Vokasi Unesa Bulletin of Engineering, Technology and Applied Science (VUBETA)
https://journal.unesa.ac.id/index.php/vubeta
Vol. 3, No. 2, 2026, pp. 315~329 DOI: 10.26740/vubeta.v3i2.46974
ISSN: 3064-0768

Leveraging State-of-the-art Deep Learning Advancements for Emotion Detection: A Comprehensive Review and Insights

Krishna Kant[1]
[1] Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charusat University, Changa, Gujarat 388421, India

Article Info

Article history:

Received November 9, 2025
Revised February 26, 2026
Accepted April 21, 2026

Keywords:

Affective Computing
Emotion Recognition
Human Face
Facial Expression
Deep Learning

ABSTRACT

Emotion recognition is a fundamental aspect of affective computing, focusing on identifying and interpreting human emotional states. Among various modalities, facial emotion recognition has gained significant attention due to its non-intrusive nature and extensive applicability across domains such as e-learning, healthcare, marketing, e-commerce, and psychology. A wide range of approaches has been employed to address the challenges inherent in facial emotion classification. There remains a lack of a holistic, structured framework that critically evaluates both the advantages and shortcomings of deep networks while introducing attention-based and Transformer-driven models. Therefore, to address this gap, this paper presents a systematic review of peer-reviewed FER studies of deep learning models published between 2022 and 2025. This paper presents the study of advanced deep learning architectures for facial emotion detection, emphasizing the predominance of Deep Learning models including Transformer-based architectures, hybrid CNN–Transformer models, spatiotemporal learning approaches, and novel attention mechanisms. This research work provides analysis of deep learning model architectures, learning strategies, datasets, evaluation protocols, and performance metrics reported in state-of-the-art FER research. It identifies common issues including computational complexity, real-world robustness, generalization across datasets, and data imbalance. It also analyzes current research challenges, limitations and their practical significance. Furthermore, this research work identified and discussed the possible opportunities, unresolved issues of human facial emotion recognition and provided the future directions. The objective of this study is to provide actionable insights for researchers and practitioners, guiding future research toward more robust, accurate, and interpretable FER systems.

1. INTRODUCTION

Affective computing integrates technology and psychology to enable intelligent systems to understand and react to human emotions effectively. It is an interface between AI and human affective behavior, directing the development of models that recognize and understand emotion from facial expression. Affective computing has evolved to enable machines to understand human emotions. Hence, affective computing not only enhances deep learning research but also promotes the formulation of a cognitive framework capable of handling natural human interactions. Emotion detection is relevant to affective computing, human-computer interaction, and mental health monitoring because it enables a system to understand and respond to human emotion in real time. Emotion recognition has revealed a rich set of emotion signals that accurately capture the affective states of humans across multiple interactions. Emotion recognition has become a fascinating area of research for many scholars due to its potential to uncover the psychological processes that generate emotions in human beings. Emotions are closely interconnected with both cognitive and physiological processes that influence human decision-making. Emotion recognition has attracted interest as an important component of human-computer

*Corresponding Author
Email: kantkrisna81@gmail.com

interaction. Deep Learning models have become the most prominent for facial emotion classification and emotion forecasting in real-world applications.

According to Ekman et al. [1] states six basic emotions that human beings express during the course of action. The basic emotions are: happy, sad, angry, fear, surprise, and disgust, according to the theory of emotions. The human face is the primary source for detecting expressions and extracting emotions from images, audio, and videos. It plays a vital role in detecting emotion and reading human psychology. Facial expression conveys a series of emotions that characterize an individual's behavior. Numerous features of the human face are extracted from facial expressions to detect emotional states. Ekman's [2], 1971 identified six basic universal expressions in the 20th century based on cross-cultural research: joy, sorrow, anger, surprise, fear, and disgust. In general, Facial Expression Recognition (FER) involves four key stages: face detection and pre-processing, facial feature extraction, emotion categorization, and classification.

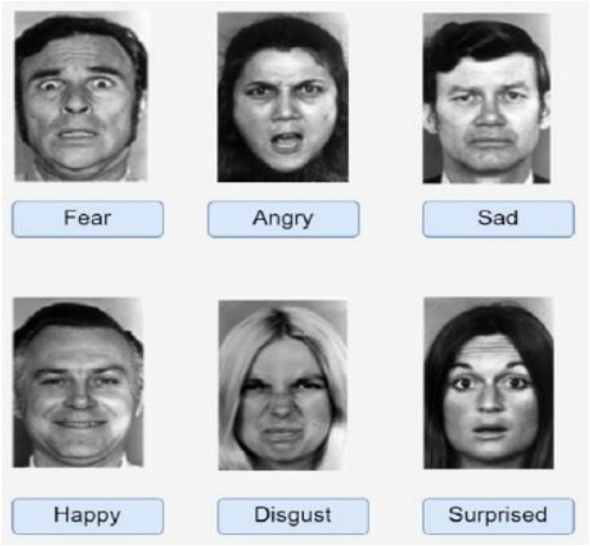

Figure 1. Ekman's Emotion.

According to Sailunaz et al. [3], and Calvo and Kim [4], the current emotion models fall into two classes: categorical and dimension-based. Nearly all of the potential human emotions have been studied by many researchers [5]-[9]. The categorical emotion model is centered on discrete emotion labels. According to the category paradigm, there are distinct emotional types. A superset of Ekman's eight fundamental bipolar emotions, plus two more, TRUST and ANTICIPATION, is defined by Plutchik [8]. These eight emotions are categorized into four bipolar sets: surprise versus anticipation, anger versus fear, trust versus disgust, and joy versus sadness. Ekman et al. [10] came to the conclusion that the six fundamental emotions are anger, disgust, fear, happiness, sadness, and surprise. Martínez-Barco [11] stated that the feelings are dispersed in a circular, two-dimensional area: Arousal and valence dimensions.

Classifying emotions is important for effective communication, which makes it a critical component of human–computer interaction. Emotion is a multifaceted psychological and physiological state triggered by internal or external events. There are three interconnected parts to it:

a) Subjective experience: The individual's inner feelings, such as joy, rage, or sorrow.
b) Body reactions like variations in heart rate, hormone levels, or brain signals are referred to as physiological responses.
c) Behavioral expression refers to acts that may be observed, like posture, vocal tones, facial emotions, and gestures.

Emotion is a complex phenomenon with behavioral, physiological, and psychological components that enable both people and animals to respond appropriately to internal and external stimuli. Fundamental to the human experience, emotions affect survival, social interaction, decision-making, and thought processes.

Triandis et al. [13] stated that early researchers hypothesized that emotion, whether simulated or natural, could not be accurately conveyed by the human face. However, this idea was refuted by Ekman's research [14], which showed that facial expressions do, in fact, serve as trustworthy indicators of emotion. Ekman acknowledged that both posed and spontaneous expressions were more accurate than chance levels and blamed earlier failures on methodological errors. The human face is the key indicator from which emotions can be extracted to understand an individual's mental state. By utilizing developments in computer vision and artificial

intelligence, facial expression recognition, or FER, has become a crucial part of emotion detection. Recent research has greatly improved the precision and generalizability of FER systems in a number of fields. Emotion detection relies heavily on feature extraction, which converts unprocessed face, audio, or physiological data into meaningful representations that a deep learning model can use to reliably identify emotions. Deep learning can automatically learn features, or features can be manually created.

This section provides well-known databases currently in use and often used in FER for both training and testing. When insufficient datasets are used for training, the system's performance suffers. Six basic emotions and neutral, which only include the frontal face with certain difficulties, such as position variation and illumination, are present in the majority of the facial expression dataset.

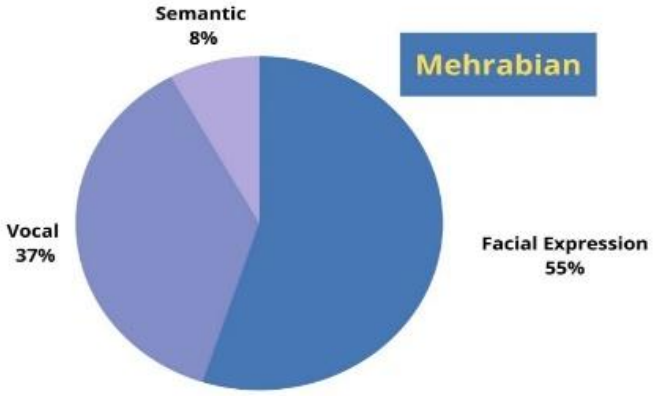

Figure 2. Facial Expression Contribution.

According to Albert Mehrabian et al. [12] stated that the emotion is conveyed 7% through verbal content, 38% through vocal, and 57% from facial expression. FER datasets are often divided into two categories: (1) spontaneous and (2) posed. Since they are recorded organically and exhibit more realistic expressions, spontaneous datasets are often regarded as authentic. Conversely, posed expressions are intentionally produced and managed by the subjects. The corresponding data set is stated in Fig. 3.

A comprehensive literature review was conducted to identify FER research that used Deep Learning techniques. Furthermore, the authors developed and assessed a key research question to pinpoint possible difficulties in FER research using DL methodologies. The analysis and retrieval of FER publications come after this initial stage. In addition to identifying potential study gaps in the designated problem area, the method provides precise instructions for academics, professionals, and businesses seeking to conduct new studies in these fields.

The present technique for deep facial expression identification is dedicated to resolving the following issues: (1) Overfitting because of insufficient training data; (2) Interference issues arise in the real world environment from other variables unrelated to expression; (3) Models have a hard time capturing the exact borders between categories since motions can overlap or appear slightly differently in various people; (4) Deep models find it difficult to preserve temporal coherence when sequential frames contain noise or transitory expressions; (5) Cultural, gender, or age-related differences in how the same emotion is expressed can weaken the trained model's resilience; and (6) Deep architectures are frequently computationally and memory-intensive, which restricts their use in real-time applications or on edge devices.

Even though the number of survey articles on facial emotion detection is increasing, the reviews now available frequently lack a cohesive analytical perspective and instead focus on discrete aspects such as model architectures, feature extraction methods, or benchmark datasets. The majority of earlier research ignores the trade-offs between computational complexity and recognition accuracy, fails to systematically assess cross-dataset generalization, and gives scant attention to practical issues such as pose variation, occlusion, illumination changes, and robustness in unconstrained environments. Moreover, methodological classifications are often limited to data representations, failing to provide significant links between innovative architecture and realistic deployment limitations. This paper proposes a comprehensive framework that integrates various analytical dimensions, including architectural paradigms, input representation techniques, performance generalization across datasets, computational efficiency, and practical applications, to overcome these restrictions. The review provides a thorough and organized overview of current developments, unresolved issues, and future research prospects in face emotion identification by synthesizing earlier work into a cohesive framework.

The main contributions of this study are summarized as follows:

a) Comprehensive Analysis of Emotion Datasets: This study presents an in-depth analysis of widely used datasets for emotion classification, offering critical insights into their representativeness, diversity, and suitability for training and evaluating emotion recognition models. This analysis assists researchers in selecting appropriate datasets for benchmarking and performance enhancement.
b) Systematic Evaluation of Deep Learning Techniques: The paper provides a comprehensive review and comparative evaluation of state-of-the-art deep learning approaches employed for emotion detection and classification, highlighting their strengths, limitations, and performance trends.
c) Assessment of Model Robustness and Reliability: The study addresses the often-overlooked aspect of model robustness by examining the sensitivity of emotion recognition systems to minor perturbations and variations in input data, thereby emphasizing reliability in real-world deployment scenarios.

Overall, this research work provided a rigorous analysis of state-of-the-art approaches to emotion classification, highlighting potential shortcomings of current methodologies and offering guidance for achieving more accurate and robust emotion recognition. The rest of this paper is organized as follows: Section 2 presents the literature review, highlighting deep learning techniques and the contributions of various researchers. Section 3 discusses the results and their interpretation. Section 4 describes the methods employed in emotion detection. Finally, Section 5 concludes the paper by summarizing the key findings and suggesting directions for future research.

2. LITERATURE REVIEW

In this study, the latest developments in CNNs for facial recognition tasks are utilized. The absence of substantial labeled data frequently hinders emotion recognition, in contrast to standard face recognition settings. With a particular focus on differences in image illumination, this work addresses these issues by proposing a novel approach that reduces the problem domain by eliminating complicating factors.

Francesco et al. [17], applied explainable AI to address potential limitations of the signal processing using Deep learning in various real-world applications. They have applied the experiment on images, and they will incorporate a multimodal data set and also improve the explainable capacity in the future. The author has improved the classification accuracy and reduced the computational cost.

M. K. Chowdary et al. [18], has applied deep learning and addressed the limitations that include differences in posture, uneven lighting, and facial accessories. They also added that transfer learning plays a vital role in emotion detection. Author has presented RESNET50, Inception V3, VGG19, and Mobile Net pre-trained networks for emotion detection on the CK+ dataset, achieving an accuracy of 97%. They have suggested implementing various modalities such as audio, video, and text to improve the performance of emotion detection and classification in the future.

Liping Lu [19] has presented CNN and LSTM for detecting the psychological state based on emotions and attained the accuracy on (1) DEAP data set as: Valence (87.6%) and Arousal (89.3%), (2) AMIGOS data set as Valence (88.5.%) and Arousal (90.1 %), and on (3) DEAP + AMIGOS data set as: Valence (91.3.6%) and Arousal (92.5%). The authors have identified computational complexity as a limitation of the proposed work. They also suggested that future studies should focus on real-time deployment and compatibility with wearable technologies.

K.Devarajan et al. [20] had implemented convolutional layers combined with graph feature representations to capture intricate intermodal dynamics, improving classification and reducing computational time. The authors have suggested that GANs are excellent for emotion detection and can leverage multi-modal, real-time evaluation in affective computing. The authors recommend applying their approach in dynamic situations in a scalable and useful way. The authors indicated that their proposed method is suitable for dynamic conditions, offering both scalability and practical applicability in affective computing.

Boitel et al. [21] proposed MIST(Motion, Image, Speech, and Text) multimodal approach for emotion detection, addressing critical challenges in Automatic Emotion Recognition to enhance the emotion recognition system. They have applied deep learning architectures (RNNs, CNNs, hybrid models) for model development on BAUM-1 and SAVEE for emotion detection. In the future, they will apply their work in Social robots, Personal assistants, Educational technologies, and enhanced human-computer interaction systems.

Abdul Aziz et al. [22] has presented affective computing by introducing a transformer framework that Integrates desired analysis with emotion recognition on the multimodal dataset MSED by addressing the conventional approach achieved good performance. The author has not validated cross-dataset validation to assess generalizability. The author claimed better results compared to the state-of-the-art approaches, having an improvement in sentiment analysis by 3%, emotion analysis by 2.2%, and desire analysis by 1%. Their work is limited to social media data and cross-cultural and multilingual generalization. They have suggested that they will include multimodal approaches, Cross-platform and cross-cultural validation, and a lightweight model.

S. Woo et al. [23] developed a deep multimodal emotion detection model that addressed critical challenges in fusing psychological signals to categorize valence and arousal levels on the AMIGOS dataset. This research work includes two key innovations: a modality-aware attention mechanism and a proxy-based multimodal loss function. The author has achieved a significant improvement in emotion-recognition performance. They will improve modalities, cross-validation of the data set, and will also incorporate facial expression and voice in psychological signals in the future.

Alisawai et al. [24] has presented a significant contribution through dataset enhancement and framework Innovation for emotion detection on the FER-213 and CK+ datasets. The author has proposed a CBAM-4CNN architecture based on an attention mechanism, achieving an accuracy of 81%. They have suggested including multimodal fusion and transformer architectures to make the FER more robust for real-time deployment in education and healthcare.

Geetha et al. [25] presented the systematic review of Deep Learning Convolutional Neural Networks (CNNs), (RNNs, LSTMs), Attention mechanisms, Transformer-based architectures for contextual understanding of hybrid models combining multiple DL techniques. Their integrated work provided 10-25% improvement over a single-modality approach on audio-visual datasets, such as IEMOCAP, RAVDESS, MSP-IMPROV, and multimodal datasets, such as CMU-MOSEI, MELD, EmotiW, and physiological datasets, such as AMIGOS, DEAP, MAHNOB-HCI, and Real-World Datasets, Such as AffectNet, FER2013, and AFEW. In the future, they will create adaptive, context-aware systems that can manage the complexity of the actual world, including individual variance, ambient noise, cultural differences, and missing modalities. Their work also highlights that to create emotionally intelligent systems that genuinely comprehend human affective states, computer scientists, psychologists, neuroscientists, and domain experts must work together across disciplinary boundaries.

Shiqing Zhang et al. [26], 2024 state that they have provided a review of deep learning techniques for multimodal emotion recognition, placing more emphasis on feature extraction and information fusion for IEMOCAP, CMU-MOSEI/MOSI, SEMAINE, MELD, and AFEW datasets. The author has reported that transformer-based and hybrid fusion networks achieved accuracies of 80-90% compared to conventional CNNs. For realistic and equitable emotion-aware systems, they have proposed combining scalable multimodal fusion frameworks, ethical AI principles, and cross-disciplinary insights.

Yoonesi et al. [27] conducted a review and meta-analysis in accordance with the PRISMA 2020 recommendations on the effectiveness of deep learning algorithms for identifying changes in facial expression associated with neurological diseases. The author has applied Convolutional Neural Networks (CNNs), along with variants such as ResNet, VGGNet, and MobileNet, to detect neurological disorders using the CK+ and FER-2013 datasets, achieving a highest accuracy of 95%. The author has proposed explainable AI techniques, multimodal integration, and established protocols as ways to develop trustworthy, morally sound neurological diagnostic systems.

Haposan et al. [28] introduced a hybrid deep learning architecture that incorporates CNNs and RNNs for emotion recognition from facial expressions in video on the Emotional Wearable Dataset 2020 and achieved the highest accuracy on InceptionV3-RNN: 66% (highest). They have identified that, although the achieved accuracy is moderate, the method makes a significant contribution to dynamic and context-aware emotion identification. The author has recommended using multimodal learning, bigger datasets, and effective structures for real-world implementation.

Dongliang Chen et al. [29] has proposed a novel framework for video-based emotion identification, called CDGT (Constructing Diverse Graph Transformers), on AffectNet-Vid and AFEW to efficiently manage temporal and spatial relationships in dynamic facial expressions. It incorporates graph structures into transformer models. The framework CDGT performs better than CNN-RNN-based techniques and state-of-the-art transformers. The author has suggested incorporating multimodal emotional cues, Lightweight graph-transformer designs, domain adaptability, cross-domain transfer learning, and nuanced affective states and micro-expressions to achieve a deeper understanding of emotions in future work.

The majority of researchers have undoubtedly resorted to deep learning because of its exceptional accuracy in categorization tasks. Deep learning is a type of machine learning that draws on how neural networks, which are part of the human brain, work. By simulating the brain's intricate functions, it seeks to create models and algorithms capable of learning and making predictions. Deep learning approaches include recurrent neural networks (RNNs), autoencoders, and artificial neural networks (ANNs). Convolutional Neural Networks (CNNs), also known as ConvNets, are widely used in FER image processing. This method's primary benefit is that it combines feature extraction and classification, significantly reducing the need for a manual feature extraction approach and its associated difficulties.

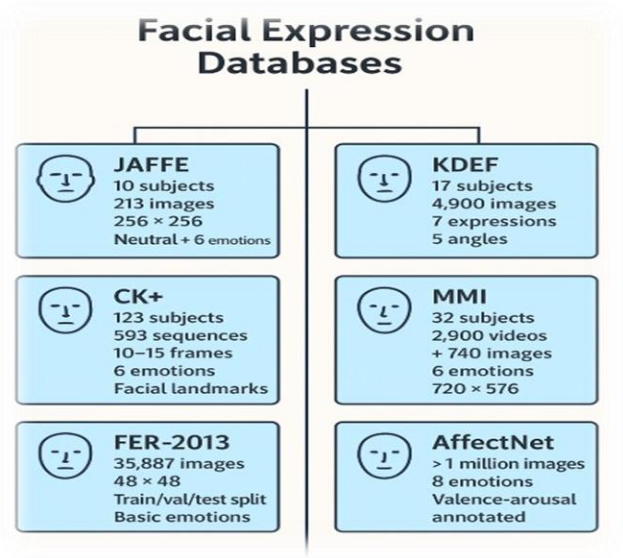

Figure 3. Facial Expression Database.

According to research, CNN [15] is a branch of neural networks that consists of two main building blocks: classification and feature extraction. Layers such as convolution, pooling, dropout, activation, and fully connected are all part of the CNN architecture. When CNN encounters overfitting, batch normalization and regularization are also used.

A feature map is the output of a convolution layer, the lowest layer, which uses an (MxM) filter to extract different information from input images. Feature maps offer details about an image's corners and edges. The main goal of the pooling layer is to reduce computational cost and feature map dimensionality. The fully connected layer is composed of neurons, weights, and bias, and is responsible for tying together several layers. It is usually positioned ahead of the CNN's output layer. The CNN architecture's overfitting issue is resolved by the dropout layer.

The activation function uses neurons to initiate connections between layers. A few popular activation functions are Sigmoid, TanH, ReLu, and Softmax. Three sets of the dataset, viz., training, validation, and testing as well as batch size and epochs, can be used to construct the CNN model [16]. A neural network usually requires a large amount of data to be trained to achieve high accuracy. When overfitting arises in the deep learning model, data augmentation techniques, including flipping, rotating, scaling, cropping, translating, and adding Gaussian noise, are employed to expand the dataset. To create the CNN architecture, many Python packages, such as TensorFlow, Keras, and PyTorch, are frequently used. Pretrained deep convolutional neural network (DCNN) models, like VGG-16, ResNet, DenseNet, and Inception, are used by FER through suitable transfer learning. These models are trained on large datasets (such as ImageNet) with many classes. The contributions of various researchers are summarized below, including the methods they employed and the datasets used in their experiments.

Table 1. Summary of Research on Emotion Detection

Authors	Methods	Data sets	Key Contribution
Bhavana N et al. [17] 2025	Transfer Learning (RESNET50, Inception V3, VGG19, MobileNet)	CK+	Psychological state detection based on emotions
Liping Lu et al. [19] 2025	CNN + LSTM	DEAP, AMIGOS, DEAP+AMIGOS	Psychological state detection based on emotions
Abdul Aziz et al. [22] 2025	Transformer framework	3% improvement in sentiment, 2.2% in emotion, 1% in desire	Integrated desire analysis with emotion recognition
Alisawai et al. [24] 2025	CBAM-4CNN (attention mechanism)	FER-2013, CK+	Dataset enhancement, framework innovation
Shiqing Zhang et al. [26] 2024	Transformer-based and hybrid fusion networks	IEMOCAP, CMU-MOSEI/MOSI, SEMAINE, MELD, AFEW	Feature extraction and information fusion review
Yoonesi et al. [27] 2024	CNNs (ResNet, VGGNet, MobileNet) for neurological diseases	CK+, FER-2013	Meta-analysis for neurological disorder detection

Authors	Methods	Data sets	Key Contribution
Cheng et al. [37] 2023	Transformer Autoencoder (TAE)	DEAP, SEED-IV	Handling missing modalities in MER
Haposan et al. [28] 204	Hybrid CNN-RNN (InceptionV3-RNN)	Emotional Wearable Dataset 2020	Dynamic context-aware emotion detection
Zhang et al. [33] 2024	Review: CNNs, RNNs, LSTMs, Attention, GNNs	DEAP, SEED, DREAMER, MAHNOB-HCI, AMIGOS	Micro-expression recognition framework

Dhavanil Bhagat et al. [30] developed a real-time facial emotion recognition using a Deep Convolutional Neural Network for emotion detection. The author uses pre-trained architectures such as VGGNet, ResNet, and EfficientNet in combination with a Haar face classifier to improve processing efficiency and detection accuracy on the FER-2013 dataset, achieving an accuracy of 82%. In the future, the author will increase resistance to environmental changes, broaden data set diversity, and ensure that deployment contexts adhere to ethical standards.

Khan et al. [31] addresses the significant drawback of contact-based devices that prevents the practical application of emotion detection technology by exploring Contactless Multimodal Emotion Detection (CMER). They have created a system that links specific modality combinations to MER needs, providing a methodology for evaluating various CMER use cases. This study provides both theoretical underpinnings and useful frameworks for putting CMER systems into practice in a variety of real-world scenarios. The review effectively converts contact-based, laboratory-constrained techniques for emotion identification into workable, deployable contactless solutions.

Mahboobeh Jafari et al. [32] explored the potential of deep learning for emotion detection from EEG signals. The author has applied deep learning architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) for emotion detection. The review examines how feature extraction methods, frequency band choices, and signal preprocessing strategies affect model performance on the DEAP, SEED, and DREAMER datasets, achieving accuracies of more than 85%. The author has stated that further work is needed to improve model interpretability, cross-subject generalization, and real-time deployment capabilities.

Zhang et al. [33] Investigate deep learning-based micro-expression recognition (MER) methods with a focus on modularity and streaming data. This work presented a thorough overview of current issues, experimental procedures, assessment criteria, and application domains, as well as a synopsis and analysis of every component of the framework. The data set verified by the researchers was DEAP, SEED, DREAMER, MAHNOB-HCI, and AMIGOS for their studies. The author has reviewed the architectures of deep learning, such as CNNs, RNNs, LSTMs, attention mechanisms, and graph neural networks, through their studies. This work has suggested incorporating multimodal fusion, transfer learning, and developing AI/DL methods to handle the complexity of the EEG signal for emotion detection.

Kumar et al. [34] has provided a comprehensive evaluation of the emotion detection system using machine learning and deep learning on multimodal data. The study focuses on the advantages of hybrid categorization methods and on how the technology can revolutionize several industries by enhancing emotional intelligence and automating decision-making. To improve emotion identification systems for a variety of real-world applications across customer-based industries, such as monitoring patients in the public, private, and healthcare sectors, as well as production companies. The author has addressed various challenges and issues in deep learning. In the future, we anticipate extensive testing of advanced machine learning algorithms based on face, speech, and textual data as well as multi-layer systems with high recognition accuracy rates for each area of emotion recognition.

K.Ezzameli et al. [35] employed a multimodal approach for emotion detection using deep learning, achieving higher accuracy. The modalities include images, text, video, speech, and psychological signals, and deep learning techniques such as CNNs, RNNs, LSTMs, Transformers, and Autoencoders Are Used for automatic feature extraction. The unimodal approach includes LBP, HOG, MFCC, and EEG spectral features combined with classifiers such as SVM, KNN, and Random Forest. The author has suggested comprehending affect in talks, as one person's feelings might influence those of other participants. The multimodal technology may simulate interpersonal emotional dependence, leading to notable progress in multimodal affect research. Also, they have stated that effective, comprehensible, and culturally sensitive systems are still needed for real-world deployment in healthcare, education, and robotics.

Pan et al. [36] has presented comprehensive analyses of datasets, pre-processing approaches, feature extraction techniques, and fusion methodologies, offering recommendations to HCI and affective computing researchers. The study finds that the most promising method for obtaining precise, reliable, and adaptable emotion recognition is deep learning-driven multimodal fusion. They have stated that SVM, Random Forest, Deep Belief Networks (DBN), LSTM, and Transformer-based architectures, with deep models, have shown clear dominance in the domain. The authors provide an extensive overview of benchmark datasets frequently

utilized in MER research. The author has recommended that their future work will include effective fusion architectures, Self-supervised learning, transfer learning, and transparent and explicable emotion recognition.

Cheng et al. [37] A unique transformer-based architecture has been developed to address the crucial problem of missing modalities in multi-modal emotion recognition (MER) on the DEAP and SEED-IV Datasets. This research contributes to the theoretical knowledge of cross-modal learning while offering a workable, deployable solution for reliable emotion recognition in demanding real-world settings. The suggested Transformer Autoencoder (TAE) exhibits amazing robustness by achieving excellent accuracy rates, surpassing 96% on complete data and retaining over 93% accuracy even with missing modalities. The author has suggested developing explainability, cross-dataset generalization, and dynamic modality selection for their future work.

Yante Li et al. [38] presented a review of deep learning-based micro-expression recognition (MER), they have also studied a thorough examination of deep learning-based micro-expression identification, creating a comprehensive taxonomy that covers input modalities, network designs, training methods, preprocessing, and datasets. This study provides researchers with a crucial point of reference, offering not just a historical summary but also a path forward for innovation in this difficult and significant field. In the future, the connection between AUs and MEs can be investigated further to enhance MER.

Ben et al. [39] contributed a methodical review of datasets, features, methods, and applications related to video-based micro-expression analysis. The study provides uniform assessments of cutting-edge techniques and proposes a new dataset (MMEW) that will serve as a starting point for further research. To support future research, especially when examining the relation between macro- and micro-expressions within the same individuals, this survey presents the MMEW dataset and offers the first thorough, cohesive comparison of micro-expression analysis techniques. Future directions are clearly identified in the paper: deeper investigation of macro-micro correlations, explainable AI for essential applications, larger standardized datasets, multi-task learning frameworks, privacy-preserving techniques, and GAN-based data augmentation.

Huilin Ge et al. [40], conducted a thorough analysis and analytical methodology for deep learning-based facial expression recognition (FER). Static facial expression recognition, which examines individual frames, and dynamic facial expression recognition, which captures temporal differences across video sequences, are the two main categories into which the authors divide FER techniques on the CK+, JAFFE, MMI, and in-the-wild datasets (FER2013, RAF-DB, AffectNet). The study shows that CNN-based and hybrid deep architectures (CNN-LSTM, GAN-based techniques) have attained high recognition accuracy, often over 95%, in controlled settings. Their focus on multimodal integration, occlusion restoration, and practical flexibility offers insightful direction for current and next affective computing research.

2.1 Advantages of Deep Networks for Recognizing Emotions in the Face

The advantages of deep neural networks for recognizing emotions on faces can be described as follows:

a) Improved Feature Representation: By learning hierarchical and discriminative face features efficiently, CNNs and Transformer-based architectures lessen the need for manually created descriptors.

b) Performance Enhancements with Advanced Architectures: By recording both local and global facial information, hybrid CNN-Transformer models and attention mechanisms increase recognition accuracy.

c) The ability to effectively analyze dynamic facial expressions in video sequences is made possible by recurrent and temporal models.

d) Adaptability Across Benchmark Datasets: A number of deep learning techniques have high performance on popular FER datasets, indicating the ability to learn under controlled circumstances.

2.2. Key Challenges and Limitations in State-of-the-Art FER (Shortcomings)

The key challenges and limitations in FER can be described as follows:

a) High Computational Complexity: Transformer-based and hybrid systems are not suitable for real-time or resource-constrained situations due to their high computational resource requirements.

b) Limited Cross-Dataset Generalization: When models are used on datasets other than those they were trained on, their performance frequently deteriorates

c) Sensitivity to Real-World Variations: In unconstrained environments, recognition robustness is greatly impacted by pose, occlusion, lighting, and backdrop variations.

d) Data Imbalance and Annotation Issues: Model reliability and fairness are diminished by uneven emotion class distribution and inconsistent labeling across datasets.

According to this study, research on emotion recognition has progressed using CNNs, RNNs, LSTMs, Transformers, and hybrid models. Nonetheless, notable deficiencies remain in multimodal integration, explainability, resource-efficient and real-time deployment, cross-dataset generalization, and resilience to individual, cultural, and environmental variability. Future studies must focus on multimodal, explainable,

adaptive, and morally sound emotion identification frameworks that operate in cross-cultural, context-aware, and real-world settings.

3. METHODS

Various methods have been employed to improve emotion detection accuracy and reduce computational cost. These methods have demonstrated the potential of emotion detection to serve society by providing efficient models in the domain. The different architectures of deep learning have distinct strengths in handling issues of varying complexity.

A comprehensive literature search was carried out across several databases, including Scopus, IEEE Xplore, and Web of Science. Duplicate entries were eliminated when the retrieved records were loaded into a reference manager. The complete text was evaluated for eligibility after titles and abstracts were checked in accordance with predetermined inclusion and exclusion criteria. A PRISMA flow diagram was used to record each step of the research selection process, including identification, screening, eligibility, and inclusion. As a result, 555 records were identified, 527 articles were assessed, and 55 studies were included in the final synthesis.

Various deep learning architectures are applied to emotion recognition for improved results and efficient computation. Table 2 shows the evolution of deep learning architectures, leveraging their computing power and capabilities, including:

Table 2. Performance Characteristics of Architectures

Architectures	Introduces	Key Strengths	Applications
CNN+LSTM Hybrid	2015	Combines spatial feature extraction and temporal sequence modeling	Multimodal emotion recognition, real-time emotion detection
Inception(Google Net)	2016	Captures multi-scale features; efficient computation	Facial emotion recognition in complex datasets
ResNet (Residual Network)	2015	Enables training of very deep networks; mitigates the vanishing gradient problem	Facial expression recognition, video-based emotion recognition
DenseNet	2017	Enhances feature reuse; efficient parameter usage	Facial emotion recognition, small datasets with high accuracy
EfficientNet	2019	Balances network depth, width, and resolution for optimal performance	Efficient facial emotion recognition in constrained environments
Transformer-Based Models	2017	Captures long-range dependencies; effective for sequential data	Emotion recognition from speech, text, and multimodal data
Graph Neural Networks (GNNs)	2018	Models relationships in graph structures; effective for structured data	Emotion detection in social networks, multimodal interactions, and physiological data
EEG-based Deep Learning Models	2022-2022	Analyzes brain wave patterns; effective for understanding emotional states	Emotion recognition from EEG signals
Ensemble Deep Learning Frameworks	2025	Combines multiple models for improved accuracy and robustness	Emotion recognition through wearable devices and multimodal physiological signal
Attention-Based Transformer Models	2024-2025	Focuses on important features; effective for sequential and multimodal data	Emotion detection from handwriting, drawing samples, and multimodal data

.

CNN+LSTM Hybrid (2015) uses temporal sequence modeling and spatial feature extraction to recognize emotions in real time and across multiple modalities. ResNet (2015), which excels at face expression and video-based emotion recognition, enables training very deep networks by addressing the vanishing gradient problem.

Multi-scale characteristics are efficiently captured by Inception/GoogleNet (2016), making it perfect for facial emotion identification in intricate datasets. With effective parameters, DenseNet (2017) improves feature reuse and achieves good accuracy on small facial emotion datasets. Transformer-Based Models (2017) recognize emotions from speech, text, and multimodal sources by capturing long-range dependencies in sequential data. In 2018, Graph Neural Networks models were useful for identifying emotions in physiological data and social networks. For effective facial emotion identification in resource-constrained settings, EfficientNet (2019) strikes the ideal balance between network depth, width, and resolution. Brain wave patterns are analyzed by EEG-based Deep Learning Models (2020–2022) to identify emotions straight from EEG signals. Several models are combined in Ensemble Deep Learning Frameworks (2023) to provide reliable emotion recognition from physiological inputs and wearable technology. By focusing on key characteristics, Attention-Based Transformer Models (2024–2025) employ attention mechanisms to detect emotions in handwriting, drawings, and multimodal data.

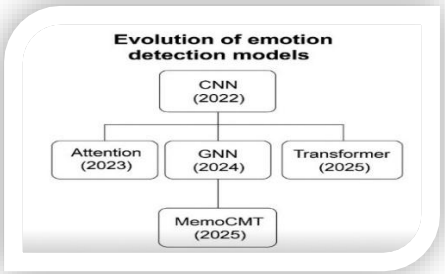

Figure 4. Emotion Detection Model

The performance of the different architectures' models is stated as:

Table 3. Performance Characteristics of Architectures

Architecture	Depth Handling	Computational Efficiency	Model Size	Memory Usage	Best For
AlexNet	Moderate (8 layers)	Moderate	Medium (60M params)	Moderate	Feature extraction basics, historical baseline
VGG	Deep layers (16-19)	Lower	Large (138M params for VGG16)	High	Straightforward implementations, feature visualization
ResNet	Very deep (50-152 layers)	Moderate	Medium-Large (25-60M params)	Moderate-High	High accuracy requirements, deep learning research
Inception	Multi-scale processing	High	Medium (23M params for V3)	Moderate	Multi-scale feature needs, efficient networks
MobileNet	Optimized depth	Very High	Small (3-4M params)	Low	Resource-constrained devices, mobile applications
Xception	Depth-separable	High	Medium (22M params)	Moderate	Accuracy + efficiency balance, mobile deployment
DenseNet	Dense connections	Moderate	Compact (8-14M params)	Moderate	Feature reuse optimization, parameter efficiency
EfficientNet	Balanced scaling	Very High	Small-Medium (5-66M params)	Low-Moderate	Scalable performance needs, state-of-the-art accuracy

AlexNet, one of the main deep learning architectures, is well-suited for simple feature extraction because it is a medium-sized model with moderate depth and efficiency. VGG is best suited for simple implementations because it has deep layers, but it is less efficient because of its size. ResNet performs best for high-accuracy needs and can handle very deep networks (50–152 layers) with modest efficiency. For multi-scale feature requirements, Inception offers highly efficient multi-scale processing in a medium-sized model. MobileNet offers maximum depth and excellent efficiency in a small package. To balance accuracy and computational cost, Xception uses depthwise-separable convolutions for medium-sized models, achieving high efficiency. DenseNet achieves moderate efficiency in its compact form by improving feature reuse through dense connections. EfficientNet provides small-to-medium models with balanced scaling and extremely high efficiency, making it ideal for scalable performance across a range of computing demands. Deep learning models have strong potential to address emotion detection challenges efficiently. The table highlights innovations in emotion detection that improve the classification accuracy of these models.

Table 4. Innovations in Architectures

Architecture	Primary Problem Solved	Innovation
AlexNet	Image classification at scale	Demonstrated DL effectiveness
VGG	Network design simplicity	Uniform architecture design
ResNet	Vanishing gradient in deep networks	Residual connections
Inception	Multi-scale feature extraction	Inception modules
MobileNet	Resource constraints	Depthwise separable convolutions
Xception	Computational economy	Enhanced separable convolutions
DenseNet	Feature reuse efficiency	Dense connectivity pattern

Architecture	Primary Problem Solved	Innovation
EfficientNet	Model scaling efficiency	Compound scaling method

To analyze the development of deep learning architectures for emotion detection across several modalities, this paper adopts a model-centric, historical approach. The chosen works were examined methodically, classified by architectural innovations, publication time, and application domain.

The review begins with early hybrid designs, such as CNN-LSTM models (2015), that combine temporal sequence modeling and spatial feature extraction for video-based, real-time emotion recognition. The capacity of basic deep CNNs, such as ResNet (2015), Inception/GoogleNet (2016), and DenseNet (2017), to handle issues such as vanishing gradients, multi-scale feature learning, and feature reuse in facial emotion identification tasks was then evaluated. Graph Neural Networks (2018) are used for relational emotion analysis in physiological and social network data, while Transformer-based models (2017 onwards) are used to model long-range dependencies in sequential and multimodal emotion data. In order to assess efficiency, robustness, and adaptability in resource-constrained, wearable, and multimodal emotion detection scenarios, more recent architectures were examined, such as EfficientNet (2019), EEG-based deep learning models (2020–2022), ensemble frameworks (2023), and attention-based transformer models (2024–2025).

Table 3 summarizes the methodical comparison of performance attributes for each architecture, including network depth management, computational efficiency, model size, memory utilization, and application applicability. This methodical assessment enables a detailed comparison of architectural trade-offs and identifies key advancements that enhance the scalability and accuracy of emotion recognition.

4. RESULTS AND DISCUSSION

Compared with conventional methods, deep learning models' findings show notable gains in accuracy and efficiency. When tested on a variety of benchmark datasets, models such as CNNs and their variants, including MobileNet, Xception, DenseNet, and EfficientNet, have shown excellent performance in feature extraction and classification tasks. Xception and EfficientNet delivered improved accuracy with balanced computational efficiency, whereas MobileNet's lightweight architecture enabled it to perform well under resource constraints. DenseNet enhanced learning with fewer parameters by efficiently reusing features. All things considered, the results show that performance is significantly impacted by the architecture selection, with trade-offs between prediction accuracy, model size, and computational cost. These findings emphasize the importance of selecting the right deep learning models for the specific needs of the applications.

Table 5. Results of Emotion Detection

Author	Data Sets	Results
Bhavana N et al. [17] **2025**	CK+	97%
Liping Lu [19] **2025**	DEAP, AMIGOS, DEAP+AMIGOS	DEAP: V-87.6%, A-89.3%; AMIGOS: V-88.5%, A-90.1%; Combined: V-91.3%, A-92.5%
Abdul Aziz et al. [22] **2025**	MSED	3% improvement in sentiment, 2.2% in emotion, 1% in desire
Alisawai et al. [24] **2025**	FER-2013, CK+	81%
Geetha et al. [25] **2025**	IEMOCAP, RAVDESS, MSP-IMPROV, CMU-MOSEI, MELD, EmotiW, AMIGOS, DEAP, MAHNOB-HCI, AffectNet, FER2013, AFEW	10-25% improvement over single modality
Shiqing Zhang et al. [26] **2024**	IEMOCAP, CMU-MOSEI/MOSI, SEMAINE, MELD, AFEW	80-90%
Yoonesi et al. [27] **2024**	CK+, FER-2013	95%
Dhavanil Bhagat et al. [30] **2024**	FER-2013	82%
Mahboobeh Jafari et al. [31] **2024**	DEAP, SEED, DREAMER	>85%
Huilin Ge et al. [39] **2023**	CK+, JAFFE, MMI, FER2013, RAF-DB, AffectNet	>95% (controlled settings)

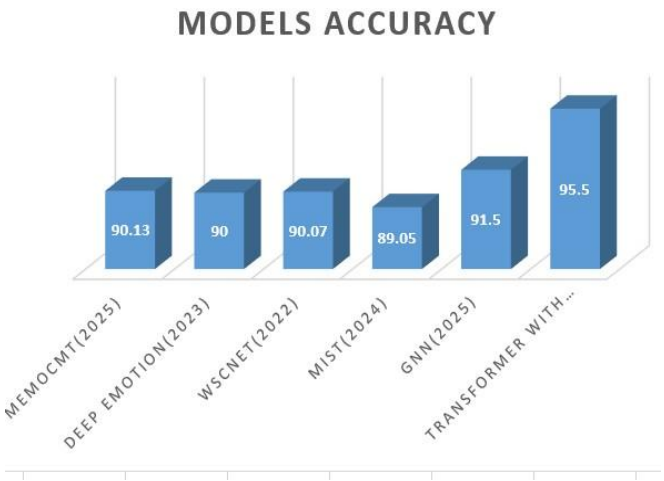

Figure 5. Model Accuracy

The bar graph shows consistent improvements in deep learning-driven emotion recognition from 2022 to 2025. Deeper CNN-based methods, such as Deep Emotion (2023), followed early models like WSCNet (2022) that improved feature attention. For dynamic emotion analysis, later models, such as MIST (2024), used spatiotemporal modeling. By 2025, hybrid models based on transformer, GNN, and memory (such as Transformer with Attention, GNN, and MemoCMT) show better contextual comprehension and accuracy, suggesting a definite trend toward transformer-driven and hybrid architectures for improved emotion recognition.

According to the reviewed studies, CNN-based architectures consistently outperform conventional methods across facial, physiological, and multimodal data, and deep learning models greatly enhance emotion identification accuracy while balancing computing economy. Performance, however, differs significantly between datasets and experimental configurations.

Reported accuracy is highly influenced by dataset features. While posed and controlled datasets like CK+, JAFFE, and FER2013 frequently show high performance, real-world difficulties, including illumination change, occlusion, and spontaneous expressions, cause in-the-wild datasets to produce poorer and less reliable results. This suggests that accuracy values from different research are not always directly comparable.

Deeper models like ResNet, Inception, and DenseNet improve robustness and feature representation, whereas CNN–LSTM hybrids efficiently capture spatial–temporal characteristics. Lightweight architectures, such as MobileNet and EfficientNet, are well-suited to real-time and resource-constrained applications because they deliver competitive performance at lower computational cost. Adaptability to multimodal and physiological emotion recognition is further enhanced by recent Transformer-, GNN-, ensemble-, and attention-based models. Overall, the findings highlight architectural trade-offs rather than perfect accuracy, showing that deployment limits, assessment settings, and dataset type all affect model applicability.

5. CONCLUSION AND FUTURE ENHANCEMENT

This research presents a comprehensive comparison of deep learning (DL) and machine learning (ML) approaches in facial expression recognition (FER) systems. The advanced deep learning architectures for face emotion detection and recognition were thoroughly examined in this study, with a focus on the increasing popularity of attention-based and Transformer-driven models. The study demonstrated the capacity of Transformer architectures to capture intricate spatial–temporal connections and enhance recognition performance by methodically analyzing hybrid CNN–Transformer frameworks, spatiotemporal learning techniques, and innovative attention mechanisms.

Furthermore, the study emphasizes the necessity of balanced model construction that incorporates robustness, efficiency, and accuracy. It also offers a structured viewpoint to direct future research toward scalable, broadly applicable, and useful emotion recognition systems. Simultaneously, the review identified enduring issues that hinder real-world implementation, including high computational complexity, susceptibility to unrestricted settings, limited cross-dataset generalization, and the impact of data imbalance on model reliability.

Through this critical analysis, the paper highlights the necessity of a balanced model design that incorporates robustness, accuracy, and efficiency. It also offers an organized perspective to guide future research toward scalable, generalizable, and useful emotion detection systems. To evaluate the effectiveness of FER systems across various datasets and address challenges encountered in real-world scenarios, we

developed a systematic review framework and formulated research questions. Traditional FER methods involve face detection, pre-processing, feature extraction, and emotion classification using ML classifiers. In contrast, deep learning algorithms demonstrate robustness when handling complex facial images, including extreme expressions, pose variations, and illumination changes. However, these techniques require substantial computational resources, such as GPUs and TPUs, due to their longer training and testing times. This study also indicates that existing FER systems achieve promising results using ensemble DL methods, GANs, LSTMs, and autoencoders. Additionally, FER datasets are analyzed in two categories: (1) posed and (2) spontaneous expressions. Most current FER systems are trained on posed datasets due to the challenges associated with spontaneous expressions. The available spontaneous datasets are highly complex, containing variations in poses, illumination, and expression intensity.

This study focuses on FER techniques and their accuracy across multiple datasets, highlighting the ongoing need to develop highly efficient FER systems with reduced computation time. Furthermore, this research summarizes the contributions of various researchers in the domain of emotion detection using deep learning. Large and deep models, such as ResNet and Transformer-based architectures, often require substantial processing power, which limits the applicability of their advertised performance and makes them less useful for real-time or resource-constrained applications. Furthermore, the majority of research uses lab-controlled or posed datasets (e.g., CK+, JAFFE), which limits the applicability of findings to real-world situations and compromises the validity of cross-study comparisons. The review highlights several open issues and potential directions for future research, including:

- Developing explainable and interpretable FER models to improve transparency and trust in real-world applications.
- Addressing cross-cultural differences in emotional expression to enhance model generalization across diverse populations.
- Designing privacy-preserving FER techniques to ensure ethical deployment in sensitive environments.
- Improving robustness under uncontrolled conditions such as occlusion, pose variation, and illumination changes.
- Creating lightweight and computationally efficient architectures suitable for real-time applications on resource-constrained devices.

REFERENCES

[1] K. Scherer and P. Ekman, "Handbook of Methods in Nonverbal Behavior Research," *Cambridge University Press*, pp. 45–90, 1982.

[2] P. Ekman and W. V. Friesen, "Constants Across Cultures in the Face and Emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971. https://doi.org/10.1037/h0030377

[3] A. Seyeditabari, N. Tabari, and W. Zadrozny, "Emotion Detection in Text: A Review," 2018. https://doi.org/10.48550/arxiv.1806.00674

[4] R. A. Calvo and S. M. Kim, "Emotions in Text: Dimensional and Categorical Models," *Computational Intelligence*, vol. 29, no. 3, pp. 527–543, 2012. https://doi.org/10.1111/j.1467-8640.2012.00456.x

[5] P. Ekman, "An Argument for Basic Emotions," *Cognition & Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992. https://doi.org/10.1080/02699939208411068

[6] P. R. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion Knowledge: Further Exploration of a Prototype Approach," *Journal of Personality and Social Psychology*, vol. 52, no. 6, pp. 1061–1086, 1987. https://doi.org/10.1037/0022-3514.52.6.1061

[7] K. Oatley and P. N. Johnson-Laird, "Towards a Cognitive Theory of Emotions," *Cognition & Emotion*, vol. 1, no. 1, pp. 29–50, 1987. https://doi.org/10.1080/02699938708408362

[8] R. Plutchik, "A Psychoevolutionary Theory of Emotions," *Social Science Information*, vol. 21, no. 4–5, pp. 529–553, 1982. https://doi.org/10.1177/053901882021004003

[9] H. Lövheim, "A New Three-Dimensional Model for Emotions and Monoamine Neurotransmitters," M*edical Hypotheses*, vol. 78, no. 2, pp. 341–348, 2012. https://doi.org/10.1016/j.mehy.2011.11.016

[10] P. Ekman, "Basic Emotions," *Handbook of Cognition and Emotion*, pp. 45–60, 1999. https://doi.org/10.1002/0470013494.ch3

[11] L. Canales and P. Martínez-Barco, "Emotion Detection from Text: A Survey," *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pp. 37–43, 2014. https://doi.org/10.3115/v1/w14-6905

[12] A. Mehrabian, "Silent Messages", *Wadsworth Belmont,* vol. 8, no. 152, 1971.

[13] H. C. Triandis and M. Fishbein, "Cognitive Interaction in Person Perception," Journal of Abnormal & Social Psychology, vol. 67, no. 5, pp. 446–453, 1963, https://doi.org/10.1037/h0038494

[14] P. Ekman, "Methods for Measuring Facial Action," *Handbook of methods in nonverbal Behavior Research*, pp. 45–135, 1982.

[15] S. Hossain, S. Umer, R. K. Rout, and M. Tanveer, "Fine-Grained Image Analysis for Facial Expression Recognition using Deep Convolutional Neural Networks with Bilinear Pooling," *Applied Soft Computing*, vol. 134, pp. 109997, 2023. https://doi.org/10.1016/j.asoc.2023.109997

[16] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended Deep Neural Network for Facial Emotion Recognition," *Pattern Recognition Letters*, vol. 120, pp. 69–74, 2019. https://doi.org/10.1016/j.patrec.2019.01.008

[17] F. D. Luzio, A. Rosato, and M. Panella, "An Explainable Fast Deep Neural Network for Emotion Recognition," *Biomedical Signal Processing and Control*, vol. 100, pp. 107177, 2025. https://doi.org/10.1016/j.bspc.2024.107177

[18] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep Learning-based Facial Emotion Recognition for Human–Computer Interaction Applications," *Neural Computing and Applications*, vol. 35, no. 32, pp. 23311–23328, 2021. https://doi.org/10.1007/s00521-021-06012-8

[19] L. Lu, L. Yuan, and L. Chen, "Deep Learning Based Emotion Recognition for Analyzing Students' Psychological States during Competitions," *Entertainment Computing*, vol. 55, pp. 101005, 2025. https://doi.org/10.1016/j.entcom.2025.101005

[20] K. Devarajan, P. Suresh, and S. Perumal, "Enhancing Emotion Recognition through Multi-Modal Data Fusion and Graph Neural Networks," *Intelligence-Based Medicine*, vol. 12, pp. 100291, 2025. https://doi.org/10.1016/j.ibmed.2025.100291

[21] E. Boitel, A. Mohasseb, and E. Haig, "MIST: Multimodal Emotion Recognition using DeBERTa for Text, Semi-CNN for Speech, ResNet-50 for Facial, and 3D-CNN for Motion Analysis," *Expert Systems with Applications*, vol. 270, pp. 126236, 2025. https://doi.org/10.1016/j.eswa.2024.126236

[22] N. K. Chowdhury, M. A. Kabir, A. N. Chy, and Md. J. Siddique, "MMTF-DES: A Fusion of Multimodal Transformer Models for Desire, Emotion, and Sentiment Analysis of Social Media Data," *Neurocomputing*, vol. 623, pp. 129376, 2025. https://doi.org/10.1016/j.neucom.2025.129376

[23] S. Woo, M. Zubair, S. Lim, and D. Kim, "Deep Multimodal Emotion Recognition using Modality-Aware Attention and Proxy-based Multimodal Loss," *Internet of Things*, vol. 31, pp. 101562, 2025. https://doi.org/10.1016/j.iot.2025.101562

[24] N. Yalçın and M. Alisawi, "Introducing a Novel Dataset for Facial Emotion Recognition and Demonstrating Significant Enhancements in Deep Learning Performance through Pre-Processing Techniques," *Heliyon*, vol. 10, no. 20, pp. e38913, 2024. https://doi.org/10.1016/j.heliyon.2024.e38913

[25] G. Vijayaraghavan, T. Mala, D. P, and E. Uma, "Multimodal Emotion Recognition with Deep Learning: Advancements, Challenges, and Future Directions," *Information Fusion*, vol. 105, pp. 102218, 2024. https://doi.org/10.1016/j.inffus.2023.102218

[26] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep Learning-based Multimodal Emotion Recognition from Audio, Visual, and Text Modalities: A Systematic Review of Recent Advancements and Future Prospects," *Expert Systems with Applications*, vol. 237, pp. 121692, 2024. https://doi.org/10.1016/j.eswa.2023.121692

[27] S. Yoonesi, R. Azar, M. Bafrani, S. Yaghmayee, H. Shahavand et al., "Facial Expression Deep Learning Algorithms in the Detection of Neurological Disorders: A Systematic Review and Meta-Analysis," *Biomedical Engineering Online*, vol. 24, no. 1, 2025. https://doi.org/10.1186/s12938-025-01396-3

[28] H. V. Manalu and A. P. Rifai, "Detection of Human Emotions Through Facial Expressions using Hybrid Convolutional Neural Network-Recurrent Neural Network Algorithm," *Intelligent Systems with Applications*, vol. 21, pp. 200339, 2024. https://doi.org/10.1016/j.iswa.2024.200339

[29] D. Chen, G. Wen, H. Li, P. Yang, C. Chen, and B. Wang, "CDGT: Constructing Diverse Graph Transformers for Emotion Recognition from Facial Videos," *Neural Networks*, vol. 179, pp. 106573, 2024. https://doi.org/10.1016/j.neunet.2024.106573

[30] D. Bhagat, A. Vakil, R. K. Gupta, and A. Kumar, "Facial Emotion Recognition (FER) using Convolutional Neural Network (CNN)," *Procedia Computer Science*, vol. 235, pp. 2079–2089, 2024. https://doi.org/10.1016/j.procs.2024.04.197

[31] U. A. Khan, Q. Xu, Y. Liu, A. Lagstedt, A. Alamäki, and J. Kauttonen, "Exploring Contactless Techniques in Multimodal Emotion Recognition: Insights into Diverse Applications, Challenges, Solutions, and Prospects," *Multimedia Systems*, vol. 30, no. 3, 2024. https://doi.org/10.1007/s00530-024-01302-2

[32] M. Jafari, A. Shoeibi, M. Khodatars, S. Bagherzadeh et al., "Emotion Recognition in EEG Signals using Deep Learning Methods: A Review," *Computers in Biology and Medicine*, vol. 165, pp. 107450, 2023. https://doi.org/10.1016/j.compbiomed.2023.107450

[33] F. Zhang and L. Chai, "A Review of Research on Micro-Expression Recognition Algorithms Based on Deep Learning," *Neural Computing and Applications*, vol. 36, no. 29, pp. 17787–17828, 2024. https://doi.org/10.1007/s00521-024-10262-7

[34] H. Kumar and A. Martín, "Artificial Emotional Intelligence: Conventional and Deep Learning Approach," *Expert Systems with Applications*, vol. 212, pp. 118651, 2023. https://doi.org/10.1016/j.eswa.2022.118651

[35] K. Ezzameli and H. Mahersia, "Emotion Recognition from Unimodal to Multimodal Analysis: A Review," *Information Fusio*n, vol. 99, pp. 101847, 2023. https://doi.org/10.1016/j.inffus.2023.101847

[36] B. Pan, K. Hirota, Z. Jia, and Y. Dai, "A Review of Multimodal Emotion Recognition from Datasets, Preprocessing, Features, and Fusion Methods," *Neurocomputing*, vol. 561, pp. 126866, 2023. https://doi.org/10.1016/j.neucom.2023.126866

[37] C. Cheng, W. Liu, Z. Fan, L. Feng, and Z. Jia, "A Novel Transformer Autoencoder for Multi-Modal Emotion Recognition with Incomplete Data," *Neural Networks*, vol. 172, pp. 106111, 2024. https://doi.org/10.1016/j.neunet.2024.106111

[38] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep Learning for Micro-Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2028–2046, 2022. https://doi.org/10.1109/taffc.2022.3205170

[39] X. Ben et al., "Video-based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. https://doi.org/10.1109/tpami.2021.3067464

[40] H. Ge, Z. Zhu, Y. Dai, B. Wang, and X. Wu, "Facial Expression Recognition Based on Deep Learning," *Computer Methods and Programs in Biomedicine*, vol. 215, pp. 106621, 2022. https://doi.org/10.1016/j.cmpb.2022.106621

BIOGRAPHIES OF AUTHORS

Krishna Kant is an Assistant Professor in the Department of Computer Science, CHARUSAT University, Anand, Gujarat, India. He has received the PhD Degree from the Computer Science from Sardar Patel University, V.V. Nagar, Anand, Gujarat, India in 2025. He completed his MCA from Pune University in 2009. His main areas of research include Computer Vision, Affective Computing and Emotion Recognition. He can be contacted via email: kantkrisna81@gmail.com