy of approximately 97.35% on the Labeled Faces in the Wild (LFW) dataset, which contains thousands of face images captured in uncontrolled environments - closely matching human performance levels (97.53%). Since then, accuracy on the LFW dataset has improved to as high as 99.80% [27]. Despite being nonlinear, appearance-based methods, DNN-based approaches are widely used in recent facial recognition studies and have shown superior performance compared to other methods.

While DNN techniques provide high accuracy for both face identification and verification tasks using images, demonstrating resilience in challenging conditions, such as varying lighting, occlusions, and facial expressions, remains an area of active research [28]. These methods show strong performance when applied to large datasets of high-quality images, even those captured in uncontrolled environments with diverse lighting, poses, and expressions. However, significant drops in recognition accuracy have been observed when images are affected by severe illumination changes, noise, or low resolution [28]. In such adverse conditions, video-based methods may offer valuable insights into facial dynamics, potentially improving recognition performance.

Recent advancements in face detection have yielded successful outcomes using deep learning techniques [29], [30], [31]. One notable approach is Faster R-CNN, which employs region proposals and was originally introduced for object detection [32], [33]. Additionally, some deep learning-based face detection techniques employ a sliding-window approach, scanning the image across various scales and positions to detect facial regions effectively [34], [35], [36]. The Single Shot Detector (SSD), initially designed for object detection, has also proven effective in face detection applications [37], [38].

B. Feature Extraction

Facial feature extraction techniques can be classified into geometric-based approaches, which represent facial points to create feature vectors from a geometrical perspective, and appearance-based techniques, such as Gabor Wavelets, which extract features from focused or comprehensive facial images. Accurate feature extraction from one face to another is a challenging task, critical for effective classification and analysis. The choice of features in FER significantly impacts performance, making feature extraction a crucial and carefully considered step.

Feature extraction efforts often incorporate the Facial Action Coding System (FACS), which monitors how facial muscles contract and relax at different intensity levels. FACS has been refined over time to improve its accuracy and robustness in recognizing subtle facial movements [39]. Other traditional methods for facial feature extraction include Local Binary Patterns (LBP) [40], Histogram of Oriented Gradients (HOG), and Local Directional Patterns (LDP) [41]. These techniques focus on capturing texture and gradient-based information from images to enhance feature extraction accuracy.

When designing deep convolutional neural networks (CNNs) for feature extraction, the choice of an effective loss function and the selection of a suitable network architecture are two critical factors to consider. CNN architectures are generally categorized as either backbone networks or multi-network systems. Following their impressive results in ImageNet competitions, models such as SENet, VGGNet, AlexNet, ResNet, and GoogleNet, which are considered standard CNN architectures, have been widely studied by researchers [7].

These networks, along with their variants, have been extensively applied in facial recognition tasks. Moreover, multi-structure networks have been developed to facilitate multi-task learning, enabling simultaneous tasks such as face recognition and other related objectives [42], [43].

Selecting an appropriate loss function is essential for training deep CNNs in face recognition. Research indicates that relying solely on softmax loss is inadequate for effective feature separation, primarily because intra-class variation often exceeds inter-class variation. Consequently, alternative loss functions have been introduced to improve feature discrimination [7]. The structure of CNNs is illustrated in Figure 1.

C. Classification

Emotion classification follows feature extraction in FER, utilizing various algorithms such as conventional learning methods and CNNs, with the latter being highly efficient and accurate. The variability of human emotions makes context-based classification challenging.

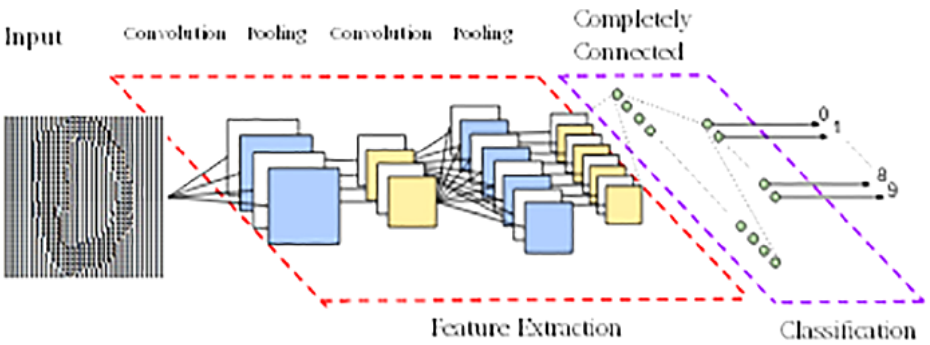

Fig. 1 Standard CNN architecture along with its feature extraction convolutional pool [39]

Deep learning (DL) has proven to be a highly effective method, primarily due to its ability to automatically extract features and classify data using architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This success has motivated researchers to leverage these techniques for human emotion recognition. Significant efforts have been made to develop deep neural network architectures, leading to impressive results in this domain. Among these, CNNs are the most widely employed classification algorithms, as they can be directly applied to input images without requiring separate facial detection or feature extraction processes. CNNs include convolutional layers that reduce the input data by acquiring relevant features and producing feature maps with the help of various feature detectors [44]. The convolutional layer condenses the input while identifying key features, resulting in feature maps utilized by different feature detectors.

This paper reviews recent advances in emotion sensing through facial expression recognition using various DL architectures. We present findings from 2014 to 2024, interpreting the challenges and contributions in this field. The paper is organized as follows: Section II presents the latest state-of-the-art techniques in facial expression recognition using deep learning; Section III introduces some publicly available databases that facilitate facial emotion recognition; and Sections IV and V provide a discussion and comparison of the methods, concluding with a summary and suggestions for future work.

II. FACIAL EMOTION RECOGNITION (FER) USING CONVOLUTIONAL NEURAL NETWORKS

FER has made significant strides with the emergence of deep learning, moving away from the traditional handcrafted feature methods it once depended on [45]. Over the past decade, there has been a growing preference among researchers for DL due to its superior ability to automatically recognize hidden patterns. This section reviews recent studies in FER that utilize CNNs to enhance detection performance. CNNs are well-suited for FER and other computer vision applications.

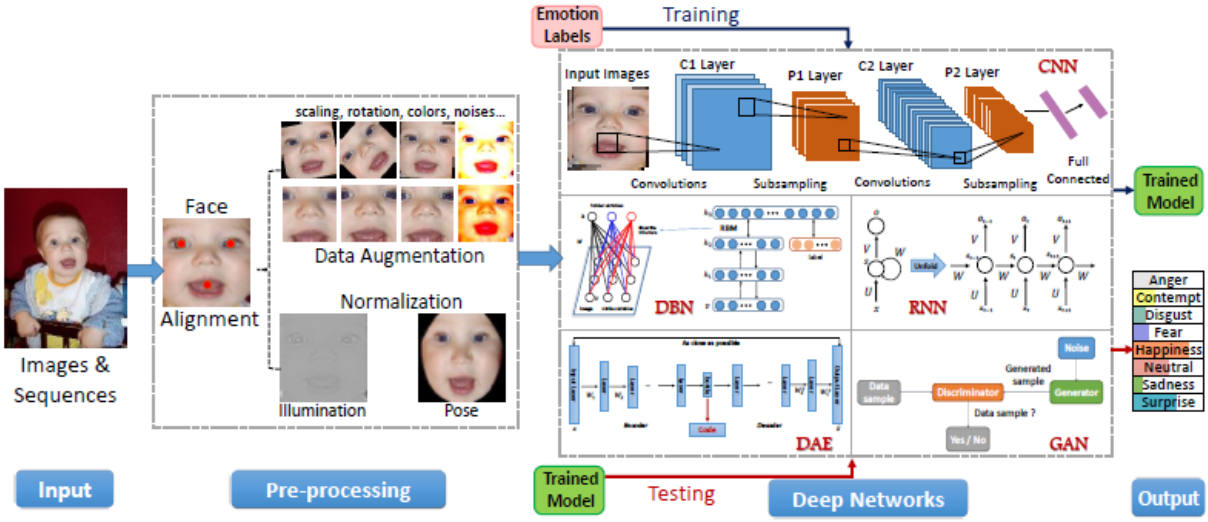

Fig. 2 The overall pipeline for deep facial expression recognition systems [46]

Early 21st-century research in the FER literature indicated that CNNs outperform multilayer perceptrons (MLPs) and provide strong results when managing scale variations and changes in face location, excelling in handling previously unseen variations in facial pose [46], [47]. CNNs, a prevalent type of DL architecture, are recognized for capturing higher-level abstractions by employing a hierarchical structure with multiple nonlinear representations and transformations [48]. Figure 2 illustrates the overall pipeline of deep facial expression recognition systems.

These methods are tested and trained on various static or sequential databases to improve accuracy and robustness in emotion recognition.

A. Popular CNN Architectures used for Facial Emotion Recognition

1. AlexNet: The DeepFace study [24] demonstrated that deep learning networks, particularly those successful in object recognition, could also excel in face recognition. This study employed an AlexNet architecture combined with the softmax loss function, trained on a large dataset of facial images. The result of 97.33% represents an impressive test performance on the Labeled Faces in the Wild (LFW) dataset, which consists of labeled facial images. This performance approaches human-level accuracy, marking a substantial leap in face recognition technology. In 2014, the DeepID2 model, built on the AlexNet architecture and employing contrastive loss, achieved an impressive 99.15% accuracy on the CelebFaces+ dataset [49].

2. VGG-Net: The study by [50] presents DeepID3, an advanced deep neural network architecture for face recognition that utilizes stacked convolutional and inception layers, inspired by VGGNet and GoogLeNet, to enhance feature extraction. Preprocessing involves adding joint face identification-verification supervisory signals and employing data augmentation techniques, such as varying positions, scales, and color channels of face regions, to improve robustness. The architecture's increased depth addresses performance issues by enabling more complex feature extraction, resulting in an impressive 99.53% accuracy for face verification and 96.0% for face identification using the LFW dataset, surpassing previous models like DeepID2+. However, the study calls for further research to assess the model's effectiveness in diverse and uncontrolled environments.

3. MTCCNN: In Zhang *et al.,* [51], a novel framework was introduced that employs multi-task learning to combine face detection and alignment, utilizing cascaded CNNs. The framework consists of three stages: the Proposal Network (P-Net), which generates candidate windows; the Refine Network (R-Net), responsible for filtering out non-face candidates; and the Output Network (O-Net), which ensures precise facial landmark localization. The framework incorporates an online hard sample mining strategy that dynamically selects informative training samples, leading to significant improvements over state-of-the-art techniques on benchmarks like the WIDER FACE dataset, FDDB, and AFLW. While the results demonstrate enhanced accuracy and efficiency in challenging conditions, critiques highlight potential model complexity and reliance on large annotated datasets, suggesting a need for further research into unsupervised learning techniques and the integration of contextual information to improve performance in diverse environments.

4. GoogleNet: FaceNet [26], introduced by Schroff *et al.,* in 2015, revolutionized face recognition by leveraging GoogleNet to map face images to a compact Euclidean space through deep convolutional neural networks (CNNs) and a triplet loss function. The model incorporates preprocessing techniques, including tight cropping of face thumbnails, and applies data augmentation strategies such as flipping and random cropping to improve robustness against variations in pose and lighting conditions. Its architecture features interleaved layers of convolutions and pooling, optimizing performance by reducing parameters while maintaining accuracy. FaceNet achieved an impressive accuracy of 99.63% on the Labeled Faces in the Wild (LFW) dataset, surpassing previous methods and proving its effectiveness in real-world face recognition applications.

5. VGG-Net 16: Parkhi *et al.,* [52] proposed VGGFace, implementing preprocessing steps like 2D affine alignment and augmenting the data using random cropping and flipping techniques to enhance training data diversity. The model architecture is based on the VGGNet-16 network with triplet-loss embedding, effectively addressing performance issues related to discriminative learning. This approach led to a significant 68% reduction in error rates on benchmarks like YouTube Faces and the Labeled Faces in the Wild (LFW), demonstrating the effectiveness of simpler architectures in achieving state-of-the-art results. The findings highlight the critical balance between data quality and model design for optimal performance.

6. ResNet: The authors in [53] investigate the effectiveness of the Congenerous Cosine (COCO) loss function in enhancing facial recognition through a ResNet architecture, achieving a face verification accuracy of 99.86% on the Labeled Faces in the Wild (LFW) dataset and 76.57% in the MegaFace challenge with 1 million distractors. By optimizing cosine distances, COCO improves intra-class similarity and inter-class variation, demonstrating significant advantages over traditional loss functions. The authors conclude that COCO provides a stable and effective approach for learning discriminative features. Since 2017, many proposed methods have favored the use of the ResNet architecture and its variations [54], [55], [56].

7. 3D CNNs: Numerous researchers [57], [58] have employed deep 3D convolutional networks (3D CNNs) and contributed to refining their architecture for emotion recognition, owing to their advanced capabilities in image recognition [59]. For example, various methods have utilized

RGB frames fused by leveraging the temporal dimension using the eNTERFACE05 dataset [60]. 3D CNNs utilize 3D convolutional filters to capture essential features from compressed sequences of RGB frames, allowing for the extraction of information in both the spatial and temporal domains [61].

8. Hybrid CNN: A hybrid approach combining Long Short-Term Memory (LSTM) networks and convolutional neural networks (CNNs) has been employed for facial emotion recognition using a video dataset. In this approach, the CNN is responsible for feature extraction in individual frames, whereas the LSTM captures the temporal relationships by integrating these features across multiple frames [62].

B. Related Work

Simonyan [26] introduced a two-stream convolutional neural network (ConvNet) architecture designed to effectively capture both temporal and spatial information, enhancing performance in action recognition. The spatial stream processes individual video frames for appearance features, while the temporal stream analyzes dense optical flow for motion. Utilizing multi-task learning across datasets like UCF-101 and HMDB-51, the model achieved significant performance improvements, with a 6% increase over the temporal stream and a 14% increase over the spatial stream when fused. Cheng *et al.,* [27] tackled the challenge of facial expression recognition under partial occlusion, proposing a new approach that combines Gabor features with a deep learning framework. The preprocessing steps include segmenting and normalizing facial images. The architecture consists of three hidden layers designed to effectively compress high-dimensional Gabor features, addressing performance issues related to occlusion by fine-tuning the model through gradient descent. Evaluation was performed using the JAFFE database, where the proposed method achieved an overall accuracy of 85.71%, with specific accuracies of 82.86% for mouth occlusion and 81.45% for eye occlusion, demonstrating significant improvements over traditional methods.

Li *et al.,* [64] introduced a novel approach for facial expression recognition (FER) using a 2-channel CNN, combining a convolutional autoencoder with a standard CNN for feature extraction. By leveraging unsupervised learning, the model achieved impressive results on the JAFFE dataset, surpassing previous methods with an average accuracy of 95.8% in a leave-one-out experiment and 94.1% in a ten-fold cross-validation experiment. Zhang *et al.,* [65] investigated domain-specific data augmentation in face recognition through a novel face synthesis method applied to the CASIA WebFace dataset. By introducing pose, shape, and expression variations, a CNN was trained to achieve state-of-the-art results comparable to systems trained on millions of images. The approach demonstrated significant performance improvements, with notable enhancements in metrics such as True Accept Rate (TAR) at False Accept Rate (FAR) of 0.01 and 0.001.

Wang *et al.,* [17] proposed a network architecture that includes four inception layers and two convolutional layers with max pooling, designed to efficiently extract intricate facial expression features. The study demonstrated improved performance compared to traditional methods, revealing enhanced recognition accuracy across multiple datasets. In [18], the system they developed employs a CNN architecture consisting of two convolutional layers, two sub-sampling layers, and one fully connected layer designed for the extraction of higher-level visual features essential for expression recognition. The proposed system integrated preprocessing steps such as correcting rotated data, cropping, down-sampling, and intensity normalization. Additionally, data augmentation was employed to enhance the database size and improve the model's robustness. By combining these elements, the system effectively addresses limited data challenges, learns intricate patterns, and achieves promising accuracy improvements in facial expression recognition tasks.

Sahli *et al.,* [66] proposed a Deep Fusion Convolutional Neural Network (DF-CNN) for multimodal 2D+3D FER, integrating both 2D and 3D facial data through six distinct types of 2D attribute maps. The architecture utilizes convolutional layers, ReLU, and pooling layers for feature extraction, followed by fusion layers that merge the extracted features to produce a 32-dimensional fused deep feature. Preprocessing involves generating attribute maps from 3D scans, while data augmentation techniques are applied to enhance training. The CNN architecture addresses performance by taking advantage of pre-trained models and random initialization.

In [67], a hybrid model was introduced that combines CNNs and recurrent neural networks (RNNs) for FER. The CNN component handled the spatial features by extracting them from individual frames, while the RNN was responsible for capturing the temporal dependencies present across the sequences of frames. An overall accuracy of 91.20% was initially achieved by the model, which later improved to 94.46% with the incorporation of Rectified Linear Units (ReLU) activation. The CNN architecture has six convolutional layers with increasing filter sizes (8, 16, 32, 64, 128, and 256) and employs max-pooling layers to reduce dimensionality, addressing overfitting issues through dropout techniques.

In [12], a CNN model was developed for recognizing students' facial expressions using the FER 2013 database. This model, comprised of four convolutional layers and four max-pooling layers, achieved a 70% accuracy rate after 106 training epochs. It effectively identified happy and surprised expressions but struggled with fearful expressions, often misclassifying them as sad. Zhai *et al.,* [68] explored the automatic recognition of students' cognitive states in e-learning environments through facial expressions using a hybrid CNN model. By combining CNN-extracted features with manually engineered pose estimator features, the accuracy achieved by the model was 51.9% on a real-time e-

learning dataset, surpassing existing methods. Performance varied across datasets, with percentages ranging from 33.15% to 53.42% for DAiSEE, 51.28% to 99.95% for CK+, and 21.42% to 85.7% for JAFFE.

Zhang *et al.,* [69] proposed a method for FER by combining a Local Gravitational Force Descriptor with deep convolutional neural networks (DCNN). The technique involved preprocessing steps such as image rotation and data augmentation to enhance dataset diversity. The DCNN architecture featured two branches for extracting local and holistic features, achieving impressive results with average recognition accuracies of 78% for FER2013, 98% for JAFFE, 98% for CK+, 96% for KDEF, and 83% for RAF databases.

In [70], advancements in facial emotion recognition (FER) were explored through the application of transfer learning (TL) within DCNNs. The authors introduced a novel FER method that incorporated TL with a pipeline training strategy, resulting in impressive accuracy of 99.52% on the JAFFE dataset and 98.78% on the KDEF dataset, demonstrating the effectiveness of their approach. The architecture of the CNN involved using pre-trained models with modified upper dense layers tailored for emotion recognition, which effectively addressed performance issues by leveraging learned features and reducing overfitting through fine-tuning.

In [71], a novel CNN architecture was presented to improve facial expression recognition, particularly in challenging conditions involving occlusions and head tilts. The model leveraged the Viola-Jones algorithm for face detection, while Local Binary Patterns (LBP) were used for feature extraction. It demonstrated impressive accuracy rates of 92.66% and 94.94% in two different experimental setups using the CK+ and JAFFE datasets, respectively. The architecture consists of five convolutional layers, a fully connected layer, and a softmax layer, which work together to effectively extract features and classify emotions, ultimately enhancing the overall performance of the system.

In [72], a novel facial recognition technique was introduced that merges a one-dimensional DCNN with linear discriminative analysis (LDA). This method involves preprocessing images from the MUCT dataset through grayscale conversion and histogram equalization, followed by face detection using the Viola-Jones algorithm and feature extraction via LDA. The model was trained on 70% of the dataset, and evaluation was conducted on the remaining 30%, attaining flawless performance metrics: 100% accuracy, precision, recall, and F-measure. These findings illustrate the model's proficiency in managing various appearances in facial conditions.

In [73], a framework was created that combines Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) for feature extraction, followed by a CNN structure that includes three convolutional layers with max pooling, dense layers, and a softmax classifier. This architecture effectively addresses performance challenges by incorporating dropout layers to minimize overfitting and enhance generalization. The model achieved impressive accuracy rates of 98.48% with the HOG-CNN model and 97.96% for the SIFT-CNN model on the CK+ dataset, as well as 91.43% and 82.85% on the JAFFE dataset, respectively, showcasing notable improvements in the reliability of emotion detection.

Finally, in [74], emotion recognition was improved by combining facial expressions with imaging photoplethysmography (IPPG) signals in a multimodal framework. The study employed machine learning algorithms, including SVM, K-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF) for processing IPPG signals, along with deep learning models like VGG16 and Vision Transformer for analyzing facial expressions. It investigated two fusion strategies: decision-level fusion and feature-level fusion. Preprocessing steps included extracting and reconstructing IPPG signals from facial videos, followed by face detection using the RetinaFace algorithm. The results demonstrate significant improvements in accuracy, with feature-level fusion achieving 72.37% for arousal and 70.82% for valence, while SVM recorded the highest accuracy for IPPG at 61.09% for arousal.

In [75], a hybrid CNN-RNN model for FER was created utilizing the Emognition Wearable Dataset 2020, which includes a variety of emotions like amusement, awe, liking, and enthusiasm. The CNN architecture incorporates time-distributed layers that handle sequential video frames, effectively capturing both spatial and temporal features. This design improves performance by reducing parameters and enhancing inference speeds. The tailored CNN-RNN model attained an accuracy of 66%, outperforming traditional methods and highlighting its potential for more nuanced emotion detection.

III. DATASETS FOR FACIAL EMOTION RECOGNITION FROM IMAGES OR VIDEOS

Datasets designed for the recognition of emotion through facial images or videos have a long history, beginning with early studies. The demand for computational methods to analyze emotions has led to the creation of numerous facial expression datasets, which vary by acquisition environment, recognizable expressions, and geographical regions [39]. Most existing datasets primarily consist of 2D video sequences or static images, while a few incorporate 3D images. Each dataset varies significantly in terms of image quantity and size [3]. These datasets usually categorize the six basic emotions: neutral, happiness, disgust, anger, fear, surprise, and sadness, and they can be gathered in either controlled or natural settings [22]. This section presents an overview of several prominent and frequently utilized datasets for emotion recognition, accompanied by a summary in Table I. Over time, these datasets have also incorporated factors such as variations in illumination, facial poses, gender, demographics, ethnicity, age, image quality, and participant count, all of which influence their quality and the effectiveness of emotion recognition algorithms.

TABLE I OVERVIEW OF FREQUENTLY USED FER DATASETS

Dataset and Reference were Used	Data Configuration	Types of Emotions	Recognition Algorithms Used
CK+ [71]	(i) 593 video sequences (ii) 123 unique participants of diverse genders and backgrounds, aged between 18 and 50 (iii) Recorded at 30 FPS with a resolution of either 640 × 480 or 640 × 490	Six basic emotions, along with neutrality and contempt	An innovative architecture featuring five convolutional layers, a fully connected layer employing ReLU activation, and a SoftMax layer.
FER-2013 [12]	(i) A collection of 35,887 grayscale images sourced using google Search (ii) Facial images scaled to 48×48 pixels, showcasing a variety of expressions	Six basic emotions and neutral	A CNN featuring 4 convolutional and max-pooling layers, and 2 fully connected layers
AffectNet [76]	Over 440,000 images gathered from the internet.	Six basic emotions and neutral	A CNN integrating a squeeze-and-excitation network with ResNet
DISFA Denver Intensity of the Spontaneous Facial Action [13]	(i) Stereo videos featuring 12 females and 15 males from diverse ethnic backgrounds (ii) Image resolution of 1024 × 768 (iii) 66 facial landmark points (iv) DISFA+ with 5 levels of intensity for twelve FACS actions (v) Extension of the DISFA dataset.	Intensity of 12 AUs coded	A hybrid model combining DenseNet201 and MobileNet V3
KDEF	(i) A collection of 4,900 images depicting human facial emotions from 70 individuals (ii) Taken from five distinct angles (iii) Comprising 35 males and 35 females aged between 20 and 30 (iv) Each image has a resolution of 562 × 762 (v) Subjects are without beards, eyeglasses, earrings, or mustaches, and have minimal makeup.	Six basic emotions and neutral	
DISFA+[9]	(i) This dataset is an extension of the DISFA database. (ii) It features manually labeled frame-based annotations that categorize 12 FACS facial actions with 5 levels of intensity.	5-level intensity of twelve FACS	Hybrid CNN
Oulu-CASIA [77]	2,880 videos recorded under three different lighting conditions.	Six basic emotions	Fine-tuned VGG-Face Model
BU-3DFE [18]	2,500 3D facial images captured from two angles: -45° and +45°.	Six basic emotions and neutral	Novel CNN
JAFFE [64]	(i) 213 images showcasing different facial expressions (ii) Ten distinct Japanese females (iii) Image resolution of 256 × 256	Six basic emotions and neutral	Multi-channel Convolutional Neural Network (MC-CNN)
GEMEP FERA	289 images sequences	Sadness, Happy fear Anger, Relief	
MMI [67]	2,900 videos categorized by neutral, onset, apex, and offset.	Six basic emotions and neutral	Hybrid CNN-RNN model with the ReLU
SFEW [78]	700 images featuring varying ages, occlusions, head poses and lighting conditions.	Six basic emotions and neutral	A novel end-to-end CNN architecture
MultiPIE [79]	Over 750,000 images taken from 15 angles and under 19 lighting conditions.	Anger, Happy, Scream, Disgust, Neutral, Squint, Surprise	CNN
RAFD-DB [23]	30,000 images sourced from the real world.	Six basic emotions and neutral	Transfer Learning CNN

IV. DISCUSSION OF ADVANCEMENTS AND METHODOLOGY ANALYSIS

Table II presents a summary of the advancements in facial emotion recognition (FER) methodologies, focusing on the various techniques employed and their performance outcomes. It illustrates the effectiveness of different approaches in improving emotion detection capabilities within the field.

TABLE II SUMMARY OF ADVANCEMENTS IN FACIAL EMOTION RECOGNITION METHODOLOGIES

Author/Date and Dataset Used	Advancements and Methodology Analysis
K. Simonyan [63]	The findings of K. Simonyan [63] underscore the complementary nature of spatial and temporal features in action recognition tasks. The significant performance improvements highlight the effectiveness of the two-stream architecture in leveraging both types of information, which are crucial for accurately recognizing actions in videos. The use of multi-task learning further demonstrates the potential for improved model performance through the integration of diverse training data.
Y. Cheng *et al.,* [80]	Y. Cheng *et al.,* [80] reported enhanced robustness in expression recognition for real-world applications, where occlusion is common. However, the study's limitations include a small and homogeneous dataset, which may affect generalizability, and the simulated nature of the occlusions, which do not fully represent real-world scenarios. Future research should focus on expanding the dataset to include diverse expressions and demographics, as well as exploring the effects of real-world occlusions and alternative feature extraction methods to further improve recognition accuracy and robustness.
D. Hamester *et al.,* [64]	In this work, the findings demonstrate the capability of unsupervised learning in facial emotion recognition (FER) tasks. The results underscore the effectiveness of synthesizing face images for training, providing a cost-effective alternative to extensive data collection while maintaining competitive performance levels.
A. Mollahosseini *et al.,* [17]	In their study, A. Mollahosseini *et al.,* [17] proposed an architecture that achieved superior accuracy in subject-independent and cross-database evaluations compared to traditional CNN methods. Experimental results validate the proposed network's effectiveness in achieving higher accuracy while reducing the computational complexity required for training, emphasizing its potential for advancing facial emotion recognition (FER) technology.
A. T. Lopes *et al.,* [18]	In their study, A. T. Lopes *et al.,* [18] included preprocessing steps and data augmentation to effectively address limited data challenges, learn intricate patterns, and achieve promising accuracy improvements in facial expression recognition tasks.
H. Li *et al.,* [66]	In their study, H. Li *et al.,* [66] results show that the Deep Fusion Convolutional Neural Network (DF-CNN) outperforms alternative models, achieving an accuracy of 86.86%, indicating its potential for accurate facial expression recognition.
N. Jain *et al.,* [67]	The findings by N. Jain *et al.,* [67] are noteworthy, as the achieved results highlight the model's potential to reduce errors in emotion detection, making it valuable for applications in monitoring mental health and interactions between machines and humans. The architecture effectively captures spatial features, while the RNN component manages temporal dependencies, significantly improving emotion detection performance.
I. Lasri [12]	The findings of I. Lasri [12] highlight the potential of using deep learning techniques to enhance classroom dynamics by allowing educators to adapt their teaching strategies based on real-time emotional feedback from students.
K. P. Rao *et al.,* [68]	The findings of K. P. Rao *et al.,* [68] revealed the potential of leveraging facial expressions to provide personalized feedback to instructors, enhancing teaching effectiveness and student engagement in online learning.
K. Mohan *et al.,* [69]	In their study, K. Mohan *et al.,* [69] reported that their model outperforms twenty-five state-of-the-art methods, emphasizing the effectiveness of integrating locally obtained and holistic features in deep learning models for facial expression tasks. By effectively combining shallow and major DCNN designs and integrating local and holistic features, the model demonstrated superior accuracy compared to traditional methods.
M. A. H. Akhand *et al.,* [70]	M. A. H. Akhand *et al.,* [70] reported results that significantly outperform traditional feature-based methods, particularly in recognizing emotions from non-frontal or angularly taken images, indicating the method's potential for real-world applications.

A. S. Qazi *et al.,* [71]	The study by A. S. Qazi *et al.,* [71] demonstrates that their architecture addresses performance issues by employing regularization techniques and optimizing hyperparameters, resulting in impressive accuracy. These results highlight the model's potential for applications in machine-human interaction, diagnosing mental health, and assistive technologies, demonstrating its effectiveness in recognizing emotions despite common challenges.
J. Mohammed *et al.,* [72]	The study by J. Mohammed *et al.,* [72] shows that their architecture of the 1D-DCNN is designed to effectively classify the 1D feature vectors extracted by linear discriminative analysis (LDA), addressing performance issues by optimizing feature representation and reducing dimensionality.
C. Gautam *et al.,* [73]	In the work conducted by C. Gautam *et al.,* [73], the results highlight the effectiveness of combining handcrafted features with deep learning, significantly improving emotion detection reliability, which is crucial for various real-world applications.
X. Tao *et al.*, [74]	The findings of X. Tao *et al.*, [74] highlight the effectiveness of combining modalities for emotion detection, offering a reliable framework applicable in various real-time scenarios. Moreover, the CNN architecture employs a multi-view feature fusion approach, effectively capturing complementary information from both modalities, which addresses performance issues related to single-modality systems.
H. V. Manalu [75]	H. V. Manalu [75] focused on addressing the challenge of distinguishing closely related emotions. The significance of these findings lies in their implications for enhancing human-computer interactions across various sectors, including healthcare and consumer analytics.

V. CONCLUSION AND FUTURE DIRECTION

This paper reviewed a decade of trends and perspectives on advancements in Facial Emotion Recognition (FER), highlighting developments in convolutional neural network (CNN) architectures. We examined various databases, including those with spontaneous and lab-generated images (see Table I), to enhance emotion detection accuracy. The discussion emphasizes that natural human-computer interaction is becoming increasingly seamless, as evidenced by the high accuracy achieved by researchers, which highlights the growing capability of machines to interpret emotions effectively. However, FER systems are still limited to basic emotions and may not capture the full complexity of human feelings. Future research should focus on creating larger databases and developing advanced deep learning architectures to recognize both basic and nuanced emotions. The adoption of deep neural networks (DNNs) marked a significant breakthrough in face recognition technology. Systems employing DNNs have attained accuracy levels exceeding 99%, even when tested on extensive datasets of facial images gathered in uncontrolled settings. However, several recent studies since 2018 [27], [28], [81] have revealed that the performance of these systems declines when processing images captured under challenging conditions, such as those with low resolution, significant lighting variations, blur, or noise, commonly referred to as semantic adversarial attacks [82]. Consequently, there is a growing need for research aimed at enhancing the robustness of deep learning methods in such adverse conditions. New approaches are also emerging to verify the resilience of these models against semantic disruptions [82]. Additionally, a shift from unimodal to multimodal approaches in emotion recognition enhances the reliability and accuracy of detection systems [74]. Researchers are advancing multimodal deep learning approaches, such as those combining audio and visual inputs studied by [48]. Integrating advanced feature extraction methods with CNN architectures improves performance and generalization [73]. Furthermore, there is a growing focus on applying FER technologies to real-world settings, including human-computer interaction, mental health diagnostics, and educational environments [14], [68].

REFERENCES

[1] A. J. Shawon, A. Tabassum, and R. Mahmud, "Emotion detection using machine learning: An analytical review," *J. Inf. Technol. Manag.*, vol. 4, no. 1, pp. 32-43, 2024.

[2] A. V. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions," *Inf. Fusion*, vol. 105, no. December 2023, 2024, doi: 10.1016/j.inffus.2023.102218.

[3] W. Mellouk and W. Handouzi, "Facial emotion recognition using deep learning: Review and insights the future," *Procedia Comput. Sci.*, vol. 175, pp. 689-694, 2020, doi: 10.1016/j.procs.2020.07.101.

[4] N. Ahmed, Z. Al Aghbari, and S. Girija, "A systematic survey on multimodal emotion recognition using learning algorithms," *Intell. Syst. Appl.*, vol. 17, no. December 2022, p. 200171, 2023, doi: 10.1016/j.iswa.2022.200171.

[5] J.-H. Byun, S.-P. Kim, and S.-P. Lee, "Multi-modal emotion recognition using speech features," *Appl. Sci.*, 2021.

[6] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion recognition from multimodal physiological signals for emotion aware healthcare systems," *J. Med. Biol. Eng.*, vol. 40, no. 2, pp. 149-157, 2020, doi: 10.1007/s40846-019-00505-7.

[7] M. Taskiran, N. Kahraman, and C. E. Erdem, "Face recognition: Past, present and future (a review)," *Digit. Signal Process.*, vol. 106, p. 102809, 2020, doi: 10.1016/j.dsp.2020.102809.

[8] D. Hutchison, *Survey High-Performance Modelling and Simulation for Big Data Applications*, 2019, doi: 10.1007/978-3-030-16272-6.

[9] A. Pise, H. Vadapalli, and I. Sanders, "Facial emotion recognition using temporal relational network: An application to e-learning," *Multimed. Tools Appl.*, vol. 81, no. 19, pp. 26633-26653, 2022, doi: 10.1007/s11042-020-10133-y.

[10] S. AlZu'bi *et al.,* "A novel deep learning technique for detecting emotional impact in online education," *Electronics*, vol. 11, no. 18, pp. 1-24, 2022, doi: 10.3390/electronics11182964.

[11] T. Kumar Arora et al., "Optimal facial feature based emotional recognition using deep learning algorithm," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/8379202.

[12] I. Lasri, "Facial emotion recognition of students using convolutional neural network," in *Proc. 2019 Third Int. Conf. Intell. Comput. Data Sci.*, pp. 1-6, 2019.

[13] A. Rahaman and W. Sait, "Developing a pain identification model using a deep learning technique," vol. 3, pp. 1-9, 2024, doi: 10.57197/JDR-2024-0028.

[14] X. Peng, "Research on emotion recognition based on deep learning for mental health," *Inform.*, vol. 45, no. 1, pp. 127-132, 2021, doi: 10.31449/inf.v45i1.3424.

[15] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, no. March 2019, pp. 103-126, 2020, doi: 10.1016/j.inffus.2020.01.011.

[16] X. Chen, S. Member, and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, 2014.

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

[18] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610-628, 2017, doi: 10.1016/j.patcog.2016.07.026.

[19] T. Le Paine and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proc. IEEE Int. Conf. Image Process.*, pp. 19-27, 2015.

[20] P. V. Rouast, M. T. P. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 524-543, 2021, doi: 10.1109/TAFFC.2018.2890471.

[21] P. Viola and M. J. Jones, "Robust real-time face detection," in *Proc. IEEE Int. Conf. Comput. Vision* (ICCV), 2004.

[22] J. X. Y. Lek and J. Teo, "Academic emotion classification using FER: A systematic review," *Hum. Behav. Emerg. Technol.*, vol. 2023, 2023, doi: 10.1155/2023/9790005.

[23] H. Zhou et al., "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *Proc. 2019 IEEE Int. Conf. Multimodal Intell.* (ICMI), 2019.

[24] Y. Taigman, M. A. Ranzato, T. Aviv, and M. Park, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), 2014, doi: 10.1109/CVPR.2014.220.

[25] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), pp. 1891-1898, 2014, doi: 10.1109/CVPR.2014.244.

[26] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), pp. 815-823, 2015, doi: 10.1109/CVPR.2015.7298682.

[27] M. Wang and W. Deng, "Deep face recognition: A survey," *arXiv preprint*, pp. 1-31, 2019.

[28] K. Grm, V. Struc, A. Artiges, M. Caron, and H. K. Ekenel, "Strengths and weaknesses of deep learning models for face recognition against image degradations," *IET Biometrics*, vol. 7, no. 1, pp. 81-89, 2018, doi: 10.1049/iet-bmt.2017.0083.

[29] Z. Zhang, W. Shen, S. Qiao, Y. Wang, B. Wang, and A. Yuille, "Robust face detection via learning small faces on hard images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1361-1370, 2020.

[30] R. Ranjan *et al.,* "Deep learning for understanding faces," *arXiv preprint*, no. January, 2018.

[31] Y. Li, B. Sun, T. Wu, and Y. Wang, "ConvNet and a 3D model," *arXiv preprint* arXiv:1606.00850v3, pp. 1-16, 2016.

[32] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proc. 2017 IEEE Winter Conf. Appl. Comput. Vision*, pp. 1-6, 2017.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 1-14, 2016.

[34] S. S. Farfade, "Multi-view face detection using deep convolutional neural networks," *Master's thesis*, 2015.

[35] H. Li, Z. Lin, X. Shen, and J. Brandt, "A convolutional neural network cascade for face detection," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015.

[36] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 3, pp. 3676-3684, 2015.

[37] S. Yang, Y. Xiong, C. C. Loy, and X. Tang, "Face detection through scale-friendly deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

[38] T. Zheng and W. Deng, "Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments," *arXiv preprint*, 2018.

[39] F. Z. Canal *et al.,* "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Inf. Sci. (Ny)*, vol. 582, pp. 593-617, 2022, doi: 10.1016/j.ins.2021.10.005.

[40] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803-816, 2009, doi: 10.1016/j.imavis.2008.08.005.

[41] T. Jabid, M. H. Kabir, and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI J.*, vol. 32, no. 5, pp. 784-794, 2010, doi: 10.4218/etrij.10.1510.0132.

[42] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," *arXiv preprint*, 2017.

[43] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4068-4074, 2015.

[44] Y. H. Liu, "Feature extraction and image recognition with convolutional neural networks," *J. Phys. Conf. Ser.*, vol. 1087, no. 6, 2018, doi: 10.1088/1742-6596/1087/6/062032.

[45] C. Dev and A. Ganguly, "Sentiment analysis of review data: A deep learning approach using user-generated content," *arXiv preprint*, vol. 12, no. 2, pp. 28-36, 2023.

[46] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195-1215, 2022, doi: 10.1109/TAFFC.2020.2981446.

[47] A. Hussain, N. Saikia, and C. Dev, "Advancements in Indian sign language recognition systems: Enhancing communication and accessibility for the deaf and hearing impaired," *arXiv preprint*, vol. 12, no. 2, pp. 37-49, 2023.

[48] Z. Farhoudi and S. Setayeshi, "Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition," *Speech Commun.*, vol. 127, pp. 92-103, 2021, doi: 10.1016/j.specom.2020.12.001.

[49] Y. Sun, Y. Chen, and X. Wang, "Deep learning face representation by joint learning," *arXiv preprint*, pp. 1-9, 2014.

[50] X. Tang, "DeepID3: Face recognition with very deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2-6, 2015.

[51] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, and R. Chellappa, "Joint face detection and alignment using multi-task cascaded convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 1-5, 2016.

[52] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *BMVC*, vol. 3, no. Section 3, 2015.

[53] Y. Liu, H. Li, and X. Wang, "Rethinking feature discrimination and polymerization for large-scale recognition," in *Advances in Neural Information Processing Systems*, NIPS, 2017.

[54] X. Zhang, Z. Fang, Y. Wen, and Z. Li, "Range loss for deep face recognition with long-tailed training data," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5409-5418, 2017.

[55] R. Ranjan, C. D. Castillo, and R. Chellappa, "L2-constrained softmax loss for discriminative face verification," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[56] J. Milgram, "Von Mises-Fisher mixture model-based deep learning: Application to face verification," unpublished, pp. 1-16.

[57] S. Prasanna, T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks," in *Proceedings of the*

IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 1-8.

[58] L. Zhao, Z. Wang, and G. Zhang, "Facial expression recognition from video sequences based on spatial-temporal motion local binary pattern and Gabor multiorientation fusion histogram," *Comput. Intell. Neurosci.*, vol. 2017, 2017, Art. no. 7206041, doi: 10.1155/2017/7206041.

[59] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 30-40.

[60] X. Pan, G. Ying, G. Chen, H. Li, and W. Li, "A deep spatial and temporal aggregation framework for video-based facial expression recognition," *IEEE Access*, vol. 7, pp. 48807-48815, 2019, doi: 10.1109/ACCESS.2019.2907271.

[61] H. Ye, Z. Wu, R. Zhao, X. Wang, Y. Jiang, and X. Xue, "Evaluating two-stream CNN for video classification," unpublished, 2015.

[62] X. Pan, W. Guo, X. Guo, W. Li, J. Wu, and J. Jinzhao, "Deep temporal-spatial aggregation for video-based," *Symmetry*, vol. 11, no. 1, Art. no. 52, 2019, doi: 10.3390/sym11010052.

[63] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1-9, 2014.

[64] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, 2015, pp. 1787-1794.

[65] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Do we really need to collect millions of faces for effective face recognition?" in *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016.

[66] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network," *IEEE Trans. Multimed.*, vol. 19, no. 12, pp. 2816-2831, 2017, doi: 10.1109/TMM.2017.2713408.

[67] N. Jain, S. Kumar, A. Kumar, and P. Shamsolmoali, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognit. Lett.*, vol. 115, pp. 101-106, 2018, doi: 10.1016/j.patrec.2018.04.010.

[68] K. P. Rao, M. V. P. Chandra, and S. Rao, "Recognition of learners' cognitive states using facial expressions in e-learning environments," *J. Univ. Shanghai Sci. Technol.*, vol. 22, no. 12, pp. 93-103, 2020.

[69] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-13, 2021, doi: 10.1109/TIM.2020.3031835.

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

[71] A. S. Qazi, M. S. Farooq, F. Rustam, M. G. Villar, C. L. Rodríguez, and I. Ashraf, "Occlusions and tilt in facial recognition systems," *Appl. Sci.*, vol. 12, no. 1, 2022.

[72] J. Mohammed, E. I. Abbas, and Z. M. Abood, "A facial recognition using a combination of a novel one dimension deep CNN and LDA," *Mater. Today Proc.*, vol. 80, pp. 3594-3599, 2023, doi: 10.1016/j.matpr.2021.07.325.

[73] C. Gautam and K. R. Seeja, "Facial emotion recognition using handcrafted features and CNN," *Procedia Comput. Sci.*, vol. 218, pp. 1295-1303, 2023, doi: 10.1016/j.procs.2023.01.108.

[74] X. Tao *et al.,* "Facial video-based non-contact emotion recognition: A multi-view features expression and fusion method," *Biomed. Signal Process. Control*, vol. 96, 2024.

[75] H. V. Manalu and A. P. Rifai, "Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm," *Intell. Syst. Appl.*, vol. 21, p. 200339, 2024, doi: 10.1016/j.iswa.2024.200339.

[76] Z. Y. Huang, C. C. Chiang, J. H. Chen, Y. C. Chen, and H. L. Chung, "A study on computer vision for facial emotion recognition," *Sci. Rep.*, pp. 1-13, 2023, doi: 10.1038/s41598-023-35446-4.

[77] W. Dias and F. Andal, "Cross-dataset emotion recognition from facial expressions through convolutional neural networks," *J. Vis. Commun.*, 2023.

[78] P. M. Ferreira and A. N. A. Rebelo, "Physiological inspired deep neural networks for emotion recognition," *IEEE Access*, vol. 6, pp. 53930-53943, 2020, doi: 10.1109/ACCESS.2018.2870063.

[79] S. Saxena, S. Tripathi, and T. S. B. Sudarshan, "An intelligent facial expression recognition system with emotion intensity classification," *Cogn. Syst. Res.*, vol. 74, pp. 39-52, 2022, doi: 10.1016/j.cogsys.2022.04.001.

[80] Y. Cheng, B. Jiang, and K. Jia, "A deep structure for facial expression recognition under partial occlusion," in *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, 2014, doi: 10.1109/IIH-MSP.2014.59.

[81] S. Ullah, J. Ou, Y. Xie, and W. Tian, "Facial expression recognition (FER) survey: A vision, architectural elements, and future directions," *PeerJ Comput. Sci.*, 2024, doi: 10.7717/peerj-cs.2024.

[82] J. Mohapatra, T.-W. Weng, P.-Y. Chen, S. Liu, and L. Daniel, "A family of semantic perturbations," unpublished, 2020.