IJIE

INTERNATIONAL JOURNAL OF INTEGRATED ENGINEERING
ISSN: 2229-838X e-ISSN: 2600-7916

Vol. 18 No. 4 (2026) 22-39
https://publisher.uthm.edu.my/ojs/index.php/ijie

Facial Emotion Recognition: Methods, Applications and Future Challenges

Anbananthan Pillai Munanday[1], Norazlianie Sazali[1,2,3]*, Arigela Satya Veerendra[4], Kumaran Kadirgama[5]

[1] *Faculty of Mechanical and Manufacturing Engineering, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, MALAYSIA*
[2] *Centre of Excellence for Advanced Research in Fluid Flow (CARIFF), Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuhraya Tun Razak, Gambang, Kuantan, Pahang, MALAYSIA*
[3] *Mechanical Engineering Department, Sepuluh Nopember Institute of Technology Kampus Keputih Sukolilo, Surabaya 60111, INDONESIA*
[4] *Department of Electrical and Electronics Engineering, Manicipal Institute of Technology, Manicipal Academy of Higher Education, Manipal, Karnataka 576104, INDIA*
[5] *Faculty of Mechanical & Automotive Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, MALAYSIA*

*Corresponding Author: azlianie@uthm.edu.my
DOI: https://doi.org/10.30880/ijie.2026.18.04.002

Article Info

Received: 9 November 2025
Accepted: 23 April 2026
Available online: 17 June 2026

Keywords

Facial Emotion Recognition, deep learning, affective computing, human-computer interaction, multimodal emotion analysis

Abstract

Facial Emotion Recognition (FER) has become a significant problem to solve at the border of computer vision, psychology, and artificial intelligence. As a result, FER technologies that are used in healthcare, automotive safety, human-computer interaction, and security, among others, progressively improve. This review article synthesizes in, depth the recent changes in image, based, video, based, and multimodal approaches to FER. As opposed to experimental research, this work does not provide new results directly but rather thoroughly reviews existing methods, datasets, and model architectures. The paper compares novel methods such as convolutional neural networks (CNN), vision transformers, and temporal modelling frameworks with commonly used datasets like FER2013, Real, World Affective Faces Database (RAF-DB), AffectNet, and Aff, Wild2. The survey extensively covers the techniques advantages, disadvantages, and comparative trends by looking rigorously at robustness, generalization, and deployment feasibility. The authors recognize the problems with dataset imbalance, cultural variability, occlusion, and ethical issues and point to the important gaps in research. Moreover, the article predicts further advances in self, supervised learning, lightweight FER for edge devices, explainable models, and multimodal fusion strategies. Such a review is expected to help as a reference point for researchers and practitioners consolidating their understanding of the progress, challenges, and future opportunities in facial emotion recognition.

1. Introduction

Emotions are integral to human cognition, communication, and behaviour. Hence, the capability of machines to automatically recognize and interpret emotions has become a key challenge that lies at the intersection of computer vision, affective computing, and human-computer interaction. FER is aimed at detecting the emotional states of the humans by analysing either static images or video sequences of their faces, thereby, enabling machines to be more natural and empathetic in their interactions [1]. The architecture of FER systems depends heavily on psychological theories of emotion, especially single category models like Ekmans basic emotions and dimensional representations such as valence arousal frameworks, which not only help in annotation by indicating the emotional states but also play a role in computational systems by suggesting how these states are to be interpreted and evaluated. In the past ten years, the problem of FER has been attracting more researchers due to a vast number of potential applications, such as healthcare diagnostics, driver fatigue monitoring, education, security, entertainment, and human-robot collaboration [2, 3].

Fig. 1 illustrates the methods of early FER systems that used manually crafted features like local binary patterns (LBP) or Gabor filters combined with conventional classifiers. These techniques, although successful to some extent in controlled environments, were very sensitive to changes in lighting, pose, and occlusion [4]. The breakthrough in deep learning especially CNNs has changed the face of FER by making it possible to perform feature extraction and classification in an end, to, end fashion directly from large, scale datasets. Currently, transformer, based architectures, temporal attention mechanisms, and hybrid multimodal approaches are attaining significant improvements in accuracy and robustness by being able to capture both global spatial dependencies and contextual information across modalities [5]. Meanwhile, video, based FER has also been heavily influenced by the developments in recurrent neural networks, 3D CNN, and temporal self, attention frameworks that can represent the time, varying nature of facial expressions [6].

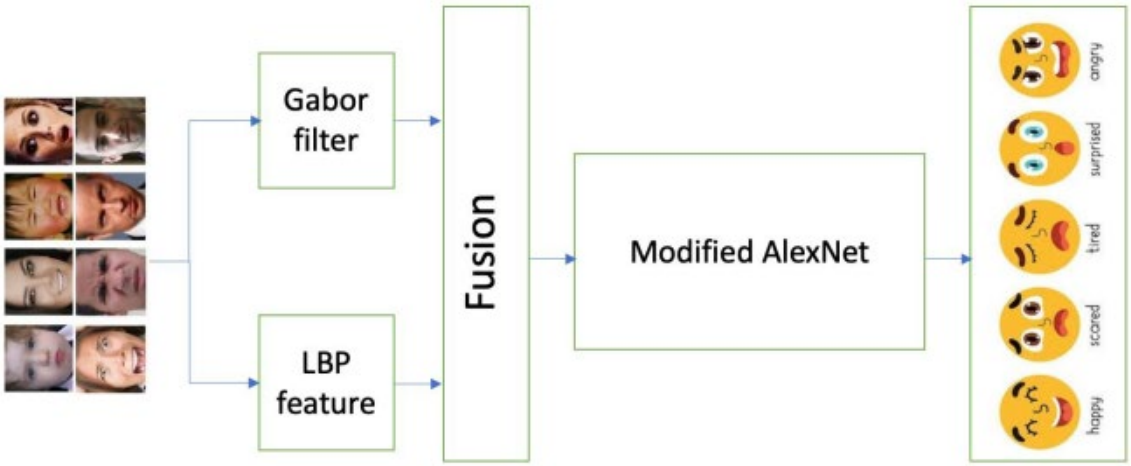

Fig. 1 *Traditional feature extraction pipeline: Utilization of Gabor filters and Local Binary Patterns (LBP) as inputs for an AlexNet-based classification model* [6]

Although these advances have been made, several unresolved challenges still hinder the deployment of systems in the real world. FER systems are often affected by dataset imbalance, limited cultural diversity, and annotation subjectivity, which lowers their generalization ability to different populations. Besides, conditions that are encountered in the real world such as wearing face masks, extreme head poses, or low, resolution surveillance videos can still make the performance of these systems worse [7]. Apart from technical obstacles, there are increasing concerns about privacy, algorithmic fairness, and the possible misuse of emotion inference technologies that have led to an intensification of ethical issues being discussed in relation to FER systems. As FER is progressively moving towards multimodal emotion recognition by integrating facial cues with speech, physiological signals, or contextual information, it is still very important to secure accuracy, transparency, and trustworthiness for the deployment in a responsible manner in the real world [8].

Surveys have recently been conducted and have discussed various aspects of FER and multimodal emotion recognition. Nevertheless, most of the review articles either concentrate on unimodal FER and do not consider the latest developments such as vision transformers, attention, based fusion, and explainable artificial intelligence (AI). Consequently, there exists a need for a comprehensive and up, to, date review that merges the methodological changes with the insights obtained from the applications and considers the challenges and ethical considerations [9].

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 22-39

The article is a review and does not bring any new experimental results. It basically goes through, compares, and integrates the results of different research works on FER. This review offers a comparative framework of traditional handcrafted methods, deep learning models, and recently proposed transformer, based architectures along with their strengths and weaknesses. Moreover, it scrutinizes the most employed FER datasets and benchmarks, shedding light on their features, biases, and how well they can be used for the evaluation in the real world. Furthermore, the paper identifies the main application sectors of FER and, in addition, pinpoints the challenges that have been the major obstacles to the deployment of FER in the wild issues such as robustness, generalization, and ethical considerations. Lastly, the paper points to the upcoming research avenues, highlighting self, supervised learning, domain adaptation, lightweight and interpretable models, and adaptive multimodal fusion strategies as the key factors to the development of the future FER systems.

2. Related Studies

FER has changed drastically over the last few years studies, moving away typical CNN, based methods to more advanced transformer architectures and multimodal integration. Most of the recent papers in this area point to three major directions, such as transform-based FER models, robustness improvements under occlusion, bias, and real, world variations, and multimodal fusion approaches that integrate audio, temporal, or physiological data [10]. This part is a survey of recent contributions and a summary of them in a comparative format.

2.1 Transformer and Hybrid Architectures

One of the major distinctions that Vision Transformers (ViTs) have is their incorporation of self, attention mechanisms. Consequently, ViT, based architectures can locate local as well as global dependencies in facial features, thus, they can handle pose variation, occlusion, and class imbalance more effectively [10]. To decide the best compromise between performance and computational power, Data-efficient image Transformers (DeiT), Swin Transformers with hierarchical windowed attention, and hybrid CNN-ViT architectures have been extensively used in FER [12]. These models beat the CNN baselines, as they increase the accuracy and F1, scores by about 3-7% thus, they have become the new benchmark methods when tested on standard datasets like RAF-DB, FERPlus, and AffectNet in noisy and occluded scenarios [12, 13].

Most recent ViT-based FER variants very clearly demonstrate that the incorporation of additional auxiliary information for example, action unit (AU) annotations or pseudo, AU supervision helps to remove dataset bias and improve generalization, thus resulting in accuracy increases of 2-4% on imbalanced datasets such as AffectNet [14]. Dual, stream architectures, which combine local salient feature extraction with global self, attention as shown in Fig. 2, are always better than single, stream transformers. These networks not only retain the minute facial features but also keep the holistic context, thus they are able to reach the reported accuracies of more than 90% on controlled benchmarks [15]. Comprehensive studies indicate that while performance of transformer, based models can be better under noisy or occluded conditions, they are usually accompanied by higher computational costs and longer inference times, which are still the main challenges for real, time application [16, 17].

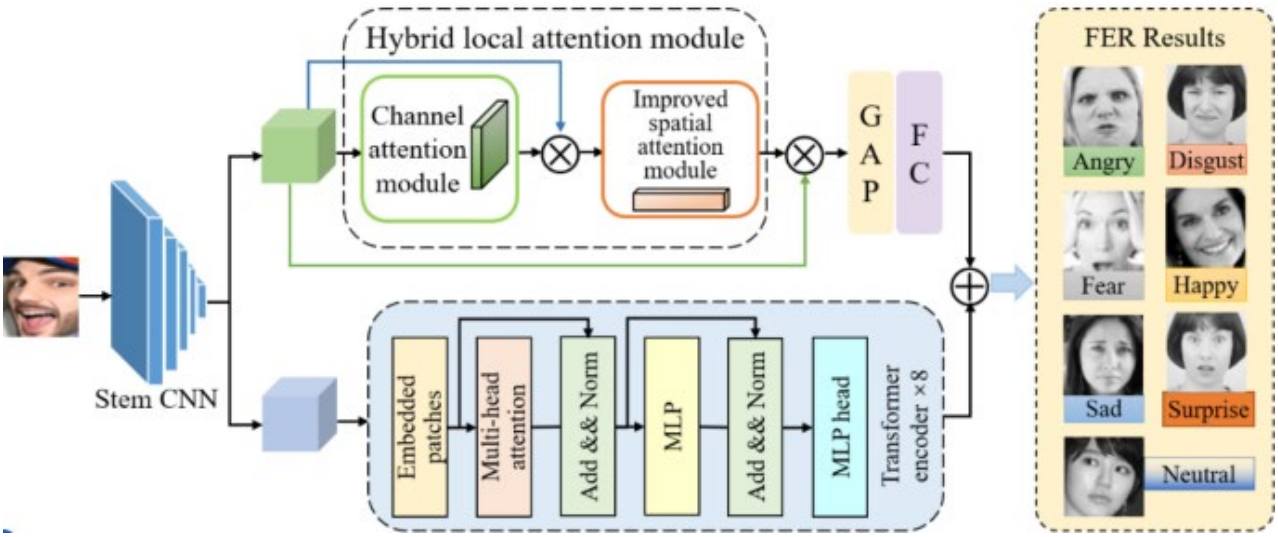

Fig. 2 *Schematic of the Hybrid Local-Attention Vision Transformer (HLA-ViT) network architecture* [18]

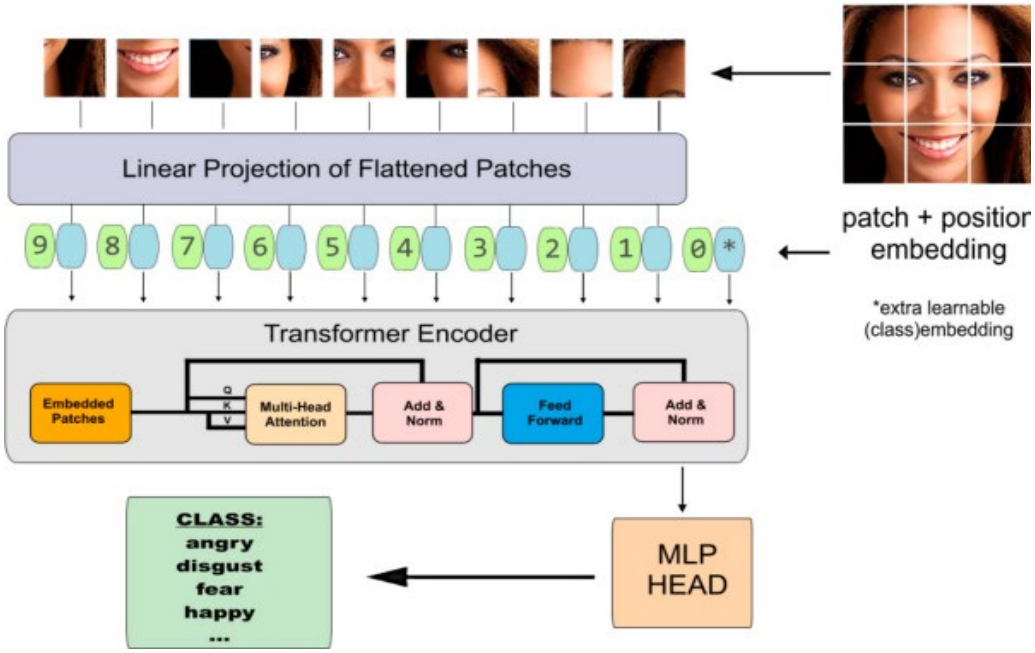

Fig. 3 *Internal structure of a base ViT model, showing the sequence of patch embedding, linear projection, and transformer encoder blocks for global dependency modelling.* [19]

2.2 Multimodal Fusion Approaches

Compared to unimodal FER, multimodal systems that combine facial, vocal, and physiological cues have been very successful by using the complementary affective information from different modalities. Fusion strategies are usually divided into early fusion, late fusion, and hybrid or adaptive fusion categories. Early fusion merges feature from several modalities at the representation level before classification, thus allowing cross, modal feature interactions but often having problems with high dimensionality and synchronization. Late fusion merges decisions or confidence scores of unimodal classifiers, thus allowing great modularity and robustness to missing modalities, however, with a drawback of limited cross, modal interaction [20]. Hybrid and adaptive fusion strategies, on the other hand, which are often realized through attention mechanisms or gating networks, determine the modality contributions dynamically depending on their reliability and thus can achieve better robustness in unconstrained environments.

Audio, visual fusion techniques, specifically, have shown steady performance improvements on benchmark datasets like Ryerson Audio, Visual Database of Emotional Speech and Song (RAVDESS), Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), and Aff-Wild2 by capturing both facial dynamics and vocal prosody, as visualized in Fig. 4. On these datasets, multimodal systems usually report absolute accuracy improvements of about 4-8% on RAVDESS and CREMA-D when compared to unimodal facial or speech models, with the in, the wild datasets such as Aff-Wild2 showing modest yet consistent gains of around 3-5%. Recognition performance is further improved by adaptive fusion mechanisms [21] such as cross, attention and reliability, aware gating that can detect noisy or partially missing modalities and hence emphasize the most informative signals at inference time. Multimodal systems that combine facial geometry with motor activity cues, for instance, have obtained accuracy levels of over 90%, thus, showing enhanced robustness against difficult in, vehicle conditions in the driver monitoring scenario [22, 23].

Hence, this comparative evidence is an overall indication that multimodal FER techniques regularly perform better than unimodal ones both in control environments and in the real world. Nevertheless, these improvements in performance are most of the time accompanied by increased computational complexity, dependency on sensors, and synchronization requirements that still pose limitations on scalability and real-time deployment. Such trade-offs are the reason for continuous research into lightweight and adaptive fusion architectures that can achieve a balance between robustness and efficiency.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 22-39

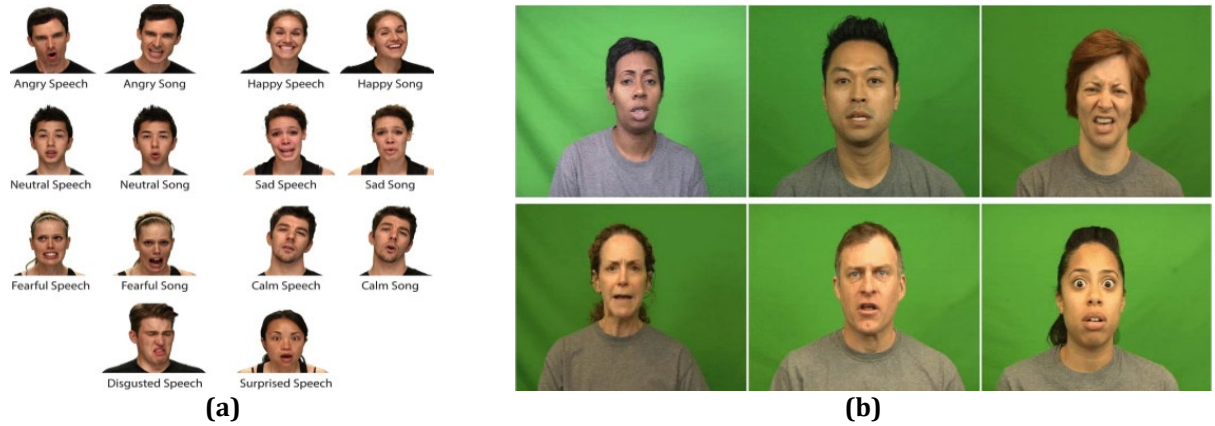

Fig. 4 *Representative video frame samples from standard multimodal datasets: (a) RAVDESS, focusing on speech-audio prosody* [21]; *and (b) CREMA-D, illustrating diverse emotional expressions in varying intensities* [16]

2.3 Surveys and Meta-Analyses

Several recent surveys and meta, analyses have summarized the gradual advances made in emotion recognition from both unimodal and multimodal sources, thus offering valuable insights into the trends of architectures and the challenges that remain. The different surveys agree that CNN coupled with attention mechanisms and hybrid CNN-transformer architectures are consistently reported to yield results that are superior to those of traditional handcrafted, feature, based methods in static image, based FER, whereas transformer, based and temporal attention models are able to achieve higher performance in video, based scenarios due to their capability of capturing long, range spatial and temporal dependencies [22]. Regarding the multimodal field, surveys consistently arrive at the conclusion that adaptive fusion strategies, especially attention, based and reliability, aware fusion, yield better results than early or late fusion approaches because they can handle modality noise and missing signals in a more efficient way.

Despite these improvements, surveys consistently pinpoint several issues that have not been resolved. One significant drawback is the inadequate representation of complex affective states, for instance, micro, expressions, compound or blended emotions, and subtle affective transitions, which are hardly detected by existing datasets and models. Furthermore, demographic and cultural biases continue to exist as most of the most commonly used benchmarks for which there is a lack of diversity in ethnicity, age, and contextual settings have been resulting in limited generalization of the model in real, world scenarios [23]. Surveys also acknowledge the difficulty of having large, scale, balanced multimodal datasets and the high computational cost of the latest models, which limit the performance of these models in real-time. In summary, these results exemplify that while there has been improvement in architectural performance, progress in dataset diversity, emotion coverage, and efficiency is still lacking, thus, the research gaps that are motivated and discussed in the following sections.

2.4 Summary of Comparative Studies

Table 1 summarizes key recent studies in FER and multimodal emotion recognition, which allows one to easily gauge how methodical advances align with the review goals. One consistent trend that can be seen from the table below is the move from typical CNN, based architectures to transformer, based and multimodal frameworks, which is largely driven by the objective of solving the issues that have been around for a long time, such as robustness to occlusion, class imbalance, and performance degradation in unconstrained environments [2]. Among the various transformer, based models, including vision transformers with auxiliary supervision and dual-stream attention mechanisms, improvements in recognition accuracy and robustness are quite evident across the standard benchmarks such as RAF-DB, FERPlus, and AffectNet, which implies that global attention mechanisms help to some extent in alleviating the problems of pose variation and occlusion [24].

Importantly, the comparative studies also demonstrate that newer methods specifically focus on resolving certain limitations of earlier FER systems, albeit with some trade-offs. Data balancing strategies and auxiliary action unit supervision have the effect of dataset bias removal and generalization improvement; thus they directly respond to the objective of robustness enhancement over imbalanced datasets. Adaptive multimodal fusion techniques, on the other hand, exceed the performance of unimodal models by being able to dynamically compensate for the modalities that are unreliable, thus they demonstrate the progress toward the solution of the real-world problems such as occlusion and noisy inputs. Nevertheless, this table also indicates that these improvements are very often accompanied by the increase in computational complexity and inference latency, which thereby limit the possibilities of deployment in real, time or resource, constrained applications. From an

objective, driven perspective, it is asserted that recent transformer-based and multimodal methods have significantly improved FER accuracy and robustness, however, they still do not solve the key deployment barriers completely which have been identified in this review. The issues of high computational cost, dependence on auxiliary annotations, and domain-specific assumptions are still there. Therefore, the comparative analysis leads to the conclusion that the current state, of, the art methods are incremental rather than complete solutions, thus the necessity of further research in lightweight architectures, adaptive fusion, and generalizable learning strategies that coordinate performance with efficiency and real, world applicability is still there.

Table 1 *Summary of recent FER and multimodal emotion recognition studies*

Method / Architecture	Datasets	Strengths	Weaknesses	Computational Cost	Accuracy Range	Study
ViT + Action Unit auxiliary supervision	RAF-DB, AffectNet, FERPlus	High accuracy (91% RAF-DB), robust to occlusion, reduces dataset bias	Relies on AU labels or pseudo-AUs; increased model complexity	High	86-91%	[24]
Dual stream (local attention + global ViT)	RAF-DB, FERPlus, AffectNet	Strong across datasets; good for occlusion/pose variation	More complex architecture; slower inference	High	88-92%	[18]
Comparative ViT study with data balancing	RAF-DB, FER2013	Shows importance of class balance; identifies top ViTs	Limited to ViTs; computationally heavy	Moderate-High	82-89%	[19]
Comparative study: Transformer vs CNN	FER datasets under perturbations	Transformers more robust to noise/occlusion	Higher inference cost; slower deployment	High	80-88%	[25]
Audio-visual fusion + LSTM temporal modeling	RAVDESS, CREMA-D, SAVEE	Captures temporal dynamics; boosts multimodal accuracy	Less generalizable to wild/noisy data	High	85-93%	[16]
Adaptive gating + recursive cross-attention	AffWild2	Handles weak modality complementarity; robust in wild data	Complex training; higher latency	High	86-90%	[12]
Facial geometry + motor activity fusion	Driver simulation datasets	Very high accuracy (~96%); domain-specific robustness	Limited to simulation; small emotion set	Moderate	92-96%	[26]
Survey of multimodal deep learning	Multiple	Highlights gaps in datasets and modalities	Not experimental	Low	N/A	[27]
FER survey: architecture & future	Multiple	Covers CNN/ViT hybrids, deployment issues	Limited multimodal focus	Low	N/A	[28]
Multimodal ER in conversations survey	IEMOCAP, CMU-MOSEI	Comprehensive review of conversational multimodality	Mostly survey; not new models	Low	N/A	[29]

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 22-39

2.5 Research Gaps

From this analysis, three major issues have been highlighted that are still the main reasons for lack of progress in FER. Firstly, the generalization of FER techniques to wild and in-the-wild datasets is still very weak, as models can hardly maintain their accuracy in real, world, imbalanced, or noisy conditions, although they achieve impressive results on curated benchmarks. In fact, in such natural settings as those in AffectNet and Aff-Wild2, where occlusion, head pose variation, and cultural diversity challenge robustness, the performance of models drops drastically. Secondly, the robustness of fusion across modalities has not been sufficiently researched yet, as most multimodal systems that fuse information from different modalities assume that all inputs are consistently reliable, however, in the real world, facial cues may be occluded, or speech signals may be noisy, thus the need for adaptive, reliability, aware fusion strategies becomes more evident. Lastly, problems related to efficiency and deployment remain, as a great number of cutting, edge transformer, based or ensemble methods are computationally intensive, thus they can hardly be used in real, time applications or environments with limited resources such as mobile health monitoring or automotive systems. Bridging these divides calls for not only methodological breakthroughs but also cross, disciplinary benchmarking to close the gap between academic performance metrics and industry adoption.

3. Review Framework and Classification Criteria

The classification criteria are designed to highlight the fundamental design decisions that have the most significant impact on recognition capability, robustness, and deployment feasibility, instead of merely concentrating on network architecture or dataset characteristics. In this review, methods for FER are classified first by the input modality (image, video, or multimodal) and the temporal modelling strategy. These criteria were selected because they represent the most extensively used and conceptually significant dimensions for differentiating FER systems across applications. Input modality directly determines the type of emotional information available to the model, while temporal modelling controls the capture of expression dynamics to the model both of which are essential factors in real, world emotion recognition. Different classification schemes based on specific network architectures such as CNN versus transformer, dataset type, or real, time feasibility are considered as secondary features and thus they are discussed within each category rather than being used as the primary organizing principle. Based on this rationale, existing studies are grouped into three broad categories:

a) Image-Based FER: These approaches operate on single static facial images and typically employ CNNs, attention-enhanced CNNs, or vision transformers to extract spatial features. Image-based FER methods are computationally efficient and well suited for real-time or resource-constrained applications. However, they are inherently limited in capturing temporal dynamics and subtle expression transitions.
b) Video-Based FER: Video-based methods incorporate temporal information from facial image sequences to model the evolution of expressions over time. Common techniques include recurrent neural networks (RNNs), long short-term memory (LSTM) networks, 3D CNNs, and transformer architectures with temporal encoding. These approaches improve recognition of subtle and spontaneous emotions but generally incur higher computational cost, which can complicate real-time deployment.
c) Multimodal FER: Multimodal FER integrates facial cues with complementary modalities such as speech, audio prosody, physiological signals Electroencephalogram (EEG) and Electrocardiogram (ECG), or contextual information. Fusion strategies are commonly categorized as early fusion (feature-level), late fusion (decision-level), or hybrid adaptive fusion. While multimodal approaches offer enhanced robustness under occlusion or noisy conditions, they introduce additional challenges related to sensor availability, synchronization, and computational complexity.

In each category, the studies that have been reviewed are analysed comparatively based on the datasets used (FER2013, RAF-DB, AffectNet, Aff-Wild2, CK+), the metrics of the evaluation (accuracy, F1, score, robustness indicators) and the target application domains, for example, healthcare, automotive monitoring, human-robot interaction, and education. The present review framework serves as an instrument for a systematic comparison of methodological trends and at the same time, it is conducive to the locating of the most frequent constraints and the open research questions that are dealt with in the sections following.

3.1 Image-Based FER

Image-based FER mainly depends on a single static facial image, from which discriminative spatial features of facial expressions are extracted. Previously, manually designed features like Local Binary Patterns (LBP) and Gabor filters were the main tools, but recently, their usage has been mostly replaced by deep learning methods, especially CNNs and ViTs, that allow end-to-end feature learning [30]. In general, CNN-based models yield recognition accuracies of about 70-75% on FER2013 and 85-90% on RAF-DB, thus making a significant leap over

the performance of the handcrafted feature, based methods on these datasets, as can be seen in Fig 5. This is evidenced by the results reported on standard benchmarks like FER2013, RAF-DB, and AffectNet [31].

(a) **(b)** **(c)**

Fig. 5 *Comparative visualization of sample images from widely used FER benchmarks: (a) FER2013 (low-resolution grayscale)* [19]; *(b) RAF-DB (real-world diverse poses)* [19]; *and (c) AffectNet (large-scale in-the-wild expressions)* [32]

CCNN-based structures remain the primary standard, often upgraded with residual connections and attention mechanisms. For instance, Attention-augmented CNNs have demonstrated accuracy levels exceeding 88% on RAF-DB while maintaining stability under partial occlusion [33]. Lightweight models, such as those utilizing MobileNet-inspired depth wise separable convolutions, significantly reduce computational costs while staying within 2–3% accuracy of larger models [33].

Recently, ViT models have gained prominence due to their self-attention capabilities. Specialized variants like DeiT and Swin Transformers have achieved 3–6% accuracy gains over CNN baselines on "in-the-wild" datasets like AffectNet [12, 13]. Furthermore, Hybrid CNN-ViT architectures, which prepended a CNN backbone for local feature extraction, have reached accuracies of 90% and above on controlled benchmarks like RAF-DB and FERPlus [34].

However, image-based FER has inherent limitations in reconstructing temporal changes since static frames cannot depict the development of facial expressions. In addition, the problem of generalization to different cultures, age groups, and naturally occurring emotional expressions is still limited by dataset bias and annotation subjectivity. The desire to overcome these restrictions drives the incorporation of temporal modelling and multimodal information, which is elaborated in the following sections on video-based and multimodal FER [31].

3.2 Video-Based FER

Video-based FER moves beyond the capabilities of static image analysis by directly incorporating the temporal changes of facial expressions in the videos. Emotions are dynamic and they change over time; thus, temporal modelling allows to detect even the slight changes of the intensity and micro-expressions that traditionally go unnoticed in the single images. Video-based FER methods usually utilize recurrent neural networks, 3D CNNs or transformer-based temporal encoders to acquire spatio-temporal representations. Fig. 6 shows architectural overview of a 3D CNN used for video-based FER. The diagram illustrates the extraction of spatio-temporal features by applying 3D convolutional kernels across consecutive video frames to capture the dynamic evolution of facial expressions.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 22-39

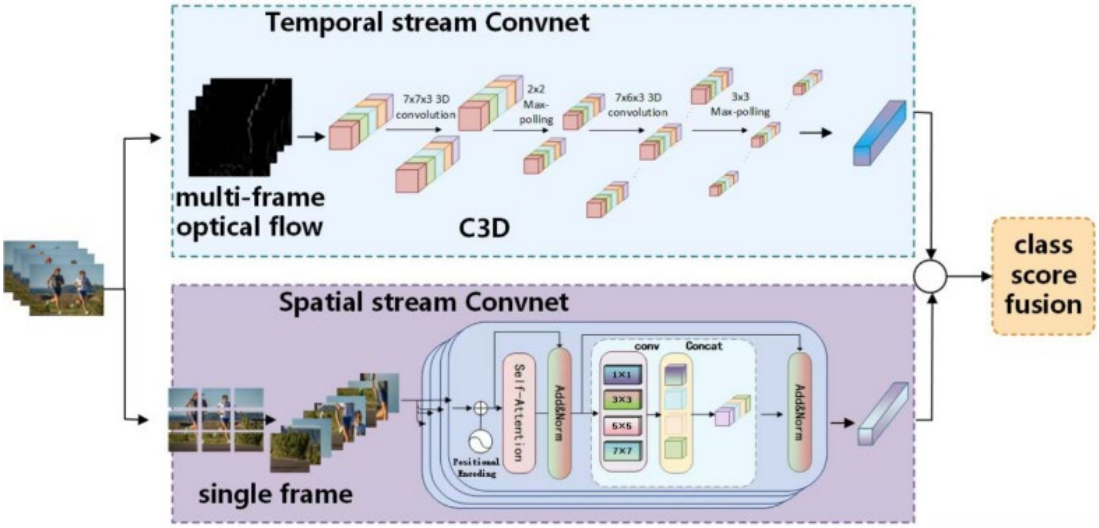

Fig. 6 *Spatio-temporal feature extraction in video-based FER: A representative architecture utilizing 3D CNNs to capture dynamic facial expression changes across consecutive frames* [35]

Early deep learning methods were mostly based on recurrent structures like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, which helped to capture sequential dependencies. LSTMs represent temporal information through gated memory cells that can still depend on long-term storage, whereas GRUs have a simplified structure with fewer parameters, thus memory usage is lower and convergence is faster. CNN-LSTM hybrid models [36] that use CNNs to extract spatial features and LSTMs to model temporal dynamics have shown accuracy improvements of around 5-8% over image, based baselines on controlled datasets like CK+ and Oulu-CASIA dataset [31]. But recurrent architectures handle frames one after another, thus inference latency is higher, and memory consumption is increased especially for long video sequences.

3D CNNs have been extensively used for video, based FER to overcome these drawbacks of single-frame models. They combine spatial and temporal information in a weighted along space, time convolutional kernels. Residual 3D CNN architectures supplemented with temporal attention mechanisms have been shown to achieve excellent results on in, the wild datasets such as Aff-Wild2 and AFEW, thus, outperforming CNN-LSTM models by up to 6% in terms of accuracy [37]. These networks are suitable for micro, expression recognition, as they follow the changes of the facial muscles that occur in very short durations of neighbouring frames. Unfortunately, 3D CNNs are very demanding in terms of computation, and hence memory and floating, point operations, which are usually several times more than those of 2D CNNs, thus, limiting the exposure of such models to the resource-constrained devices [38].

At the same time, transformer-based architectures have also been proposed for video FER, as they grant modelling of distant temporal dependencies by self, attention mechanisms. Contrary to recurrent models, temporal transformers operate on the whole frame sequences in parallel, thus, allowing faster modelling of global temporal context. Temporal vision transformer models [16] have been able to increase accuracy and F1- score by about 7% as compared to RNN and CNN, based approaches on large-scale datasets such as AffectNet, Video and Aff-Wild2. The hybrid combinations that mix 3D CNNs with transformers not only retain the fine, grained motion representation but also the long-range temporal modelling and thus, perform well in such applications as driver monitoring and human-robot interaction [14].

Despite these improvements, video-based FER is still heavily burdened with high computational and memory requirements that slow down real-time performance and scalability to a large extent. A large Graphical Processing Unit (GPU) memory and computational resources may be required to process lengthy video sequences with high frame rates; thus, it is difficult to deploy on edge devices. Moreover, because of occlusion, head pose variation, and illumination changes, spontaneous in-the-wild videos continue to be a difficult problem. Such compromises underscore the continuous demand for efficient temporal modelling methods that can maintain micro-expression sensitivity while lowering the computational load [39].

3.3 Multimodal FER

Multimodal FER is a technology that identifies facial emotions by combining facial cues with other modalities such as speech, audio prosody, or physiological signals like EEG, ECG, and even contextual information to give a more robust and naturalistic understanding of the emotion. The main reason behind this is that emotional expression

is always multimodal, and a sole dependence on facial information usually results in a decrease of performance when the face is occluded, there is a change in illumination or cultural variation.

Fusion strategies utilized in multimodal FER can be roughly divided into early fusion, late fusion, and hybrid or adaptive fusion. In the case of early fusion, features derived from each modality individually such as facial embeddings, speech spectrogram features, physiological descriptors are combined into one single joint representation before the classification. This method allows for direct cross, modal interaction at the feature level and has been shown to achieve higher accuracy on controlled datasets like Interactive Emotional Dyadic Motion Capture (IEMOCAP). Nevertheless, early fusion is affected by the lack of synchronization in time between modalities and usually has a problem with dimensionality explosion, thus, it becomes unstable for training and computationally cost [40].

Late fusion is a decision-level method that combines predictions or confidence scores from unimodal classifiers. For example, it could be averaging or weighting emotion probabilities obtained from facial and speech models. This method provides more flexibility and is less vulnerable to the absence of certain modalities, as individual classifiers can function independently. Research on datasets such as AffectNet and CREMA-D shows that ensemble-based late fusion consistently outperforms unimodal baselines [16]. However, late fusion cannot capture detailed cross, modal interactions and its performance can be greatly affected if a modality, for instance, speech, is dominant but becomes noisy or suffers from signal loss [41]. Recently, the emphasis of research has been on hybrid and adaptive fusion strategies that have the advantages of both early and late fusion. Typically, these methods use attention mechanisms or gating networks to dynamically determine the weights of different modalities based on their reliability. For example, cross-modal attention frameworks [30] enable the system to focus on facial cues when the audio is noisy and vice versa, thereby achieving better generalization in open scenarios. Adaptive transformer-based fusion models combining facial, audio, and physiological signals have set new performance levels on large, scale multimodal benchmarks such as Aff-Wild2, Multi and MSP-IMPROV. They report accuracy and F1-score improvements of around 5-10% compared to unimodal and static fusion baselines [42].

However, multimodal FER has several practical challenges that, directly, affect its performance and deployment. The synchronization of heterogeneous signals is very important, as a temporal mismatch between facial and speech cues may result in conflicting emotional evidence and thus recognition accuracy may decrease. Missing or corrupted modalities, which are common in real-world situations, can cause such a dramatic drop in the performance of fusion models that assume complete inputs that these models become unusable. Moreover, adaptive fusion architectures and multimodal transformers have very high computational and memory costs because of the parallel modality processing and attention operations and, therefore, they very often require a substantially larger number of resources than unimodal systems. These factors, thus, fuel a continuous effort of researchers in solving the issues of lightweight, reliability, aware fusion mechanisms and scalable multimodal datasets that can be used for robust and efficient FER in different application domains [43].

4. Comparative Review and Discussion

4.1 Comparative Performance Across FER Approaches

Studies have been consistent in showing that the three major categories of FER which are image-based, video-based, and multimodal, perform differently depending on dataset complexity, emotion granularity, and environmental conditions. Image-based methods continue to be very close in performance on constrained datasets with controlled acquisition settings. To illustrate, CNN and transformer-based models usually attain accuracy levels of around 70-75% on FER2013 and 85-90% on RAF-DB, with the most recent ViT-based architectures achieving accuracies up to 91% on RAF-DB under balanced training conditions [33]. Nevertheless, the performance is significantly lower when the methods are tested on in-the-wild datasets like AffectNet and Aff-Wild2, with the results going down to the 55-65% range most of the time, thus indicating that the methods are very sensitive to pose variation, occlusion, illumination changes, and spontaneous expression intensity [44].

Video-based FER methods overcome such constraints by introducing temporal dynamics as a key factor in their models. CNN-LSTM hybrids, 3D CNNs, and temporal transformers-based methods are consistently above the image baselines by around 3-7% in accuracy or F1-score on different benchmarks like CK+, Oulu-CASIA, and AFEW [16, 39]. As an example, 3D CNN, based models report a leap in the accuracy of their results from about 88% to over 93% on CK+, whereas the temporal transformer models obtain performance enhancements of 4-6% on difficult datasets like AFEW and Aff-Wild2 [35]. These enhancements are mainly exemplified in the case of micro-expression recognition, where the temporal modeling can more effectively capture the short-duration muscle movements. However, these improvements are associated with increased computational and memory demands, which may impose a limitation on the possibility of real-time applications [12].

Typically, multimodal FER systems attain the most robust aggregate performance by fusing the most informative affective cues derived from facial expressions, speech, audio prosody, physiological signals, or

contextual information. On controlled multimodal benchmarks such as IEMOCAP, MSP-IMPROV, and Aff-Wild2-Multi, accuracies and weighted F1-scores reported are often above 90%, which correspond to improvements of 5-10% over the respective unimodal baseline [11, 12, 31]. Nevertheless, their deployment in uncontrolled environments is still limited by various feasible issues such as sensor synchronization, missing or noisy modalities, and significantly higher computational overhead. These compromises emphasize the necessity of determining the most suitable FER methods depending on the application scenario, available resources, and robustness requisites [45]. As illustrated in Fig. 7, modern multimodal systems utilize adaptive reliability-aware gating mechanisms to dynamically weight signals. This allows the model to prioritize audio cues when facial landmarks are obscured by lighting or occlusion, and vice versa, ensuring stable performance in unconstrained environments.

4.2 Robustness and Generalization in Real World Settings

One of the main issues for FER systems is the problem of generalization beyond controlled laboratory environments. Models relying on images are very vulnerable to situations where parts of the face are hidden, such as when wearing a face mask, eyeglasses or having facial hair, because these things interfere with facial landmarks and texture cues that the model uses. Studies based on real-world data show that the drop in accuracy is in the range of 10-25% when the lower part of the face is covered by a mask at the same time, the drops in performance caused by occlusions due to glasses and hair are usually within the 5-15% interval, with the exact numbers depending on dataset and model architecture. Moreover, these impacts become stronger in the case of the in-the-wild datasets where occlusions are found to be oftentimes accompanied by pose variations and spontaneous expressions [47].

Noise coming from the surroundings continues to worsen the performance of FER of any model category. Changes in lighting, for example, situations of low, light or when the object is backlit, have a major impact on image, based models as they cause the contrast and facial detail to be reduced, which in turn leads to accuracy losses of around 10% or more on unconstrained datasets like AffectNet and Aff-Wild2 [48]. The presence of compression artifacts as well as low-resolution inputs that are typical of scenarios such as surveillance or web videos have the effect of changing the fine-grained facial features and simultaneously making it difficult for both CNN and transformer-based models to identify the actual face due to the noise that these features bring. On top of that, background clutter and partial visibility of the face make it even harder for feature extraction to take place since they bring along visual information that is not relevant.

Video-based FER methods alleviate some of these problems by making use of temporal continuity, thus allowing temporal smoothing that lessens the effect of transient occlusions and illumination fluctuations. Temporal models have shown increased robustness to short-term occlusion and motion blur, with improvements of 3-6% over image-based baselines in difficult video datasets. Unfortunately, their power is limited in low, frame, rate recordings, heavily compressed videos, or long-duration occlusions, where temporal cues are ambiguous or inconsistent [49].

Multimodal FER methods are the most robust in real-life conditions as they allow one modality to be the backup when another is deteriorated. For example, speech prosody or audio features can facilitate emotion recognition when facial cues are occluded or are in the dark. However, multimodal systems are still vulnerable to the following practical constraints, modality synchronization errors, missing data streams, and uneven signal quality that can drastically decrease fusion effectiveness. Demographic bias of training datasets, especially with regards to ethnicity, age, and cultural expression styles is a problem that affects all FER categories and has been a major factor in limiting generalization. This fact emphasizes the necessity of more diverse and balanced datasets to be used for the reliable deployment of FER in real, world applications.

Fig. 7 *Multimodal fusion framework for robust emotion recognition: Illustration of an adaptive reliability-aware gating mechanism that dynamically weights facial and audio signals to handle noisy or occluded inputs.* [46]

4.3 Ethical Considerations, Privacy and Social Impact

The use of FER technologies raises ethical and societal concerns which are the main challenges of the deployment and must be solved to be used responsibly. The main concern is privacy because facial data is one of the most sensitive biometric pieces of information. Without strong data governance and consent, there is a great possibility of unauthorized surveillance and data breaches. Besides that, algorithmic fairness is still a very important issue, most current models have demographic biases because their training datasets are not balanced, which leads to lower accuracy for the ethnic or age groups that are less represented. If these models are used in sectors like recruitment or security, it can lead to discriminatory outcomes. At last, the possibility of the device abuse, such as identification of the emotional state of persons without their consent, requires the creation of AI systems that are transparent and understandable not only for gaining the public trust but also for being in accord with the upcoming global data regulations.

4.4 Deployment Feasibility and Computational Efficiency

The trade-off between model complexity, accuracy, and the available computational resources is the main factor that determines whether FER systems can be deployed in real-time applications. Lightweight CNN architectures such as MobileNet and EfficientNet, based models are very good for deployment on mobile and embedded platforms because they require low computational power and memory. These models usually run at 10-30 ms per frame on modern mobile GPUs or edge devices, their memory footprints are less than 50 MB, and their computational costs are in the range of 0.3-1.0 Giga Floating-Point Operations Per Second (GFLOPs), thus, they allow real-time inference with low energy consumption [50].

On the other hand, transformer-based FER systems, although attaining state-of-the-art recognition accuracy, have very high computational costs. ViT and hybrid CNN-ViT models typically need several GFLOPs per inference and their memory usage is more than 100-300 MB, thus, the inference time is 50-150 ms per frame on standard GPUs [51]. Because of these features, they cannot be directly deployed on devices with limited resources unless model compression techniques like pruning, quantization, or knowledge distillation are used to decrease latency and energy consumption.

Video-based FER methods have exponentially increased computational demands as they require temporal modeling over several frames. Generally, architecture based on 3D CNNs or CNN-LSTM hybrids work on a few video clips rather than a single frame, thus their inference times are more than 100 ms per clip and the memory usage is several times higher than that of image-based approaches [36]. In the same way, multimodal FER systems have some additional overheads because of parallel modality processing and fusion operations. Some studies reveal that multimodal transformer, based systems can be two, three times as demanding in terms of computational resources as unimodal models, which results in higher energy consumption due to the synchronized processing of audio, visual, and physiological signals.

Int. Journal of Integrated Engineering Vol. 18 No. 4 (2026) p. 22-39

While facing such difficulties, the deployment of these systems is becoming more practical due to recent advances in efficient model design and hardware acceleration. The optimized temporal sampling strategies, the lightweight attention mechanisms, and the edge, oriented fusion architecture have largely lowered the latency and at the same time, the accuracy was preserved. These improvements together with the use of dedicated hardware accelerators or edge, computing infrastructures are turning the near real-time deployment of video-based and multimodal FER systems feasible in a wide range of applications such as driver monitoring, assistive robotics, and human-computer interaction.

4.5 Trends Observed Across Recent FER Literature

Three dominant trends result from the comparative analysis of the recent FER literature, which reflects the field's transition to more expressive, robust, and application, ready systems:

a) Shift toward contextual and temporal modeling: The transition from static image-based FER to video-based and multimodal approaches is strongly indicated by most of the recent articles. Image-based CNN and ViT models achieve good results on constrained datasets but fail to capture emotion dynamics and contextual cues in real-world scenarios. Consequently, temporal models like CNN-LSTM hybrids [36], 3D CNNs with temporal attention [39], and temporal vision transformers [16] have been utilized to show consistent accuracy improvements of around 3-7% on datasets such as CK+, AFEW, and Aff-Wild2 [37]. At the same time, multimodal frameworks combining facial, audio, and physiological signals help even more in the contextual understanding of the emotion, especially in the case of spontaneity or noise in the environment. These changes demonstrate that there is an increasing understanding that the expression of emotion is a dynamic and context, dependent phenomenon and thus the models should be able to conduct temporal and cross, modal reasoning [11, 12].

b) Performance efficiency trade, off and rise of hybrid architectures: Although transformer-based and multimodal FER systems have been able to achieve state-of-the-art results, they are usually accompanied with high computational and memory costs, thereby limiting their deployment in real-time scenarios. The study of ViT-based FER models shows that the accuracy can be improved by 4-6% as compared to the CNN baselines, but the inference latency and resource consumption are increased [52]. To solve this problem of trade-off-the latest studies have turned their attention to hybrid architectures that combine the strengths of different paradigms in a complementary way. Some instances are CNN-ViT hybrids for spatial feature extraction with global attention, 3D CNN-transformer models for spatio-temporal modeling in video FER [21], and adaptive multimodal transformer frameworks that selectively weight facial, audio, and physiological inputs. These hybrid methods can achieve accuracy at a similar level of the fully transformer, based pipelines while the computational overhead is lessened, thus, they represent a viable solution to the problem of efficient yet accurate FER systems.

c) Growing emphasis on real, -world robustness and deployment readiness: Recent research on FER emphasizes systems robustness to real-world scenarios, which can include occlusion, cultural diversity, lighting changes, and background noise. Large-scale in-the-wild datasets such as AffectNet, Aff-Wild2, and MSP-IMPROV have become the standard instruments to evaluate the models, thus reflecting the shift of the evaluation methods from lab-only to real-world conditions. Models that use temporal smoothing, cross-modal attention, or adaptive fusion techniques become more robust to occlusion and environmental noise [9]. However, comparative evaluation experiments still find that the difference between laboratory and real-world performance is very large, especially for demographic groups that are underrepresented and spontaneous emotion expressions. This difference points to the necessity of more diversified datasets, interpretable FER models, and deployment, aware evaluation protocols that can help close the gap between academic benchmarks and real-world applications.

4.6 Application Domains of FER

FER technologies are becoming less of a science to be researched in the lab, and more of a real-world application in different sectors such as healthcare, automotive safety, education, human-robot interaction, and surveillance industries. In healthcare, FER has been utilized in patient monitoring, mental health assessment, and pain detection. Clinical trials of video-based FER systems have shown that the engagement of patients and the accuracy of their emotional state assessment made by the staff can be improved around 10-15% just by the use of these systems as compared to the situation when no technology is implemented and the observation is done manually [53]. Besides that, the multimodal FER systems which combine facial cues with speech and physiological signals have demonstrated higher levels of detection of affect in hospital and telemedicine settings, however, their installation is limited by the need for privacy and regulations on data governance.

Aside from this, the use of FER in the automotive sector aims mainly at the driver monitoring systems which can detect the conditions of the driver such as fatigue, stress, and emotional distraction. The incorporation of video-based FER models with temporal analysis has led to an improvement in the accuracy of driver drowsiness

and stress detection resulting in a 5-10% increment as compared with static image, based approaches under real driving conditions. Some advanced driver-assistance systems (ADAS) and in-cabin monitoring solutions are now benefited from the implementation of FER modules to promote safety by either giving the warning signals or cooperating with the vehicle to make it behave differently according to the condition of the driver that has been detected [54].

In the realm of educational technologies, weight-reduced image-based FER models have been successfully merged with intelligent tutoring systems as well as online learning platforms for the purpose of recording student engagement and emotional reaction. Several studies show that engagement estimation gets improved by about 8-12% thus, enabling the adaptive feedback mechanisms that personalize learning content. Owing to computational limitations and privacy concerns, these systems are mostly equipped with efficient CNN-based models instead of complicated temporal or multimodal architectures [55].

However, the issues of ethics and society have a significant influence on the deployment of FER and they are still there despite the promising applications. These problems concerning privacy, consent, algorithmic bias, and fairness that are most noticeable in the cases of surveillance and clinics, constitute the core of the difficulties that have been discussed. These challenges emphasize the existence of a necessity for transparent, easily interpretable, and regulation, compliant FER systems aimed at ensuring their responsible use in various fields of application.

5. Conclusion

FER has basically changed its direction, where it has moved to dynamic video, based and multimodal architectures from static, image-based analysis. The paper has gone through these techniques in detail, showing that the improvements in ViTs, 3D CNNs, and the use of adaptive fusion mechanisms have resulted in measurable performance increases for the given standard benchmarks. Although image, based methods can still achieve very high accuracy on tightly controlled datasets like RAF-DB (up to 91%), the significant fall of their performance at in-the-wild scenarios (most of the time, it is around the 55-65% range) shows that combining temporal and multimodal data is inevitable if we want our models to be robust outside of the lab.

The comparative analysis brings to the fore the differences in accuracy, robustness, and the possibility of deployment, which are essentially the trade-offs. The temporal modelling in video-based FER has elevated the recognition of subtle and micro, expressions by 3-7%, however, it comes with the cost of higher inference latency and memory overhead. Multimodal systems combining facial, audio, and physiological signals (EEG, ECG) are the most reliable, often achieving more than 90% accuracy on controlled benchmarks. Nevertheless, these improvements are counterbalanced by the difficulties of sensor synchronization and a 2-3x increase in computational demand compared to unimodal baselines. On the other hand, lightweight models such as MobileNet-based CNNs are still the best option for the edge deployment, thus providing a sustainable balance with inference times as low as 10-30 ms per frame.

Despite these achievements, there are still significant gaps in research that have to be addressed. Current models are very sensitive to environmental noise, long, term occlusions, and demographic biases. These biases are especially visible in the differences in performance where research has found a 10-15% difference in accuracy when testing different ethnicities and age groups that is most of the time caused by the over, representation of certain populations in the training sets. Besides that, the large dependence on full data streams makes a lot of multimodal systems quite susceptible to major drops in performance when there is a loss of a signal.

To resolve those difficulties, research to be done in the future should focus on three main aspects, firstly, improving generalization by using self, supervised learning which will help in cutting down the annotation costs and alleviating the dataset imbalance problem. Secondly, creating efficient architectures like hybrid CNN-Transformer models and reliability, aware gating mechanisms that can retain state-of-the-art performance under very strict energy constraints of mobile devices, and thirdly, building trustworthy AI by instrumenting ethical frameworks such as privacy, preserving techniques, algorithmic fairness, and explainability to give FER systems a responsible deployment.

In summary, FER is becoming a real-world problem from a lab-based research area, which means a high impact technology in healthcare, automotive safety, and education. The coming generation of FER systems should still be accurate, but they also must be efficient and able to handle noise from the real-world and ethically correct to be accepted by society on a large scale.

Acknowledgement

This work was supported by the Geran Penyelidikan Pasca Siswazah (GPPS), Pusat Pengurusan Penyelidikan, Universiti Tun Hussein Onn Malaysia (Grant Code: J052). The authors gratefully acknowledge the financial support provided by Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) under Grant RDU210314. The publication of this research was made possible through financial support from the Universiti Tun Hussein Onn Malaysia (UTHM) Publisher's Office via the Publication Fund (E15216).

Conflict of Interest

Authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Anbananthan Pillai Munanday, Norazlianie Sazali;* ***data collection:*** *Anbananthan Pillai Munanday, Norazlianie Sazali, Arigela Satya Veerendra;* ***analysis and interpretation of results:*** *Anbananthan Pillai Munanday, Norazlianie Sazali,Kumaran Kadirgama;* ***draft manuscript preparation:*** *Anbananthan Pillai Munanday, Norazlianie Sazali, Kumaran Kadirgama. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Yaqoob, S., Ullah, F., AbuAli, N., Anwar, H., Saeed, N., & Hayajneh, M. (2025). AIoT for human emotion recognition: Potentials, challenges, and healthcare applications. Computer Science Review, 60, 100859. https://doi.org/10.1016/j.cosrev.2025.100859

[2] Jayaswal, R., Ansari, Mohd. A., Dixit, M., Singh, D. K., & Ahmad, S. (2025). Advances in facial expression recognition technologies for emotion analysis. Discover Computing, 28(1). https://doi.org/10.1007/s10791-025-09699-8

[3] Zaman, K., Khan, R., Zengkang, G., Khan, S. U., Ali, F., & Hussain, T. (2025). Accurately recognizing driver emotions through using CNN fused features and NasNet-large model. Alexandria Engineering Journal, 134, 177–196. https://doi.org/10.1016/j.aej.2025.12.010

[4] Aly, M., & Alotaibi, N. S. (2025). A comprehensive deep learning framework for real time emotion detection in online learning using hybrid models. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-26381-7

[5] Tagmatova, Z., Umirzakova, S., Kutlimuratov, A., Abdusalomov, A., & Im Cho, Y. (2025). A Hyper-Attentive Multimodal Transformer for Real-Time and Robust Facial Expression Recognition. Applied Sciences, 15(13), 7100. https://doi.org/10.3390/app15137100

[6] Safarov, F., Kutlimuratov, A., Khojamuratova, U., Abdusalomov, A., & Cho, Y.-I. (2025). Enhanced AlexNet with Gabor and Local Binary Pattern Features for Improved Facial Emotion Recognition. Sensors, 25(12), 3832.https://doi.org/10.3390/s25123832

[7] Kim, C.-L., & Kim, B.-G. (2023). Few-shot learning for facial expression recognition: a comprehensive survey. Journal of Real-Time Image Processing, 20(3). https://doi.org/10.1007/s11554-023-01310-x

[8] Mattioli, M., & Cabitza, F. (2024). Not in My Face: Challenges and Ethical Considerations in Automatic Face Emotion Recognition Technology. Machine Learning and Knowledge Extraction, 6(4), 2201–2231. https://doi.org/10.3390/make6040109

[9] Kus, I., Kocak, C., & Keles, A. (2025). A systematic review of vision transformer and explainable AI advances in multimodal facial expression recognition. Intelligent Systems with Applications, 29, 200615. https://doi.org/10.1016/j.iswa.2025.200615

[10] Ketan Sarvakar, & Rana, K. (2025). Revolutionizing facial emotion recognition: in-depth analysis of cutting-edge models, methodologies, and datasets. Discover Artificial Intelligence, 5(1). https://doi.org/10.1007/s44163-025-00553-w

[11] Elharrouss, O., Yassine Himeur, Mahmood, Y., Saed Alrabaee, Abdelmalik Ouamane, Faycal Bensaali, Yassine Bechqito, & Ammar Chouchane. (2025). ViTs as backbones: Leveraging vision transformers for feature extraction. Information Fusion, 102951–102951. https://doi.org/10.1016/j.inffus.2025.102951

[12] Wang, A. (2025). Vision Transformers (ViTs): A New Era in Computer Vision – A Review. Journal of Computing and Electronic Information Management, 18(3), 23–30. https://doi.org/10.54097/41t0vx90

[13] Huang, Z.-Y., Chiang, C.-C., Chen, J.-H., Chen, Y.-C., Chung, H.-L., Cai, Y.-P., & Hsu, H.-C. (2023). A study on computer vision for facial emotion recognition. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-35446-4

[14] Mao, S., Li, X., Zhang, F., Peng, X., & Yang, Y. (2025). Facial Action Units as a Joint Dataset Training Bridge for Facial Expression Recognition. IEEE Transactions on Multimedia, 1–12. https://doi.org/10.1109/tmm.2025.3535327

[15] Jeon, S., Lee, J., & Lee, Y.-J. (2025). Dual-Stream Former: A Dual-Branch Transformer Architecture for Visual Speech Recognition. AI, 6(9), 222–222. https://doi.org/10.3390/ai6090222

[16] Salas-Cáceres, J., Lorenzo-Navarro, J., Freire-Obregón, D. et al. Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics. Multimed Tools Appl 84, 27327–27343 (2025). https://doi.org/10.1007/s11042-024-20227-6

[17] Ramakrishnan, S., & Babu, D. (2025). Improving Multi-Label Emotion Classification on Imbalanced Social Media Data with BERT and Clipped Asymmetric Loss. IEEE Access, 1–1. https://doi.org/10.1109/access.2025.3557091

[18] Tian, Y., Zhu, J., Yao, H., & Chen, D. (2024). Facial expression recognition based on Vision Transformer with hybrid local attention. Applied Sciences, 14(15), 6471. https://doi.org/10.3390/app14156471

[19] Bobojanov, S., Kim, B. M., Arabboev, M., & Begmatov, S. (2023). Comparative Analysis of Vision Transformer Models for Facial Emotion Recognition Using Augmented Balanced Datasets. Applied Sciences, 13(22), Article 12271. https://doi.org/10.3390/app132212271

[20] Moorthy, S., & Moon, Y.-K. (2025). Hybrid Multi-Attention Network for Audio–Visual Emotion Recognition Through Multimodal Feature Fusion. Mathematics, 13(7), 1100–1100. https://doi.org/10.3390/math13071100

[21] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLOS One 13(5):e0196391. https://doi.org/10.1371/journal.pone.0196391

[22] Ahmed, N., Aghbari, Z. A., & Girija, S. (2023). A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications, 17, 200171. https://doi.org/10.1016/j.iswa.2022.200171

[23] Serrano, S., Serghini, O., Esposito, G., Carbone, S., Mento, C., Floris, A., Porcu, S., & Atzori, L. (2025). Review and Comparative Analysis of Databases for Speech Emotion Recognition. Data, 10(10), 164. https://doi.org/10.3390/data10100164

[24] Mao, S., Li, X., Wu, Q., & Peng, X. (2022). AU-Aware Vision Transformers for Biased Facial Expression Recognition. arXiv preprint. https://doi.org/10.48550/arXiv.2211.06609

[25] Are transformer-based models more robust than CNN-based models? (2024). Neural Networks, 175, 106–118. https://doi.org/10.1016/j.neunet.2023.12.009

[26] Praveen, R. G., & Alam, J. (2025). Handling weak complementary relationships for audio-visual emotion recognition. arXiv preprint. https://arxiv.org/abs/2503.12261

[27] Jia, M., & Sun, Z. (2024). A survey of multi-modal emotion recognition based on deep learning. Highlights in Science, Engineering and Technology, 119, 533-540.

[28] Ullah, S., Ou, J., Xie, Y., & Tian, W. (2024). Facial expression recognition survey: A vision, architectural elements, and future directions. PeerJ Computer Science, 10, e2024. https://doi.org/10.7717/peerj-cs.2024

[29] Wu, C., Cai, Y., Zhu, P., Xue, Y., Gong, Z., Hirschberg, J., & Ma, B. (2025). Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. ResearchGate Preprint.

[30] Nanni, L., Lumini, A., & Brahnam, S. (2012). Survey on LBP based texture descriptors for image classification. Expert Systems with Applications, 39(3), 3634–3641. https://doi.org/10.1016/j.eswa.2011.09.054

[31] Sajjad, M., Ullah, F. U. M., Ullah, M., Christodoulou, G., Alaya Cheikh, F., Hijji, M., Muhammad, K., & Rodrigues, J. J. P. C. (2023). A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. Alexandria Engineering Journal, 68, 817–840. https://doi.org/10.1016/j.aej.2023.01.017

[32] Z. Chen, L. Yan, H. Wang, and B. Adamyk, "Improved Facial Expression Recognition Algorithm Based on Local Feature Enhancement and Global Information Association," Electronics, vol. 13, no. 14, p. 2813, Jul. 2024, doi: https://doi.org/10.3390/electronics13142813.

[33] Li, N., Huang, Y., Wang, Z., Fan, Z., Li, X., & Xiao, Z. (2024). Enhanced Hybrid Vision Transformer with Multi-Scale Feature Integration and Patch Dropping for Facial Expression Recognition. Sensors, 24(13), 4153. https://doi.org/10.3390/s24134153

[34] Chen, X., Zheng, X., Sun, K., Liu, W., & Zhang, Y. (2023). Self-supervised vision transformer-based few-shot learning for facial expression recognition. Information Sciences, 634, 206–226. https://doi.org/10.1016/j.ins.2023.03.105

[35] Liu, M., Li, W., He, B., Wang, C., & Qu, L. (2025). Human Action Recognition Based on 3D Convolution and Multi-Attention Transformer. Applied Sciences, 15(5), 2695. https://doi.org/10.3390/app15052695

[36] Kim, T.-Y., Yun, H.-S., Yoon, H.-M., & Lee, S.-J. (2025). Comparative Analysis and Validation of LSTM and GRU Models for Predicting Annual Mean Sea Level in the East Sea: A Case Study of Ulleungdo Island. Applied Sciences, 15(20), 11067. https://doi.org/10.3390/app152011067

[37] Kopalidis, T., Vassilios Solachidis, Vretos, N., & Daras, P. (2024). Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets. Information, 15(3), 135–135. https://doi.org/10.3390/info15030135

[38] Liu, Z., Chow, P., Xu, J., Jiang, J., Dou, Y., & Zhou, J. (2019). A Uniform Architecture Design for Accelerating 2D and 3D CNNs on FPGAs. Electronics, 8(1), 65. https://doi.org/10.3390/electronics8010065

[39] Yaddaden, Y. (2025). Efficient Dynamic Emotion Recognition from Facial Expressions Using Statistical Spatio-Temporal Geometric Features. Big Data and Cognitive Computing, 9(8), 213. https://doi.org/10.3390/bdcc9080213

[40] Chaves-Villota, A., Jimenez-Martín, A., Jojoa-Acosta, M., Bahillo, A., & García-Domínguez, J. J. (2025). Deep feature representations and fusion strategies for speech emotion recognition from acoustic and linguistic modalities: A systematic review. Computer Speech & Language, 96, 101873–101873. https://doi.org/10.1016/j.csl.2025.101873

[41] Xue, Z., & Radu Marculescu. (2023). Dynamic Multimodal Fusion. ArXiv (Cornell University). https://doi.org/10.1109/cvprw59228.2023.00256

[42] Sridhar, K., & Busso, C. (2022). Unsupervised Personalization of an Emotion Recognition System: The Unique Properties of the Externalization of Valence in Speech. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.07876

[43] Shehu, H. A., Browne, W. N., & Eisenbarth, H. (2025). Emotion categorization from facial expressions: A review of datasets, methods, and research directions. Neurocomputing, 624, 129367. https://doi.org/10.1016/j.neucom.2025.129367

[44] Gómez-Sirvent, J. L., López de la Rosa, F., López, M. T., & Fernández-Caballero, A. (2023). Facial Expression Recognition in the Wild for Low-Resolution Images Using Voting Residual Network. Electronics, 12(18), 3837. https://doi.org/10.3390/electronics12183837

[45] Mostert, W., Kurien, A., & Djouani, K. (2025). Multi-Modal Emotion Detection and Tracking System Using AI Techniques. Computers, 14(10), 441. https://doi.org/10.3390/computers14100441

[46] H. Liu et al., "HRHPE: FRoI Guides Heterogeneous Relationship Representation Learning for Precise Head Pose Estimation," Neurocomputing, vol. 647, Art. no. 130623, 2025, doi: https://doi.org/10.1016/j.neucom.2025.130623.

[47] Alyuz, N., Gokberk, B., & Akarun, L. (2013). 3-D Face Recognition Under Occlusion Using Masked Projection. IEEE Transactions on Information Forensics and Security, 8(5), 789–802. https://doi.org/10.1109/tifs.2013.2256130

[48] Rodríguez-Rodríguez, J. A., López-Rubio, E., Ángel-Ruiz, J. A., & Molina-Cabello, M. A. (2024). The Impact of Noise and Brightness on Object Detection Methods. Sensors, 24(3), 821. https://doi.org/10.3390/s24030821

[49] Panigrahi, U., Sahoo, P. K., Panda, M. K., Parija, S. R., Jain, P., Lu, L., & Liu, H. (2025). An enhanced artificial learning framework for moving object detection in challenging video scenes. Alexandria Engineering Journal, 129, 329–345. https://doi.org/10.1016/j.aej.2025.05.082

[50] Cordova-Cardenas, R., Amor, D., & Álvaro Gutiérrez. (2025). Edge AI in Practice: A Survey and Deployment Framework for Neural Networks on Embedded Systems. Electronics, 14(24), 4877–4877. https://doi.org/10.3390/electronics14244877

[51] Li, Z., Paolieri, M., & Golubchik, L. (2025). A Study on Inference Latency for Vision Transformers on Mobile Devices. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2510.25166

[52] Yang, Q., He, Y., Chen, H., Wu, Y., & Rao, Z. (2025). A Novel Lightweight Facial Expression Recognition Network Based on Deep Shallow Network Fusion and Attention Mechanism. Algorithms, 18(8), 473–473. https://doi.org/10.3390/a18080473

[53] Shehada, D., Tawfik, H., Bouridane, A., & Hussain, A. (2025). An Explainable Framework for Mental Health Monitoring Using Lightweight and Privacy-Preserving Federated Facial Emotion Recognition. Sensors, 25(23), 7320. https://doi.org/10.3390/s25237320

[54] Chew, Y. X., Razak, S. F. A., Yogarayan, S., & Sayed Ismail, S. N. M. (2024). Dual-Modal Drowsiness Detection to Enhance Driver Safety. Computers, Materials and Continua, 81(3), 4397–4417. https://doi.org/10.32604/cmc.2024.056367

[55] Yu, C., & Yu, X. (2025). Application of mobile learning system based on convolutional network technology in students' open teaching strategies. Scientific Reports, 15(1), 41646–41646. https://doi.org/10.1038/s41598-025-25532-0