STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–30.
Published online in International Academic Press (www.IAPress.org)

From Facial Emotion Recognition to Emotion Authenticity Detection: A Comprehensive Review of Datasets, Methods, Challenges, and Future Directions

Sushanta Kumar Sahu*, Chakrapani Ghadai

Odisha University of Technology and Research, Bhubaneswar, India

Abstract Facial expressions are one of the most important forms of nonverbal human communication. Automatic detection of posed (not felt) versus felt facial emotions is still an open problem for the computer vision community. Although conventional facial emotion recognition (FER) systems have achieved remarkable performance in classifying basic emotional categories, distinguishing genuine (felt) emotions from posed or manipulated facial expressions remains significantly more challenging due to subtle facial dynamics, limited annotated datasets, and poor cross-domain generalization. Unlike existing FER surveys that primarily focus on emotion classification, this review systematically synthesizes recent advances in emotion authenticity detection by integrating datasets, feature representations, deep learning architectures, evaluation strategies, challenges, and emerging research directions within a unified framework. We provide a critical comparison of state-of-the-art deep-learning architectures (CNNs, LSTMs, Transformers) along with graph-based and multimodal learning approaches and benchmark their performance on laboratory-collected as well as in-the-wild datasets. Additionally, we create a taxonomy of laboratory-controlled, in-the-wild, dynamic, multimodal, and micro-expression/veracity annotated datasets and identify gaps in diversity and number of classes within these datasets. Furthermore, we critically analyze the strengths, limitations, and applicability of existing methods under different acquisition conditions while highlighting the factors affecting robustness and generalization. We conclude with a discussion on emerging directions such as multimodal fusion and physiological analysis, explainable artificial intelligence, trustworthy affective computing, and robust cross-domain emotion authenticity verification to help narrow the validity gap. The review further identifies current research gaps and presents a future research roadmap to facilitate the development of reliable and trustworthy emotion authenticity detection systems for next-generation affective computing applications.

Keywords Facial Emotion Recognition, Genuine and Fake Emotions, Facial Expression Analysis, Emotion Authenticity Detection, Deep Learning, Computer Vision, Affective Computing

DOI: 10.19139/soic-2310-5070-4298

1. Introduction

Facial expressions are one of the most innate forms of expressing human feelings and thoughts [1]. Moreover, human emotions can be easily understood through facial expressions. Detecting facial emotions correctly is gaining importance in the fields of computer vision and artificial intelligence. Facial emotion detection systems have several applications, including healthcare, surveillance, education, human–computer interaction, and psychology. Recent advancements in deep learning architectures and image processing have attracted considerable attention towards developing robust facial emotion recognition

*Correspondence to: Sushanta Kumar Sahu (Email: sksahu@outr.ac.in). School of Electronic Sciences, Odisha University of Technology and Research. Bhubaneswar, Odisha, India.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

(FER) systems. Researchers have been trying to develop intelligent systems capable of recognizing facial emotions for many decades; however, it has proven to be a very difficult problem owing to the complexities involved with human emotions and their facial movements.

Humans communicate emotions primarily through two channels: verbal and nonverbal communication. Verbal emotions can be interpreted through speech, whereas nonverbal emotions are communicated through facial expressions, body gestures, and tone of voice. Each individual can portray multiple emotions, which are commonly classified into seven categories: happy, sad, angry, fearful, disgusted, surprised, and neutral. Emotions that are a combination of these seven categories are known as compound emotions. The rapid proliferation of synthetic facial media, commonly referred to as "deepfakes," has introduced new challenges for conventional facial emotion recognition systems [2]. In this review, *genuine* facial expressions refer to spontaneous emotional displays arising from an individual's actual affective state, whereas *posed* facial expressions are intentionally produced and may not reflect the underlying emotion. In contrast, *deepfake facial expressions* are synthetically generated or manipulated facial contents created using generative artificial intelligence techniques, and *micro-expressions* are involuntary facial movements of very short duration that may reveal concealed emotions. Although these concepts are closely related, they represent distinct research problems with different datasets, annotation protocols, learning objectives, and evaluation criteria. Throughout this review, emotion authenticity detection refers to the task of distinguishing genuine from posed or manipulated facial expressions rather than conventional facial emotion classification.

Whereas the first generation of generative architectures failed to create photo-realistic outputs, present-day frameworks empowered by Generative Adversarial Networks (GANs) and diffusion models have surpassed the "unidentifiable threshold." Current algorithms produce near-realistic faces with highly consistent facial geometries and temporally coherent facial dynamics. These artificially generated faces pose significant challenges to conventional facial emotion recognition systems because correctly classifying an emotional category does not necessarily imply that the displayed facial expression is genuine. As synthetic media generation and authenticity concerns have expanded, recent research has increasingly considered authenticity verification alongside conventional emotion classification. However, these are fundamentally different tasks with distinct objectives, datasets, annotation protocols, and evaluation strategies. Therefore, current evidence supports authenticity verification as a complementary research direction rather than a replacement for traditional FER. Potential cues include temporal facial behavior, multimodal consistency, physiological responses, and synthetic-media artifacts, although their effectiveness remains highly dependent on the acquisition conditions and benchmark datasets.

One of the major challenges in facial emotion recognition is determining whether the displayed facial expression genuinely reflects the underlying emotional state. Although humans can often perceive subtle inconsistencies between genuine and posed expressions, psychological studies suggest that human judgement is itself imperfect and strongly influenced by contextual, cultural, and individual differences. Consequently, developing robust computational models capable of reliably distinguishing between authentic and posed facial expressions remains an active research challenge. Identifying deceptive or manipulated emotional expressions has become increasingly important in digital communication, online social media, and intelligent human–computer interaction systems, where reliable emotion understanding directly influences user trust, decision making, and interaction quality.

Emotion-aware systems can better assist users by adapting their responses according to the perceived emotional state. Potential applications have been explored in healthcare, human–computer interaction, intelligent surveillance, media authentication, digital forensics, and cybersecurity. Nevertheless, these application domains involve consequential decisions and should not be interpreted as evidence that facial analysis can reliably determine deception, testimony credibility, or an individual's underlying mental state. Reliable deployment, therefore, requires rigorous domain-specific validation, informed consent, fairness assessment, privacy preservation, and meaningful human oversight before practical adoption [3].

1.1. Comparative Positioning and Gap Analysis

Although numerous studies have independently reviewed general facial emotion recognition, microexpression analysis, and deepfake detection, a comprehensive synthesis that systematically connects conventional facial emotion recognition with emotion authenticity detection remains mostly unavailable. Existing reviews generally emphasize either emotion classification or synthetic media detection without jointly examining datasets, annotation strategies, learning paradigms, evaluation protocols, deployment challenges, and future research directions. To clearly position the present work, Table 1 compares this review with representative recent surveys across several key analytical dimensions and highlights the unique scope and contributions of this review.

Table 1. Comparative positioning of this review against existing FER and affective computing surveys.

Survey Reference	Basic FER	Micro-Expressions	Posed vs. Genuine	Deepfake Veracity	Multimodal Fusion	PRISMA Scoring
Saleh et al. (2021) [4]	✓	×	Partial	×	×	×
Abdeldayem et al. (2025) [1]	✓	×	×	×	Partial	×
Ramanaharan et al. (2025) [5]	×	×	×	✓	Partial	✓
Yakut et al. (2026) [6]	Partial	×	×	Partial	✓	×
This Review	✓	✓	✓	✓	✓	✓

1.2. Conceptual Boundaries of Facial Veracity

The terminology used in this review is deliberately separated because the underlying tasks answer different scientific questions:

- **Basic emotion recognition:** identification of the emotion category conveyed by a facial expression, such as happiness, sadness, anger, fear, surprise, or disgust.
- **Emotion authenticity analysis:** estimation of whether a displayed expression is genuine/felt or posed/acted according to the ground-truth definition adopted by a dataset. This task should not be interpreted as a general-purpose lie detector.
- **Micro-expression recognition:** detection or classification of brief, subtle facial movements. Micro-expressions may provide temporal evidence relevant to authenticity analysis, but their presence alone does not establish that an expression is genuine or deceptive.
- **Deepfake/synthetic-media detection:** identification of media that has been digitally manipulated, generated, or synthesized. A synthetic face may display either a genuine-looking or posed expression, while an authentic recording may contain a posed expression; therefore, synthetic-media detection and emotion authenticity are related but distinct problems.

Accordingly, the performance values for these tasks are not pooled or directly compared unless the task, dataset, metric, and evaluation protocol are sufficiently aligned.

1.3. Major Contributions of This Review

The primary contributions of this review are summarized as follows:

- We present a comprehensive comparative analysis of existing surveys on facial emotion recognition, emotion authenticity detection, micro-expression analysis, and deepfake detection to clearly establish the novelty and scope of this review.
- We propose a unified taxonomy covering publicly available datasets, feature representations, handcrafted descriptors, deep learning architectures, multimodal learning strategies, and application domains for emotion authenticity analysis.

- We critically synthesize and demonstrates how the present review extends beyond existing surveys through a critical synthesis of datasets, methodological advances, evaluation protocols, and future research challenges.
- We systematically identify the major challenges affecting practical deployment, including dataset bias, limited cross-dataset generalization, micro-expression recognition, computational complexity, explainability, ethical considerations, fairness, and privacy.
- We outline future research directions emphasizing multimodal affective computing, trustworthy artificial intelligence, explainable learning, cross-domain generalization, and robust emotion authenticity verification for next-generation intelligent systems.

2. Literature Review for Background

Several key publications have shaped the evolution of facial emotion analysis, beginning with conventional facial emotion recognition and extending to recent advances in the analysis of emotion authenticity. This section provides the necessary background by reviewing the evolution of FER, facial feature representations, processing pipelines, artificial intelligence algorithms, publicly available datasets, and recent transformer- and multimodal-based architectures used in FER. Rather than merely summarizing previous studies, we critically discuss their strengths, limitations, and remaining research gaps that motivate the need for a comprehensive emotion authenticity analysis.

2.1. Basic and Authentic Emotion Detection by FER Systems

The field of Facial Emotion Recognition has rapidly evolved from a narrow Computer Vision task to a prevalent area of research within Artificial Intelligence. Classical literature focused almost exclusively on identifying the six basic emotions: happiness, sadness, anger, fear, surprise, and disgust [7]. Contemporary work has expanded upon this work. Facial expressions are now commonly classified with respect to authenticity; that is, whether they are genuine or deceptive [8].

Emotion authenticity analysis has potential applications in intelligent surveillance, healthcare, human–computer interaction, media authentication, and cybersecurity [3]. However, these applications should not be interpreted as evidence that facial analysis can reliably determine deception, credibility, or an individual's underlying mental state. Reliable deployment requires rigorous domain-specific validation, fairness assessment, and meaningful human oversight.

Despite advances in deep learning, modern FER systems still rely heavily on handcrafted solutions. FER frameworks can capture minute changes in facial topology. However, the detection of authentic emotions presents a unique challenge. When an individual poses a facial expression, it is typically exaggerated and deliberate. Meanwhile, genuine emotions are produced by subconscious muscle movements that are difficult to replicate.

Several experimental studies have reported that Duchenne smiles, produced during genuine happiness, activate the *orbicularis oculi* muscles surrounding the eyes, whereas voluntarily posed smiles often lack this characteristic muscular activation [9]. Consequently, authenticity analysis requires FER systems to model not only the facial appearance, but also the subtle temporal dynamics and muscle activation patterns [4].

2.2. Background of Facial Features Used in Authenticity Analysis

Detecting emotional authenticity requires robust facial feature extraction. The literature classifies facial features into two primary categories: geometric and appearance-based descriptors [10].

Geometric features were used to quantify the relative spatial positioning of the facial structures. They establish the distances, angles, and motions between fiducial points, such as the extremities of the mouth, nose, and eyebrows. As a person experiences changes in emotion, these points move closer together or further apart to accommodate facial deformation.

This technique has been leveraged to identify inconsistent emotions. Genuine emotions tend to gradually increase in intensity and slowly dissipate. Conversely, when faces are purposefully made to display a particular emotion, transitions between emotions are typically abrupt [9]. As such, the relative timing and symmetry of facial feature movements serve as a cue for authenticity analysis.

Appearance-based features attempt to capture variations in facial texture. When facial muscles contract, they create wrinkles, furrows, and micro-expressions on the skin's surface. Textural analysis can be performed using computer vision techniques such as local binary patterns (LBP) and histograms of oriented gradients (HOG) [11,12].

For instance, crow's feet or wrinkles that form near the eyes are strong indicators of genuine joy. Other methods for processing appearance-based features exist, many of which cannot be captured by geometric facial landmarks alone. Facial topology can be easily mimicked by those attempting to deceive; however, texture-related facial cues are generally more difficult to reproduce voluntarily than coarse geometric facial movements, making them valuable complementary indicators for authenticity analysis [3].

2.3. Structural Taxonomy and Processing Pipeline

The field of affective computing is large and multifaceted. Fig. 1 illustrates a proposed taxonomy to structure existing and future work. This hierarchical structure is divided into four major components: Data Acquisition, Feature Extraction, Architectural Designs, and Contemporary Application.

Data Acquisition encompasses all modalities of affective computing, ranging from 2D facial images to 4D volatile scans of facial activity. Feature Extraction encapsulates both handcrafted and deep representation learning. Architectural Designs comprise both traditional machine learning algorithms and modern deep neural networks. Finally, Contemporary Application showcases the active industries that affective computing expands into. Representative application domains include healthcare, human–computer interaction, intelligent surveillance, education, media authentication, and digital forensics, illustrating the broad applicability of affective computing technologies.

It should also be noted that there is a critical distinction to be made within this taxonomy. Traditional FER classifies primary emotions into several discrete categories. The analysis of emotional authenticity detects the "Veracity Gap" between genuine and posed facial expressions.

Fig. 2 depicts an idealized workflow of FER systems. Earlier conventional FER systems primarily utilized geometric facial landmark detectors for their analyses. State-of-the-art systems implement deep representations learned using CNNs and transformer models.

However, there exists a major obstacle that FER systems must overcome. Authentic emotion detection requires the analysis of facial dynamics. Observing the onset, apex, and offset of a smile can provide additional temporal evidence for authenticity analysis; however, temporal observation alone cannot definitively establish whether modelingexpression is genuine. Contemporary frame-based approaches may therefore benefit from temporal modelling, subject-independent validation, and uncertainty-aware decision rules. Although remarkable progress has been achieved in FER, existing approaches continue to face several challenges, including limited cross-dataset generalization, demographic bias, insufficient interpretability, and the scarcity of large-scale datasets specifically annotated for emotion authenticity. These limitations motivate the methodological developments discussed throughout the remainder of this review.

2.4. AI Algorithms Applications for FER Systems

Facial emotion recognition (FER) leverages various AI algorithms with two dominant approaches being Deep Learning (DL) and Machine Learning (ML). DL frameworks excel at modeling static imagery inputs, whereas ML algorithms are preferred when working with dynamic video streams [13].

Convolutional Neural Networks (CNNs): CNN-based networks are considered state-of-the-art for FER tasks because these algorithms excel at extracting high-level geometric representations directly from pixel intensities without the need for manual feature extraction or domain expertise [14]. Recent studies show that using custom CNN architectures with auxiliary attention modules yields accuracies

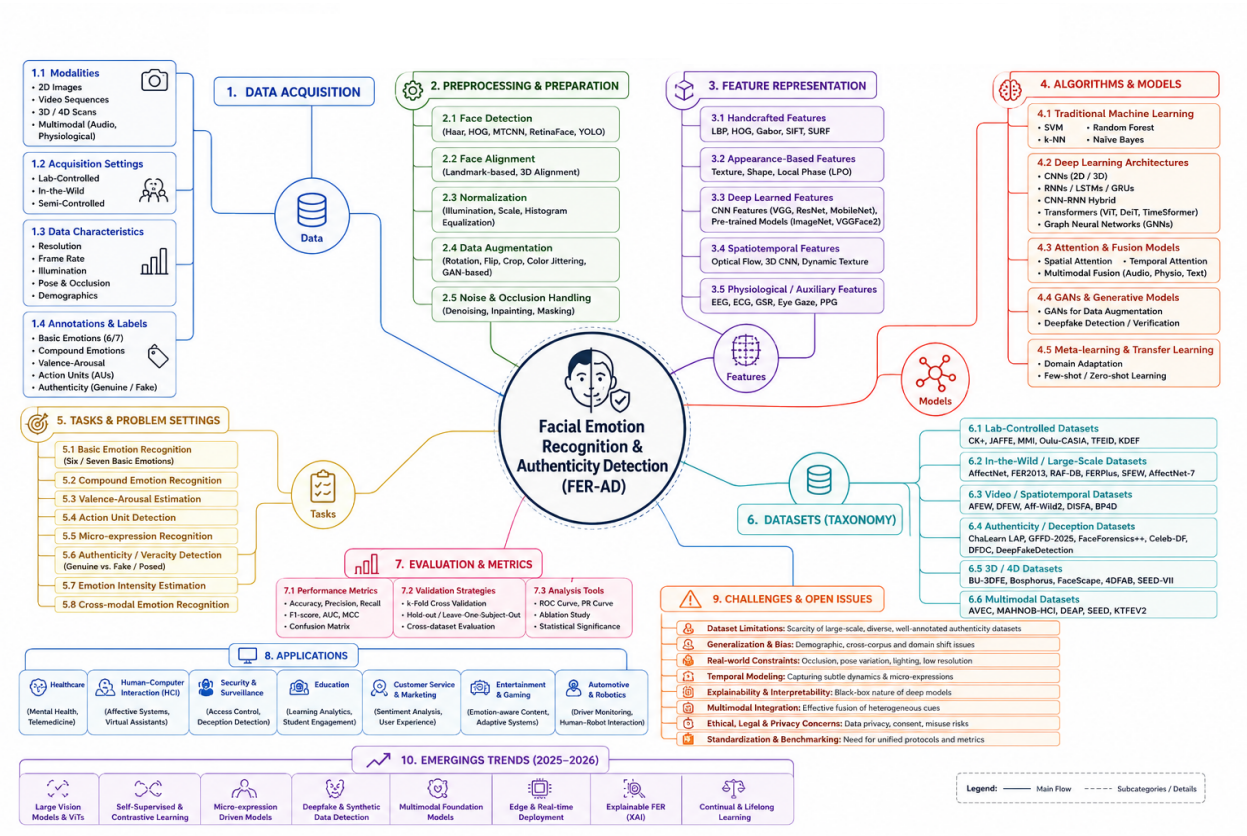


Figure 1. Facial Emotion Recognition and Authenticity Detection taxonomy.

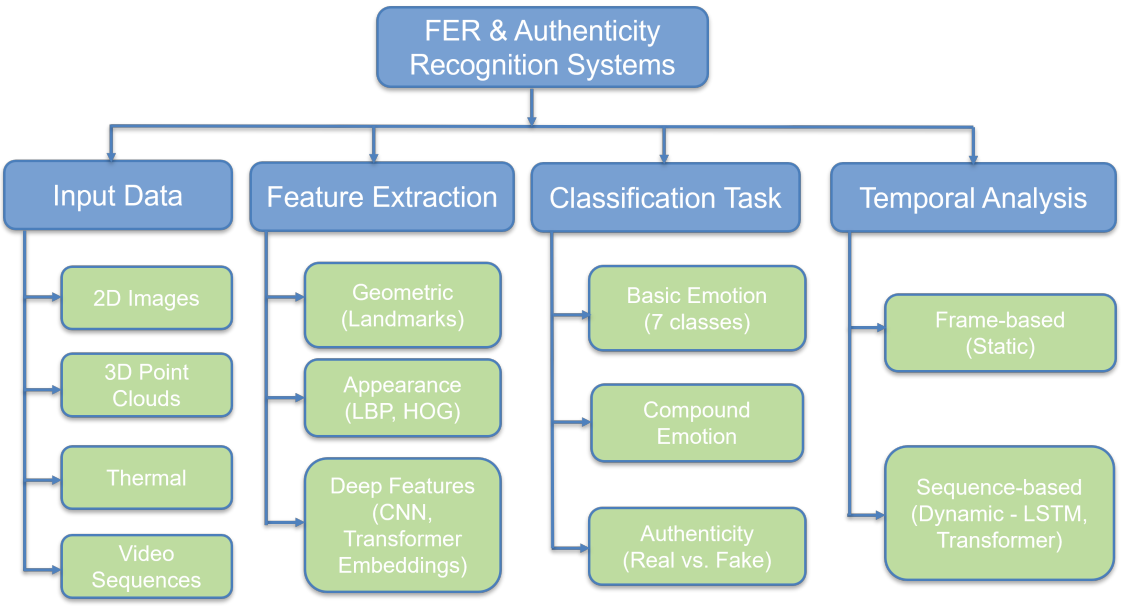


Figure 2. Standard processing pipeline for FER and Authenticity Recognition Systems.

above 90% on large-scale datasets for basic emotion classification tasks [15]. For example, modified CNN architectures have demonstrated success when independently masking ROIs such as eyes and mouth to capture subtle cues associated with deceptive micro-expression detection [16].

Recurrent Neural Networks (RNNs) & LSTMs: While CNNs remain powerful tools for spatially analyzing facial imagery, identifying whether a facial expression is being conveyed authentically often requires temporal processing. LSTM frameworks have shown promise for processing video streams by learning the temporal context surrounding an expression (onset-peak-offset) [17]. Hybrid networks that use CNN feature-extraction branches followed by LSTM sequencing layers were some of the first to introduce solutions that differentiate impulsive "real" reactions from voluntary, posed, "fake" facial movements [9].

Support Vector Machines (SVMs): Although the past decade has seen the rise of neural network-based architectures for solving almost every ML problem, SVMs paired with handcrafted feature extractors (i.e., Local Binary Patterns [LBP]) continue to be relevant. Recent works have demonstrated SVMs trained on manually extracted LBP histograms are effective at binary classification tasks such as "Real vs. Fake" smile detection and often outperform CNNs with far fewer trainable parameters [18, 19]. Although CNNs, recurrent networks, and traditional machine learning methods have significantly advanced FER research, their performance often deteriorates when evaluated across heterogeneous datasets owing to variations in illumination, pose, demographic diversity, and recording conditions. Recent research therefore increasinmodelinges on transformer-based architectures, graph neural networks, and multimodal learning frameworks capable of modelling long-range dependencies and complementary affective cues.

2.5. Distinguishing Emotion Authenticity from Deepfake Detection

A critical conceptual boundary is maintained between human expression authenticity and synthetic media detection. Emotion authenticity analysis determines whether a facial expression is genuine or posed according to the ground-truth protocol of the dataset. Deepfake detection determines whether the observed media have been digitally manipulated, generated, or synthesized. These tasks may share spatial, temporal, or multimodal features, but success in one does not guarantee success in the other.

For example, an authentic recording may contain a deliberately posed smile without any digital manipulation, whereas a synthetically generated video may depict an apparently genuine-looking emotional expression. Consequently, the two tasks should be reported separately unless a study explicitly defines and evaluates a joint objective that encompasses both tasks. This distinction is maintained throughout this review.

2.6. Database and Demographics in FER Systems

Although novel network architectures continue to emerge, training data remain the single most important factor influencing the performance of a facial emotion recognition system. Throughout the development of FER, research has transitioned datasets from small-scale controlled environments to large-scale unconstrained collections. This trend is due in part to the deep learning model's requirement for large quantities of training data and generalizability issues across varying demographics. As such, modern datasets range from 2D and 3D facial image sequences, multimodal sensor recordings (thermal, physiological), varying degrees of environmental restrictions (in-lab vs. in-the-wild), along with diverse subject groups (age, ethnicity, and gender).

Because selecting a dataset often defines the limits of a model's deployment context, we reviewed and categorized publicly available FER databases based on training objectives and underlying demographics.

2.6.1. Controlled and Laboratory-Based Benchmarks Benchmark datasets collected in laboratory settings have played a fundamental role in developing computer vision methods for facial expression analysis. These datasets typically allow researchers to isolate specific factors by enforcing front-lit, tightly cropped images with fixed pose instructions.

CK, CK+, JAFFE: The Cohn–Kanade (CK) dataset [7] and Extended Cohn–Kanade (CK+) [20] expression benchmarks are perhaps the most widely used datasets in FER-related literature. CK+ consists of 593 image sequences captured from 123 subjects (ages 18–50), whereas the original CK dataset is comprised of 327 video sequences from 123 subjects profiling the transition from neutral-to-peak facial expression with facial action unit (FAU) annotations. While CK+ expanded the number of demographic classes with higher age variability, the JAFFE dataset [21] was constructed to support demographic-specific modeling. JAFFE contains 213 grayscale images (256×256 resolution) of 10 Japanese females with each image rated by 60 human observers to verify semantic agreement.

Pose-Invariant: The Karolinska Directed Emotional Faces (KDEF) dataset (mentioned in [22]) provides RGB images of 70 Caucasian subjects (ages 20–30) across 5 viewing angles. All subjects were instructed to maintain strict visual consistency: no facial hair, glasses, or jewelry. Radboud Faces Database (RaFD) [23] expands on this concept by capturing 67 different models from five camera angles and three directions of gaze simultaneously to encourage pose variability.

Large-Scale "Controlled Variability": CAS-PEAL [24] was created with the intent to introduce large-scale variability under strict "Pose, Expression, Accessory, Lighting (PEAL)" settings. The benchmark consists of 99,594 training images of 1,040 subjects. Similarly, CMU Multi-PIE database used in [25] contains around 750,000 high-resolution images taken from 337 subjects across 15 viewpoints (45 Degree increments) under 19 different lightning configurations.

Controlled Evaluation Sets: FEI Face Database [26] contains 2,800 images of 200 subjects from Brazil. Caltech Faces dataset contains 450 images evenly distributed across 27 individuals; however, subjects were instructed to adopt different poses and "attitudes". Finally, The NIST Mugshot Identification Database is a face identification benchmark containing 3,248 images collected from 1,573 subjects. Though these datasets have been used to evaluate a variety of facial analysis algorithms, they typically enforce strong environmental controls.

2.6.2. In-the-Wild and Large-Scale Datasets Deep neural networks often require massive datasets for training. To accommodate deep learning, there has been increased traction towards using web-crawled "In-the-Wild" datasets. Images are typically extracted from the Internet or movies and contain naturally occurring background clutter, occlusions, spontaneous expressions, and uncontrollable illumination.

Deep Learning Forerunners: FER2013 [14] contains 35,887 grayscale facial images of dimensions 48×48. FER+ [27] was crowdsourced to reduce noisy-label concerns with FER2013 by distributing each sample to 10 unique Amazon Mechanical Turk annotators (thus each image contains 10 emotion labels). With the largest number of training images today, AffectNet [28] contains over one million facial images, out of which, 440k have been manually annotated with both categorical emotion labels and continuous valence–arousal values.

Evaluating FER systems under unconstrained "in-the-wild" settings introduces substantial performance degradation owing to non-uniform illumination, head pose variations, and background clutter, requiring specialized algorithmic optimizations and feature enhancement strategies to maintain robustness in real-world environments [1].

Robustness Benchmarks: RAF-DB [15] dataset includes 29,672 images captured from 217 different subjects and annotated for seven basic and compound emotions like "Excited," "Crying," etc. Images in this dataset cover a wide range of demographic, pose, and illumination variations. ExpW dataset [12] contains over 91k images with highly non-uniform backgrounds and variations in head pose. SFEW [29] has been widely adopted for studying "static" FER performance and consists of carefully extracted video frames from movie clips.

Embeddings, AU-Focused: EmotioNet dataset [30] contains more than one million images annotated with 11 emotion classes and 20 Action Units (AU). Unlike traditional classification labels, Google FEC [31] uses triplet-based learning to learn facial expression embeddings rather than class labels.

2.6.3. Dynamic and Temporal Datasets Static frames sampled from videos often lose fidelity when recognizing and analyzing facial movements. These benchmarks use sequential frames or entire video recordings as the sample data.

Movie Clips: AFEW dataset [29] contains a total of 600 videos collected from movies, whereas DFEW dataset [32] expanded this idea by using clips from "more than 1500 movies" for training-val. DFEW includes a total of 16,372 video clips and 160 subject demographic attributes.

Continuous Annotation: Aff-Wild2 dataset [33] is one of the largest facial video datasets containing 564 videos worth of about 2.8 million frames crawled from YouTube. Each frame has been frame-by-frame annotated for valence, arousal, basic expressions, and action units by expert annotators.

High-Quality Videos: MMI database used in [9] and Oulu-CASIA [34] capture facial expressions under controlled environments with respect to illumination and pose settings. Since both datasets are webcam recorded at 25 fps, researchers have commonly adopted them for developing static FER algorithms. Affectiva-MIT dataset, commonly known as AM-FED [35] dataset is similar to Oulu-CASIA; however, faces were recorded in-the-wild while watching commercials.

2.6.4. Micro-Expression and Authenticity Datasets Micro-expression recognition benchmarks focus on identifying involuntary facial movements that occur when a person attempts to conceal an emotion. Datasets for perceived authenticity (real vs. fake) also require high sampling rates for correct classification.

Micro-Expressions: CASME II dataset [36] is widely considered one of the gold standard benchmarks for micro-expression research and contains a total of 247 samples recorded at 200 fps. SMIC dataset [37] expands this even further with 164 video clips recorded at 100 fps. USTC-Emoji dataset employs similar forced-hand-sign elicitation protocols to record genuine micro-expressions from subjects.

Intensity Analysis: DISFA dataset [38] recorded videos of 27 actors performing facial actions with varying intensity levels. Action Unit intensities are labeled on a scale of 0 to 5 (peak level). Two datasets used to evaluate "Real vs Fake" emotion analyses are Chalearn dataset [39] and GFFD-2025 subset (25%) [40]. Both datasets contain sampled videos of individuals posing/spontaneously expressing emotions; however, they are typically utilized for applications in security.

2.6.5. Multimodal and SpecialModelingsets Contemporary FER approaches leverage auxiliary modalities, and research often focuses on the challenges associated with specific demographics or scenarios.

3D Modelling: BU-3DFE dataset [41] and BU-4DFE dataset [42] were two of the first 3D (temporal) expression databases available to researchers. Building off this, BP4D-Spontaneous dataset [43] provided high-resolution 2D & 3D spontaneous facial expressions while 4DFAB [44] provides a large-scale spontaneous 4D facial expression dataset with fine-grained geometric deformation annotations. IMPA–FACE3D database used in [45] allows researchers to explore facial animation modeling.

Physiological and Thermal: SEED-VII dataset [46], ASCERTAIN dataset [47], and MAHNOB dataset [48] all contain facial RGB images fused with EEG, ECG, GSR, or eye-tracking signals to enable multimodal affect verification tasks. KTFEv2 [49] is a large-scale facial thermal dataset that correlates chest-surface temperature changes to sympathetic emotional arousal.

To overcome the limitations of purely vision-based affect recognition, multimodal approaches increasingly incorporate physiological and neurophysiological signals into their models. For instance, heuristic machine learning models applied to electroencephalogram (EEG) signals provide direct measurements of internal emotional states that effectively complement external facial muscle observations [50].

Child, Group, and Context-Specific: The problem of group-level affect analysis is tackled by GAF 3.0 dataset [51] and EmotioW dataset [52]. Recently, researchers have started building emotion recognition datasets catered to children. EmoReact dataset [53] and LIRIS-CSE dataset [54] provide caregivers with tools to analyze facial expressions in kids. DAiSEE dataset [55] was created with the intent to understand student emotional states during e-learning courses. M-LFW-FER dataset [18,56] can be applied to detect

faces with masks for face spoofing prevention. GAPED dataset [57] provides stimuli videos to induce emotions from actors rather than footage used to "train" FER systems.

As shown in Table 2, many controlled/laboratory-based datasets have been adopted throughout the early stages of computer vision research. Tables 3, 4, 5, and 6 compare modern datasets intended for in-the-wild images, videos, micro-expressions, and multimodal training, respectively.

Table 2. Controlled / Lab-based FER Datasets

Dataset	Emotions / Content	Size	Type / Subj.	Notes
CK [7]	6 Basic Emotions	327 seq	Adults / 123	Emotion progression sequences.
CK+ [20]	7 Emotions + Neutral	593 seq	Adults / 123	30 FPS, 640 × 490 resolution.
JAFFE [21]	6 Basic + Neutral	213 imgs	Adults / 10	Japanese female subjects only.
KDEF [22]	7 Emotions	4,900 imgs	Adults / 70	5 viewing angles, color images.
RaFD [23]	8 Emotions + Neutral	8,040 imgs	Mixed / 67	Multi-view and gaze directions.
CAS-PEAL [24]	PEAL variations	99,594 imgs	Adults / 1040	Large-scale Chinese dataset.
FEI Face [26]	Neutral, Smile	2,800 imgs	Adults / 200	Pose variations up to 180°.
CMU Multi-PIE [25]	Poses, lighting	750k imgs	Adults / 337	Controlled acquisition conditions.
Caltech Faces [25]	Faces with attitudes	450 imgs	Adults / 27	Attitude variations.
NIST Mugshot [25]	Frontal/profile	3,248 imgs	Adults / 1573	Metadata included.

Table 3. In-the-Wild and Large-Scale Datasets

Dataset / Ref	Emotions / Content	Size	Type / Subj.	Notes
FER2013 [14]	7 Emotions	35,887 imgs	Adults / N/A	48 × 48 grayscale images.
FER+ [27]	7 Emotions + Contempt	35,887 imgs	Adults / N/A	Crowdsourced re-annotation.
AffectNet [28]	8 Emotions	∼1M imgs	Adults / N/A	Largest in-the-wild dataset.
RAF-DB [15]	6 Basic + 12 Compound	30,000 imgs	Adults / 217	Real-world affective faces.
ExpW [12]	7 Basic Emotions	91,793 imgs	Adults / N/A	High pose/lighting variance.
EmotioNet [30]	11 Emotions + AUs	∼1M imgs	Adults / N/A	Large-scale AU annotations.
SFEW [29]	7 Basic Emotions	1,766 imgs	Adults / 68	Static movie frames.
Google FEC [31]	Similarity triplets	∼87k imgs	Adults / N/A	Embedding-based learning.

Table 4. Dynamic and Temporal Datasets (Video)

Dataset	Emotions / Content	Size	Type / Subj.	Notes
AFEW [29]	Video expressions	600 videos	Adults / N/A	Valence and arousal annotations.
DFEW [32]	7 Basic Emotions	16,372 clips	Adults / 3600	Extracted from 1500+ movies.
Aff-Wild2 [33]	Valence, Arousal, AUs	564 videos	Adults / 554	Large-scale audiovisual dataset.
MMI [9]	Expressions + AUs	2,900+ vids	Adults / 75	Onset-peak-offset phases.
Oulu-CASIA [34]	6 Basic Emotions	2,880 clips	Adults / 80	Visible and NIR recordings.
AM-FED [35]	Smiles, gestures, AUs	242 videos	Adults / N/A	Spontaneous webcam videos.

Overall, existing FER datasets exhibit substantial diversity in terms of acquisition protocols, annotation strategies, demographic composition, and environmental conditions. Laboratory datasets provide carefully controlled recordings but often suffer from limited demographic diversity and ecological validity, whereas in-the-wild datasets improve generalization at the expense of increased annotation ambiguity and environmental variability. Furthermore, publicly available datasets specifically designed for emotion authenticity remain comparatively limited in scale, diversity, and annotation consistency, highlighting an important research gap for future benchmark developments.

Table 5. Micro-Expression and Authenticity Datasets

Dataset / Ref	Emotions / Content	Size	Type / Subj.	Notes
CASME II [36]	Micro-expressions	247 samples	Adults / 26	200 FPS high-speed capture.
SMIC [37]	Pos, Neg, Surprise	164 clips	Adults / 16	100 FPS micro-expressions.
USTC-NNE [37]	Micro-expressions	100+ samples	Adults / 30	Spontaneous high-speed data.
DISFA [38]	AU intensity (0–5)	27 videos	Adults / 27	Spontaneous AU annotations.
ChaLearn LAP [39]	Real vs Fake	600 videos	Adults / 50	100 FPS genuine vs acted.
GFFD-2025 [40]	Genuine vs Fake	2,224 imgs	Adults / N/A	Authenticity-focused dataset.

Table 6. Multimodal and Specialized Datasets

Dataset / Ref	Emotions / Content	Size	Type / Subj.	Notes
BU-3DFE [41]	6 Basic (3D)	2,500 models	Adults / 100	Static 3D facial expressions.
BU-4DFE [42]	6 Basic (4D)	606 seq	Adults / 101	3D + Time (Dynamic).
BP4D [43]	Spontaneous (3D)	328 videos	Adults / 41	High-res 2D + 3D spontaneous.
4DFAB [44]	Spontaneous/Posed	$>$ 100k meshes	Adults / 180	Longitudinal 4D dataset.
SEED-VII [46]	Multimodal (6 Emo)	N/A	Adults / N/A	EEG + Eye-tracking + Face.
ASCERTAIN [47]	Arousal, Valence	2,088 imgs	Adults / 58	EEG, ECG, GSR + Face.
KTFEv2 [49]	Thermal	$\sim$14k imgs	Adults / 30	Infrared/Thermal imaging.
GAF 3.0 [51]	Group Emotion	Groups	N/A	Group-level affect analysis.
DAiSEE [55]	Engagement	9,068 clips	Adults / 112	E-learning engagement.
EmoReact [53]	Child states	1,102 clips	Children	Child-specific dataset.

2.7. The Shift from CNNs to Vision Transformers (2022–2026)

Although CNNs have been used as the standard for facial-feature extraction owing to their built-in assumption of locality and translational invariance [14], recent studies have shown that these models lack the capacity to learn long-range dependencies. Vision Transformers (ViTs), originally introduced for computer vision by adapting the transformer architecture to image analysis, have increasingly bemodelingd for facial emotion recognition owing to their ability to capture long-range spatial dependencies [28]. Compared with conventional CNNs, transformer-based architectures provide improved modelling of global contextual relationships, making them particularly attractive for analyzing subtle facial behaviors associated with authenticity verification.

Instead of slowly aggregating global information through hierarchical layers with exponentially growing receptive fields, transformers apply self-attention to model global dependencies from the first layer onward. This behavior is crucial for modeling micro-expressions, where subtle perioral muscle movement can be semantically correlated with a micro-expression occurring at the corner of a person's mouth [6] at the same time, which is considerably more difficult to capture using only local convolution kernels.

Benchmarking on AffectNet [28] and RAF-DB [15] shows that ViT-based models, such as dual-branch attention networks, outperform ResNet-50 baselines by 3–5% when classifying compound emotions. Furthermore, recent works have shown that hybrid CNN-ViT models are state-of-the-art for authenticity detection. Hybrid approaches exploit convolutional layers to extract low-level textural features (e.g., skin wrinkles) from local regions and then send learned embeddings through a transformer model to reason about global semantic consistency between facial areas (eyes vs. mouth) to detect common spatial artifacts seen in deepfakes [58].

2.8. Recent Advances in Multimodal and Transformer-based Architectures (2025–2026)

Multiple studies published in 2025 and 2026 demonstrate a trend towards multimodal CNNs and hybrid transformer architectures. Previous works generally adopted unimodal architectures searching for spatial artifacts in facial regions. Contemporary deepfake and authenticity systems leverage inconsistencies found in audio and visual modalities.

Yu et al. [59] introduce ACE-Net, a fine-grained deepfake detection framework utilizing emotional consistency analysis in both audio and visual modalities. Their approach aggregates acoustic embeddings with visual keyframes using a bidirectional cross-attention module. By jointly modeling face-video and speech pairs, ACE-Net learned subtle offsets between audio and facial motion, resulting in an Area Under Curve (AUC) score of 0.921 on benchmark datasets.

Expanding on this concept, Yan et al. [58] proposed a comprehensive multimodal deepfake detection framework with additional liveness verification. This framework jointly processes facial footage with speech using a CNN-based encoder for spatial representation learning and a temporal-conformer module for modeling inter-frame variations. An adaptive mid-late fusion strategy maintains performance robustness when one modality is corrupted. The experimental results on FaceForensics++ showed an accuracy of 97.3%.

Finally, Yakut et al. [6] highlighted the importance of transformer-based architectures for multimodal emotion analysis in multilingual video contexts. Their findings demonstrated that traditional CNN models, such as ResNet-50, fail to model higher-order cognitive and affective human states. Instead, transformer-based neural networks, such as DDAMFN and Vision Transformers (ViT), maintain affective representations distributed across time, permitting better model interpretability and stability in cross-language comparisons. Supporting this claim, Bhad et al. [60] demonstrated that multimodal audio-visual fusion plays an important role in deepfake detection reliability under real-world distortions.

Table 7. Performance Comparison: CNN vs. Transformer & Hybrid Architectures (2019–2026)

Architecture Type (Ref)	Model / Method	Dataset	Performance
Traditional CNN Approaches (2019–2024)			
CNN (Static) [11]	Hierarchical CNN	CK+ (Lab)	96.50% (Acc)
CNN (Static) [14]	Standard CNN	CK+ (Lab)	94.50% (Acc)
CNN (In-the-Wild) [28]	ResNet-50	AffectNet	≈60.0% (Acc)
CNN (Attention) [15]	RAN Attention	RAF-DB	86.90% (Acc)
Transformer & Hybrid Approaches (2025–2026)			
Multimodal Hybrid [59]	ACE-Net (Cross-Modal)	DFDC	0.921 (AUC)
CNN + Conformer [58]	Temporal Conformer	FaceForensics++	97.30% (Acc)
Vision Transformer [6]	ViT + DDAMFN	Multilingual Video	Competitive
Explainable Hybrid [61]	Multiple CNNs + XAI ttowardsfake Wild	89.0%+ (Acc)	

Overall, the reviewed literature demonstrates a clear progression from handcrafted feature engineering and conventional machine learning toward deep representation learning, transformer-based architectures, graph neural networks, and multimodal affective computing. Despite remarkable improvements in benchmark performance, current approaches continue to face challenges related to dataset bias, limited cross-dataset generalization, insufficient authenticity-oriented datasets, computational complexity,

explainability, and ethical deployment. These unresolved challenges motivate the comprehensive analysis of datasets, methodologies, evaluation protocols, and future research directions presented in the subsequent sections.

3. Methodology

This review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines. The adopted methodology provides a transparent and reproducible framework for identifying, screening, and synthesizing studies on facial emotion recognition (FER) and emotion authenticity (genuine-versus-posed) analysis.

3.1. Study Design

A systematic literature review was conducted to examine computational approaches for facial emotion recognition and the discrimination of genuine versus posed facial expressions. The review methodology was designed to identify state-of-the-art algorithms, benchmark datasets, evaluation protocols, and emerging research trends relevant to emotion authenticity analysis. The overall scope of the review was defined according to the research objectives presented in Section 1.

Component	Description
Research Question	How can computational models distinguish between genuine and fake facial emotions, and what are the current limitations in datasets and methodologies?
Population / Datasets	Datasets containing facial expressions (images and videos), including standard FER datasets (e.g., CK+, FER2013) and authenticity-focused datasets (e.g., ChaLearn LAP, GFFD-2025).
Intervention / Methods	Facial emotion recognition systems using machine learning and deep learning techniques, such as CNNs, RNNs, SVMs, and transformer-based models.
Comparison	Comparison of genuine (felt) facial expressions with posed or acted (fake) expressions.
Outcome / Metrics	Accuracy, classification performance, and the ability to reliably distinguish authentic expressions from deceptive ones.

3.2. Research Questions

This systematic review was designed to answer one primary research question together with five complementary research questions that collectively investigate the current state of facial emotion recognition (FER) and emotion authenticity detection. These questions provide the analytical framework used throughout the review to organize the literature, compare methodologies, identify research gaps, and suggest future research directions.

3.2.1. Primary Research Question **RQ:** What computational methods have been proposed for facial emotion recognition and emotion authenticity detection, and how have these methods evolved toward distinguishing genuine, posed, and manipulated facial expressions?

3.2.2. Secondary Research Questions

- **RQ1:** What machine learning and deep learning approaches have been proposed for facial emotion recognition and emotion authenticity detection, and what are their principal advantages and limitations?
- **RQ2:** Which publicly available datasets are most frequently used for FER, micro-expression recognition, genuine-versus-posed expression analysis, and deepfake detection, and what are their key characteristics and limitations?
- **RQ3:** How do existing methods evaluate emotion authenticity, and which performance metrics, validation protocols, and benchmark datasets are commonly employed?
- **RQ4:** What are the major technical challenges affecting the reliability, robustness, explainability, fairness, and cross-dataset generalization of current FER and emotion authenticity detection systems?
- **RQ5:** What emerging research directions, including transformer-based architectures, multimodal learning, explainable artificial intelligence, and trustworthy AI, are expected to advance future FER and emotion authenticity analysis?

Collectively, these research questions guided the qualitative synthesis presented in this review by examining algorithmic developments, benchmark datasets, evaluation methodologies, practical deployment challenges, and emerging research trends in facial emotion recognition and emotion authenticity detection.

The research questions collectively investigate four complementary aspects of the literature, namely algorithmic development, dataset characteristics, emotion authenticity analysis, and current research challenges. Together, they provide the analytical framework used throughout the subsequent sections of this review.

3.3. Search Strategy and PRISMA Protocol

A systematic literature search was conducted across IEEE Xplore, ScienceDirect, SpringerLink, and arXiv. The revised search window covered studies published between January 1, 2015, and August 31, 2026. The search strategy was designed to identify literature addressing facial emotion recognition, genuine-versus-posed expression analysis, temporal facial behavior, and related computational approaches. Additional databases such as the ACM Digital Library, Web of Science, and PubMed are relevant to this topic; however, they were not part of the database set used to generate the reported 945-record screening flow. Therefore, they are not represented as searched sources in the present PRISMA counts.

The search strategy combined keywords related to facial emotion recognition, emotion authenticity, deep learning, and computer vision. The primary Boolean search query was:

```
("Facial Emotion Recognition" OR FER) AND ("Emotion Authenticity" OR
"Genuine Emotion" OR "Posed Emotion" AND ("Deep Learning" OR CNN OR
Transformer OR Vision Transformer OR LSTM OR Graph Neural Network)
```

Database-specific adaptations of the query were used when required by the search interface. The search date and database set are reported explicitly to ensure that the numerical study selection flow remains reproducible. Two reviewers participated in the title/abstract and full-text screenings, and disagreements were resolved by discussion and consensus. Because the underlying individual screening decisions are not reproduced in the current manuscript, an inter-rater coefficient is not reported here; the previously stated value of $\kappa = 0.86$ has therefore been removed rather than being treated as an independently verifiable statistic.

The reference lists of the selected studies were inspected to identify relevant publications that were not retrieved through the primary database search.

3.4. Inclusion and Exclusion Criteria

The inclusion criteria were operationalized as follows:

- Studies published within the stated search period and available in English.
- Studies investigating facial-expression analysis using image, video, or multimodal facial data and reporting quantitative experimental results.
- Studies addressing basic/compound emotion recognition, temporal facial behaviour, or explicit genuine-versus-posed/authenticity analysis, provided that the relevance of the result to the review questions could be established.
- Studies using a publicly described dataset or sufficiently detailed data source to permit interpretation of the reported experiment.

The exclusion criteria were:

- Studies without an evaluable quantitative result or without sufficient methodological information.
- Studies based exclusively on non-visual modalities when no facial information was analysed.
- Duplicate records, abstracts without sufficient methodological information, and studies outside the defined scope.
- Studies whose conclusions could not be linked to a defined facial-analysis task or dataset.
- Studies lacking sufficient methodological detail to reproduce or interpret the reported results.

The exclusion of non-English literature may introduce language bias and is therefore treated as a limitation rather than as evidence that the reviewed literature represents the complete global research output.

3.5. Study Selection and Data Extraction

Screening and selection of relevant literature was performed using reference management software. All citations found during electronic database searches were imported into our library and duplicates removed based on title, authors, and year of publication. Two reviewers screened titles and abstracts of unique records according to the inclusion and exclusion criteria outlined above. Any disagreements about eligibility were resolved through discussion. Eligible articles were retrieved by the reviewers and full text analyses were performed to ensure relevance to our research questions with attention to genuine vs. posed facial expressions. Data was then extracted by reviewers using a standardized form. Any disagreements were resolved by consensus. The variables extracted and presented below were:

- **Bibliometric Data**: Authors, Year, Country of Origin.
- **Dataset:**
  - Name of dataset(s) used (CK+, RAF-DB, Chalearn, etc.)
  - Total Sample Size (Number of Images/Video Sequences).
  - Type of Acquisition (Lab-controlled or In-the-wild).
  - Age, Gender and Ethnicity (if provided).
- **Preprocessing**: Face Alignment, Normalization, Data Augmentation, etc.
- **Features**: Geometric (Landmarks), Appearance (Texture), Learned (Deep Features).
- **Algorithm/ Model**: CNN, LSTM, SVM, Hybrid Models, etc.
- **Parameters**: Learning Rate, Batch Size, Number of Epochs, Validation split, etc.
- **Performance:**
  - Overall Accuracy.
  - Accuracy by Basic Emotion (Happy, Sad, etc.)
  - Accuracy for Authenticity Detection (Genuine vs. Fake classifications)

3.6. Quality Assessment Score Card

A five-item quality-assessment framework was defined to evaluate transparency and methodological reporting across the included studies. The criteria cover dataset transparency, methodological detail, validation design, metric reporting, and baseline comparison. The framework is used as a structured risk-of-bias aid rather than as a claim that a numerical score alone establishes study validity.

1. **QA1 (Dataset Transparency):** Explicit description of sample characteristics, acquisition conditions, and subject demographics where available (1.0 pt).
2. **QA2 (Methodological Rigor):** Sufficient detail on preprocessing, architecture, training procedure, and relevant hyperparameters (1.0 pt).
3. **QA3 (Validation Protocol):** Explicit description of the train/test or cross-validation protocol, including subject independence where applicable (1.0 pt).
4. **QA4 (Metric Reporting):** Reporting of appropriate evaluation metrics for the stated task, such as accuracy, F1-score, AUC, CCC, or confusion matrices (1.0 pt).
5. **QA5 (Baseline Comparison):** Comparison with relevant baselines or established methods where appropriate (1.0 pt).

Each criterion received a score of either 1 (criterion satisfied) or 0 (criterion not satisfied), resulting in a maximum possible quality score of five for each study.

Because the study-level scoring sheet was not included in the current manuscript, the previous aggregate counts of 26 high-quality, 12 moderate-quality, and 5 low-quality studies were removed. These counts should only be restored after the completed extraction sheet has been independently checked and made available as supplementary material. This change prevents a quantitative quality synthesis from being presented without an auditable record.

3.7. Rationale on Quantitative Meta-Analysis Heterogeneity

A quantitative meta-analysis was not performed because the included studies differ substantially in task definition, dataset composition, class balance, subject split, training protocol, and evaluation metric. In particular, seven-class emotion recognition, binary genuine-versus-posed classification, micro-expression recognition, and synthetic-media detection are not statistically interchangeable outcomes. Accuracy values are therefore retained as study-level observations rather than pooled into a single effect estimate. The synthesis instead groups evidence by task, dataset setting, modality, and evaluation protocol. This approach also avoids interpreting a high accuracy obtained under a controlled laboratory protocol as direct evidence of generalization to an unconstrained authenticity-detection setting.

3.8. Mathematical Foundations of Attention Mechanisms

Vision Transformers achieve state-of-the-art results by leveraging a mechanism called Scaled Dot-Product Attention. Given an input image X of facial patches, the model maps X to three parameterized matrices: Queries (Q), Keys (K), and Values (V). Each attention head then produces an output based on a weighted summation as shown in Equation 1:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k is the dimensionality of the keys. The scaling factor in the denominator is used to avoid extremely large dot products that push the softmax function into regions where it has extremely small gradients. This can be problematic because the attention weights are difficult to learn.

In FER, self-attention enables the model to capture long-range dependencies among spatially separated facial regions. For example, interactions between the eye region and mouth region can be jointly modeled, allowing the network to learn coordinated facial musclMulti-headons that are informative for expression analysis. Such global contextual modeling is particularly beneficial for subtle facial behaviors, including micro-expressmulti-headtemporal facial consistency, and multi-headattention weights themselves should not be interpreted as direct evidence of emotion authenticity.

MUlti-head Attention (MHA) allows the model to attend to information at different positions from different representation subspaces. This can be seen in Equation 2:

$$\mathrm{MHA}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)W^O \tag{2}$$

where $head_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$. Because each attention head learns complementary feature representations, multi-head attention enables the network to model diverse spatial relationships and global context dependencies. This capability has been successfully exploited in recent transformer-based FER, micro-expression analysis, and deepfake detection frameworks [59].

4. Results

This section describes the results found at the time of conducting this review. The search results from the study identification process are presented first. Thereafter, publication and global contribution trends are analyzed. The usage of benchmark datasets and their effect on reported performance metrics are reviewed next. This is followed by a discussion on the success of various feature extraction methods as well as learning strategies. Lastly, the capability of existing FER algorithms to differentiate authentic versus posed expressions is highlighted.

4.1. Search Results

A total of 945 records were retrieved from the four sources used for the reported screening flow: IEEE Xplore, ScienceDirect, SpringerLink, and ArXiv. After the removal of 215 duplicates, 730 records remained for title and abstract screening. Of these, 642 records were excluded at the screening stage, leaving 88 articles for full-text assessment. Forty-five full-text articles were subsequently excluded, resulting in 43 studies being included in the final review. These numbers are internally consistent ($945 - 215 = 730$, $730 - 642 = 88$, and $88 - 45 = 43$) and were retained as the basis of the reported PRISMA flow.

The full-text exclusion categories are reported in Table 8. The categories must sum to 45; therefore, the table is treated as the authoritative numerical summary, and the accompanying narrative has been aligned with it. The review does not interpret the 43 included studies as a homogeneous experimental cohort because the studies address different tasks and use different datasets and evaluation protocols.

Table 8. Summary of Excluded Full-Text Papers

Exclusion Reason	**Count (n)**	**Description**
No quantitative performance metrics reported	18	Studies that discussed theoretical concepts without reporting accuracy, F1-scores, confusion matrices, or other quantitative evaluation measures.
Use of non-standard or inaccessible datasets	12	Studies that relied on private or unavailable datasets, limiting reproducibility and comparative analysis.
Lack of authenticity analysis	10	Studies that focused only on posed expressions and did not investigate the distinction between genuine and deceptive emotions.
Reliance on non-visual modalities	5	Studies using only audio signals or physiological measurements without facial images or video analysis.
Total Excluded	**45**	

In total, 43 articles were included in this review. The outcomes of these studies were used to construct reviews related to dataset characteristics, reported algorithmic performance, and methods for authenticity detection. Fig. 3 depicts the entire selection process.

4.2. Bibliometric Trends and Global Contributions

Following the selection of the final studies for review, bibliometric trends were assessed to determine the progression of research in FER. Several key takeaways can be highlighted from these results. The temporal trends are shown in Fig. 4. Research interest peaked in 2019, with eight pertinent publications.

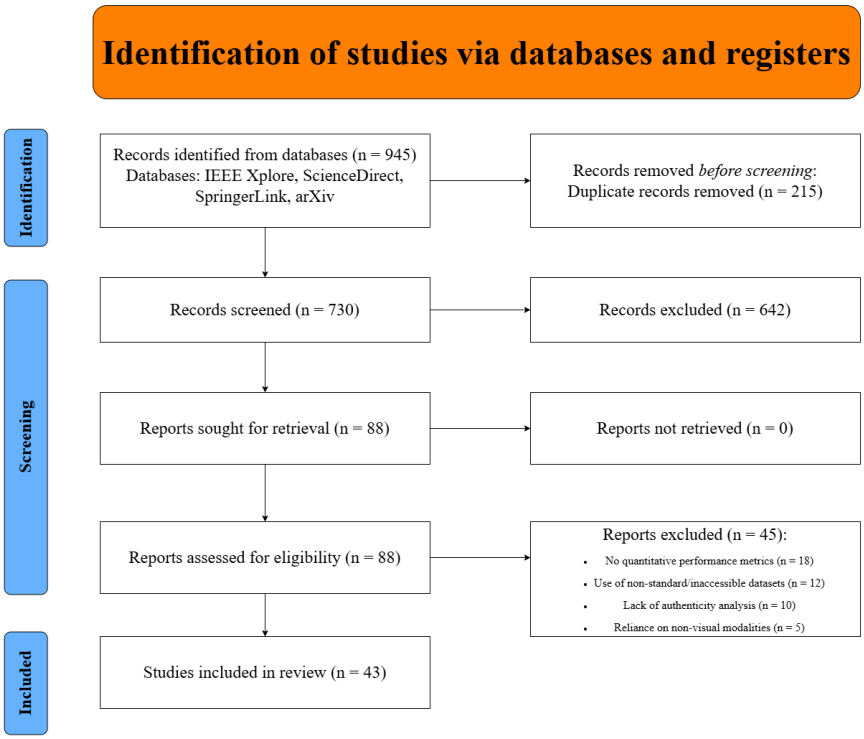


Figure 3. Flow diagram describing the study selection process.

Note that this uptick in publications coincided with the release of several large-scale datasets, including AffectNet [62].

A second increase in publications is visible in the reviewed sample during 2025–2026. Because the present review is based on a finite, search-dependent corpus, this temporal pattern should be interpreted descriptively rather than as proof of a causal relationship. The observed increase may coincide with growing interest in synthetic media, multimodal learning, and authenticity verification; however, the contribution of each factor cannot be established from publication counts alone.

It is also worth noting where the majority of past research has been conducted. The country of origin for each study is shown in Fig. 5. India produced the largest number of publications from the reviewed literature (15 articles). The USA (11 articles) and China (9 articles) followed. India contributed the largest number of publications among the reviewed studies (15 articles), followed by the USA (11 articles) and China (9 articles). These findings describe the geographic distribution of the reviewed literature only and should not be interpreted as evidence of overall global research activity or the underlying causes of publication trends.

4.3. The Validity Gap in Benchmark Datasets

The reviewed studies reveal a clear difference between the performance reported on controlled laboratory datasets and that obtained on more unconstrained real-world datasets. As shown in Fig. 6, CK+ was the most frequently used dataset in the selected studies, appearing in 12 articles. Its popularity is understandable because its controlled acquisition conditions make it suitable for reproducible experiments and model comparisons. The extracted results also show high performance on CK+, with a mean reported

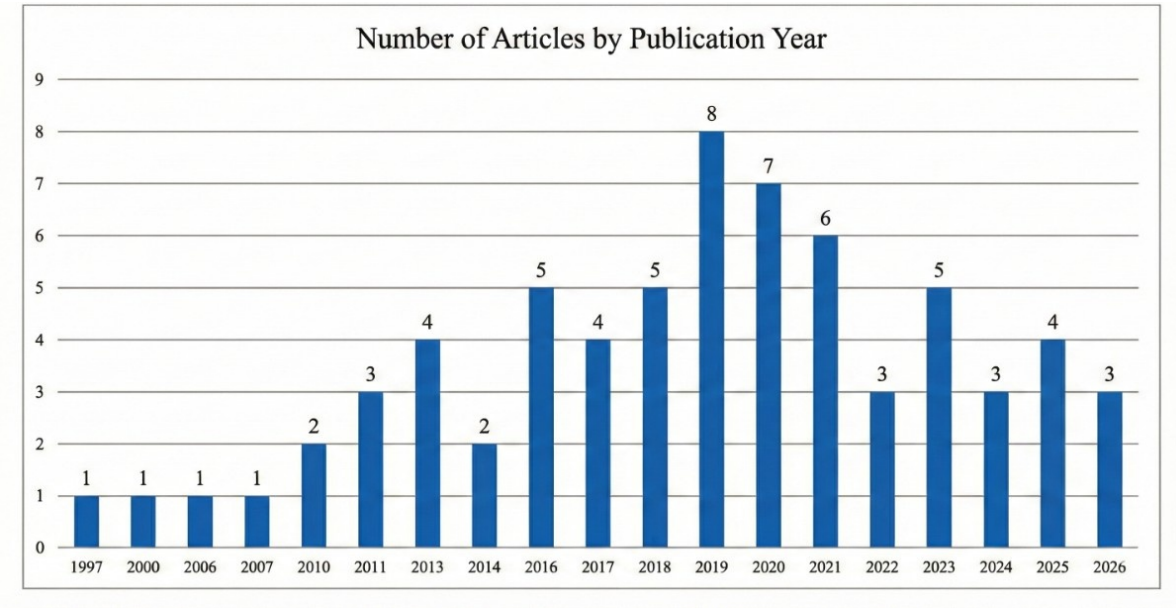


Figure 4. Distribution of reviewed articles by publication year.

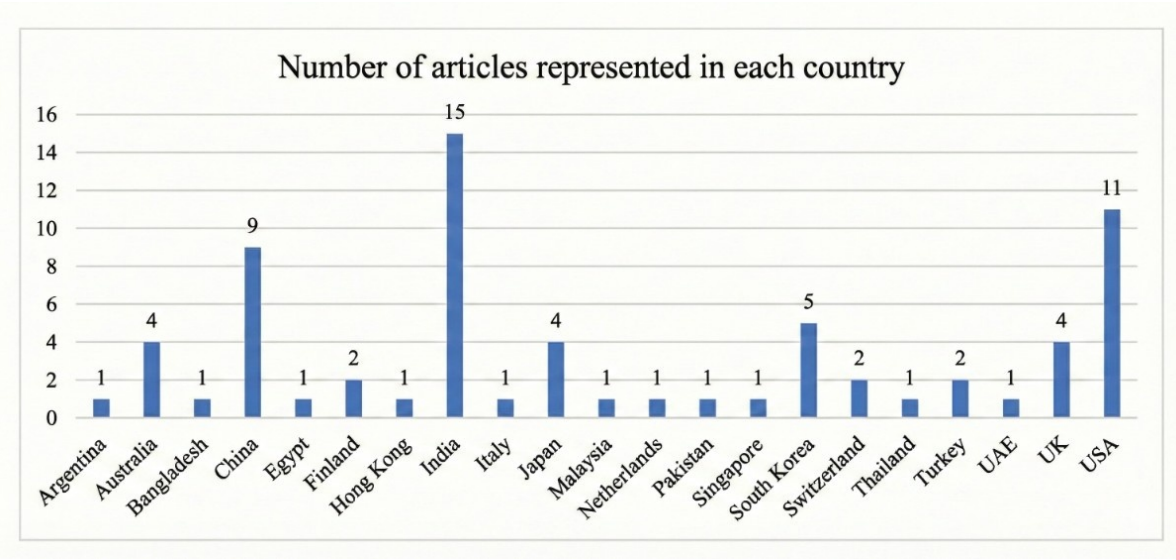


Figure 5. Geographical distribution of the primary authors in the reviewed literature

accuracy of approximately 95.3% (Fig. 7). However, this value is only a descriptive summary of the reviewed results and should not be interpreted as a pooled effect size, since the studies used different models, data splits, preprocessing methods, and evaluation protocols.

A different pattern can be observed for unconstrained datasets such as AffectNet, where the extracted mean performance is approximately 63.5%, as summarized in Fig. 7 and Table 9. This difference should not be treated as a universal 31.8-percentage-point performance drop because dataset characteristics and experimental settings vary considerably across studies. Nevertheless, the results consistently suggest that performance tends to decrease when models are evaluated under more variable conditions.

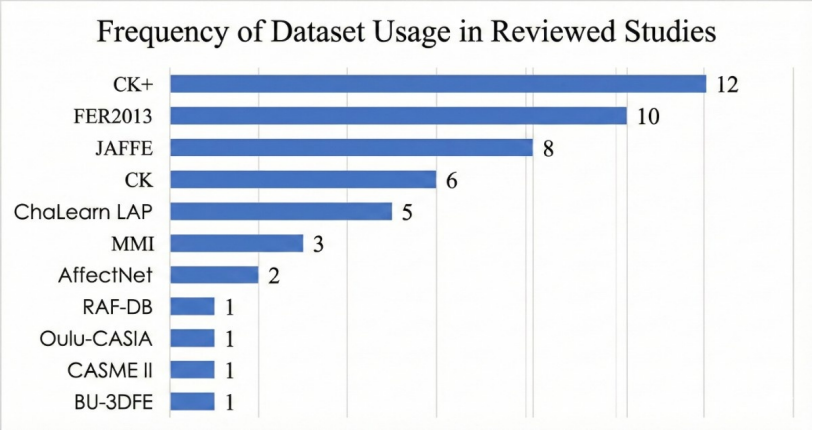


Figure 6. Frequency of each dataset included in the review.

This difference highlights what we refer to as a validity gap. Controlled datasets reduce variations in pose, illumination, background, and recording conditions, whereas in-the-wild datasets introduce greater variations in pose, occlusion, identity, illumination, cultural factors, and context. These differences are particularly important for authenticity-aware FER, in which a model must distinguish between genuine and posed expressions beyond the specific conditions observed during training. Therefore, high accuracy on a controlled benchmark does not necessarily indicate reliable performance in real-world settings.

Overall, the findings suggest that controlled benchmarks remain valuable for initial model development and comparison; however, they should be complemented by diverse and unconstrained datasets. Cross-dataset evaluation, where the test data are not used during training, would provide a more realistic measure of generalization and help determine whether models learn meaningful cues of expression authenticity rather than dataset-specific patterns.

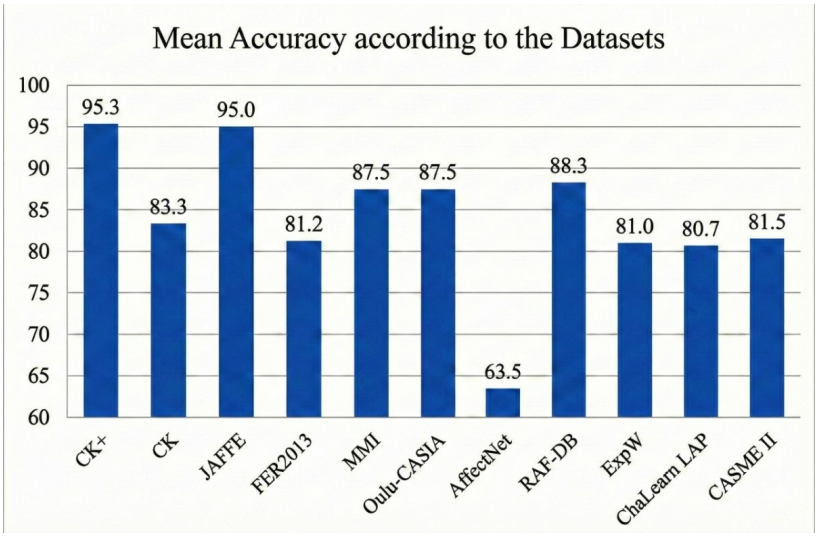


Figure 7. Mean detection accuracy reported for the datasets included in the review.

Table 9. Summary of dataset usage and reported performance in the reviewed studies.

4.4. Critical Synthesis: Handcrafted vs. Deep Feature Extraction

The extracted literature reports a historical mean accuracy of 95.12% for studies using Local Binary Patterns (LBP) and 92.32% for studies using deep-learned representations. These values should not be interpreted as evidence that LBP is intrinsically superior to CNN-based representations. The studies contributing to these averages differ in dataset size, acquisition conditions, subject splits, task definitions, and evaluation protocols.

The apparent LBP advantage is plausibly influenced by the concentration of handcrafted-feature studies on small, controlled datasets such as CK+ and JAFFE, where illumination, pose, background, and expression intensity are relatively constrained. Under such conditions, texture descriptors can provide a compact representation with a comparatively low sample complexity. Conversely, deep representations are generally evaluated across a broader range of unconstrained conditions and are capable of learning higher-level spatial features from larger datasets. Therefore, the appropriate conclusion is conditional: handcrafted descriptors can remain competitive under constrained protocols, whereas learned representations offer greater flexibility when sufficient data and appropriate validation are available.

For authenticity analysis, the relevant question is not which feature family has the highest aggregate accuracy, but whether the representation captures reproducible evidence associated with expression dynamics and veracity under subject-independent and cross-dataset evaluations. Therefore, future comparisons should report matched datasets, matched splits, identical metrics, and confidence intervals, where possible.

4.5. Efficacy of Feature Extraction Techniques

Convolutional neural networks (CNNs) were the most frequently used feature-learning strategy in the included studies, appearing in 22 studies in the current extraction (refer Fig. 8). Texture descriptors, geometric features, action units, and temporal representations were also employed. Rather than modeling these feature families using pooled accuracy, we interpreted their use according to the conditions under which each representation was informative.

Texture descriptors, such as LBP, can encode local intensity patterns and may be effective when the facial region is well aligned and the acquisition conditions are controlled. Geometric features capture relative landmark displacement and are useful for modelling facial configuration and motion. Action Units provide an interpretable representation of muscle-related facial actions, but their reliability depends on accurate AU annotation or detection. CNNs and transformer-based representations can learn higher-dimensional appearance and contextual features, while recurrent, convolutional-temporal, or transformer modules can model changes across frames.

The reviewed evidence supports a complementary rather than a winner-takes-all interpretation. For authenticity detection, combining spatial appearance, geometry, and temporal information may be more scientifically defensible than selecting a feature family solely from aggregate accuracy. Reported values such as 95.12%, 92.32%, 88.63%, and 66.50% are retained as descriptive study-level summaries (refer Fig. 9) and should not be treated as statistically equivalent estimates across heterogeneous experiments.

4.6. Techniques Used

The review articles were further grouped by technique to obtain additional insights. For each study, we identified the learning framework, evaluation protocol, and associated performance metrics.

Tables 11 through 13 group selected techniques according to dataset properties and target applications. Articles that tested algorithms on laboratory datasets are summarized in Table 11. Studies that utilized unconstrained image datasets are presented in Table 10. A summary of the video-based and multimodal approaches can be found in Table 12. Finally, approaches that incorporate micro-expression analysis, action unit activity, and 3D/4D facial representations are listed in Table 13.

The reviewed studies demonstrate considerable diversity in feature representations, learning strategies, datasets and evaluation protocols. Many studies rely on traditional handcrafted features such as LBP and

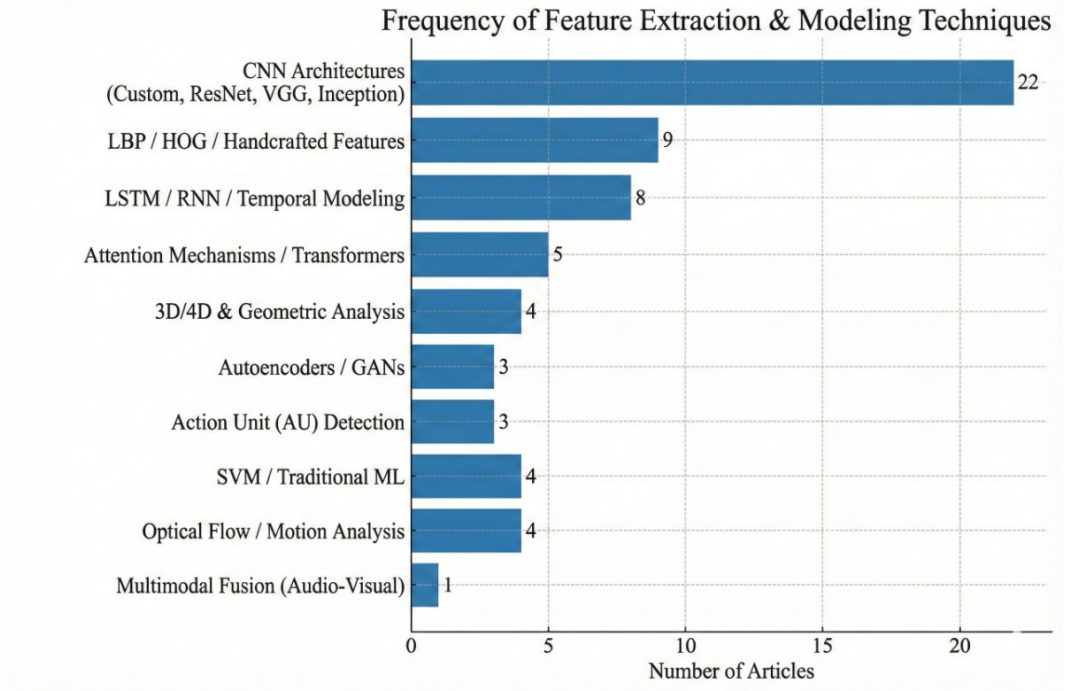


Figure 8. Frequency of each feature extraction technique.

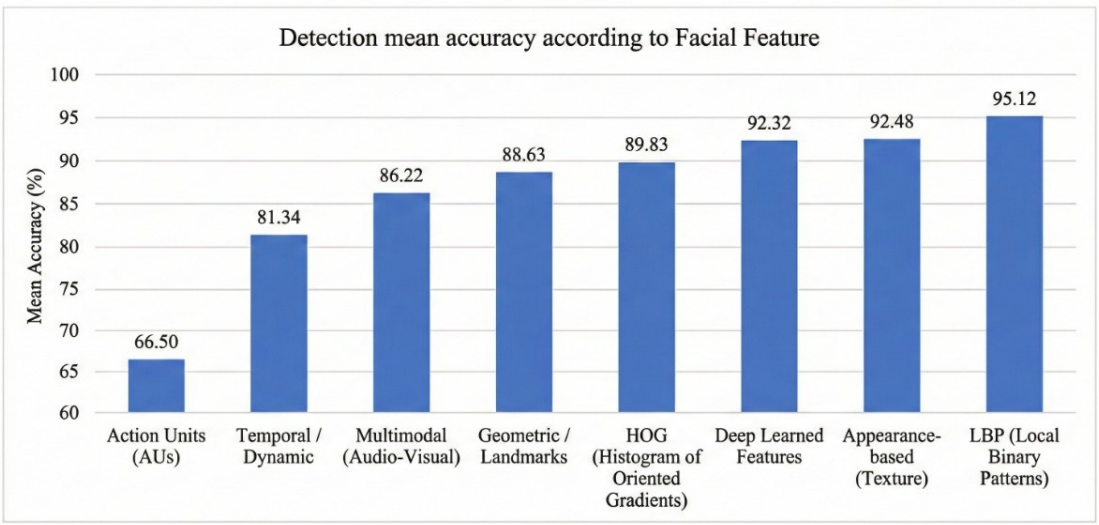


Figure 9. Mean detection accuracy when classifying each type of facial feature.

HOG descriptors. Recent studies have shifted towards using deep convolutional neural networks (CNNs) as the primary feature extractors. Hybrid models that combine CNNs and recurrent neural networks (RNNs) exploit temporal behavior, whereas attention and ST-LSTM frameworks model appearance and motion context.

The reported performance varied based on the dataset, ground-truth labels, and evaluation metrics. As mentioned previously, many benchmark datasets used to evaluate FER algorithms are collected in laboratory environments and lack realism.

4.7. Performance on Authenticity Detection

Authenticity detection is fundamentally different from conventional facial emotion recognition (FER). Although many FER models achieve recognition accuracies exceeding 95% on controlled laboratory benchmarks such as CK+, their ability to distinguish between genuine and posed expressions remains considerably more limited. To provide a comprehensive comparison, the reviewed studies are categorized according to the evaluation environment. The performance of representative methods on controlled laboratory datasets is summarized in Table 11, while the results obtained on challenging in-the-wild image datasets are presented in Table 10. Table 12 summarizes multimodal and video-based FER approaches that explicitly exploit temporal information, whereas Table 13 reviews methods designed for micro-expression, action-unit (AU)-based, and 3D/4D facial expression analysis. In contrast, the performance of methods specifically developed for emotion authenticity recognition is summarized in Table 5. The overall comparison indicates that conventional FER systems consistently outperform authenticity-aware models, highlighting that recognizing an emotional category and determining whether the emotion is genuine are two related but substantially different learning tasks.

The results reported in Table 11 demonstrate that controlled laboratory datasets remain the easiest benchmark for FER because of their constrained imaging conditions, frontal facial poses, uniform illumination, and well-defined emotional labels. Consequently, deep learning architectures such as CNNs, hierarchical CNNs, Grid-LSTMs, and hybrid CNN–feature-engineering approaches routinely achieve recognition accuracies between 93% and 98%. However, these datasets contain deliberately elicited or acted facial expressions under controlled acquisition settings; therefore, their excellent performance does not necessarily translate to practical applications where spontaneous facial behavior dominates.

The performance decreased noticeably when the models were evaluated under unconstrained conditions. As shown in Table 10, datasets such as AffectNet, RAF-DB, FER2013, FER+, and ExpW introduce substantial variations in illumination, pose, occlusion, ethnicity, age, and image quality. Although modern attention-based CNNs, transformers, and region-aware architectures have significantly improved robustness, recognition performance generally falls to approximately 80–90%, with AffectNet remaining particularly challenging because of its large intra-class variability and annotation ambiguity. These results demonstrate that environmental variability substantially affects the generalization capabimodelingER systems.

Temporal facial dynamics become increasingly important when facial expressions are represented as videos rather than static images. As summarized in Table 12, CNN–RNN, CNN–LSTM, GRU-based architectures, ConvLSTM networks, and transformer-based temporal models exploit sequential information to capture the onset, apex, and offset phases of facial expressions. Such temporal modelling generally improves emotion recognition, particularly for subtle and continuously evolving expressions. Similarly, Table 13 shows that micro-expression recognition relies on highly discriminative representations based on optical flow, action units, graph neural networks, and dynamic 3D facial geometry because these expressions are characterized by extremely small facial muscle movements and very short durations. Although these techniques significantly improve the recognition of subtle facial dynamics, the reported performance indicates that recognizing micro-expressions remains considerably more challenging than recognizing conventional posed expressions.

Despite these advances, distinguishing genuine facial expressions from posed expressions remains considerably more challenging than conventional facial emotion recognition. As summarized in Table 5, the reviewed studies generally report lower performance for authenticity recognition than for standard FER tasks, with reported accuracies typically ranging between approximately 70% and 80% on publicly available datasets such as ChaLearn. Unlike conventional FER, authenticity recognition requires models to differentiate facial expressions that often belong to the same emotion category but vary only in

subtle spatial, temporal, and physiological characteristics. In addition, authenticity-oriented datasets are generally smaller, exhibit greater class imbalance, and employ different annotation protocols and evaluation methodologies, making direct comparisons across studies difficult. These factors collectively contribute to the variability in the reported performance. Therefore, the reported accuracies should be interpreted as descriptive observations of the reviewed literature rather than as universal benchmarks for authenticity recognition. The findings also suggest that future research would benefit from larger, more diverse datasets, standardized annotation protocols, and cross-dataset evaluation strategies to improve the generalizability and reproducibility of authenticity-aware FER systems.

Several observations emerge from the comparative analysis of the reviewed Several observations emerge from the comparative analysis of the reviewed literature. First, increasing dataset complexity generally leads to a progrmodelingduction in recognition performance, moving from controlled laboratory environments (Table 11) to unconstrained in-the-wild scenarios (Table 10), followed by temporal video analysis (Table 12), micro-expression recognition (Table 13), and finally authenticity detection (Table 5). Second, temporal modelling appears to be particularly beneficial for authenticity analysis because genuine facial expressions typically exhibit natural onset, apex, and offset transitions, whereas posed expressions often display abnormal temporal evolution [9]. Third, multimodal information, facial action units, and dynamic facial representations provide complementary cues that cannot be captured from static appearance alone. Nevertheless, thettowardsing evidence does not establish that temporal information by itself is sufficient to determine emotional genuineness, and the quality of authenticity annotations remains a major factor influencing reported performance.

Figure 1 further illustrates the relationship between the different categories of FER techniques reviewed in this survey. It highlights the evolution from conventional image-based FER toward video-based, multimodal, micro-expression, and authenticity-aware emotion recognition, emphasizing the increasing complexity of the underlying learning problem. Combined with the quantitative comparisons presented in Tables 11–5, the figure demonstrates that future research should move beyond conventional emotion-classification accuracy and adopt evaluation protocols specifically designed for authenticity recognition.

Table 10. FER Techniques on In-the-Wild Image Datasets

Dataset(s)	Techniques / Methods	Accuracy (in %)
CK+, AffectNet [13]	CNN + ABC optimization; EWT features	N/A
FER+ [27]	VGG-13; ResNet; Crowd re-annotation	85–87
AffectNet [28]	ResNet-50; VGG-16; ViT; Attention	60–67
RAF-DB [15]	Deep Alignment; Region Attention; Self-Cure Net	87–89.5
ExpW [12]	Deep HOG; Relation-aware CNN	80–82
FER2013 [10]	Hybrid LBP–CNN; ResNet; Dlib landmarks	90.20
FER2013 [63]	Shallow CNN; Deep CNN	89.98

Table 11. Performance of the reviewed techniques on conventional laboratory datasets.

Dataset(s)	**Techniques / Methods**	**Accuracy / Performance**
CK+ [11]	LBP; Autoencoders; Custom CNN	96.50%
CK+ [14]	2D Grid LSTM; AAMs	94.50%
CK+, JAFFE [64]	Custom CNN (6 Conv + 3 MaxPool)	JAFFE: 95.23%; CK+: 93.24%
JAFFE, CK [65]	Deep CNN; Gaussian normalization; BatchNorm	JAFFE: 95.23%; CK: 93.24%
JAFFE, FER2013 [66]	CNN + LBP + HOG; Contrast normalization	JAFFE: 98.65%
CK+, JAFFE [25]	Hierarchical CNN; Autoencoder-based neutral synthesis	CK+: 96.46%; JAFFE: 91.27%
Oulu-CASIA [34]	LBP-TOP; P3D ResNet; Deformable Conv	85–90%
MMI [9]	Audio-visual fusion; GANs; BiLSTM	95%+ (sequence-based)

Table 12. Multimodal and Video-Based FER Techniques

Dataset(s)	Techniques / Methods	Performance
CK, FER2013 [67]	CNN (face); RNN (speech); MFCC	78.12%
AFEW [9]	CNN–RNN; Cascaded CNN; BiLSTM	74.4%
ChaLearn LAP [68]	DeepFace; VGG + LSTM-PB; GBM	97% (offline)
CK, IMPA-FACE3D [45]	C3D; TSN; ConvLSTM	67.71%
Aff-Wild2 [33]	CNN–RNN; ResNet + GRU; Transformers	CCC (V): 0.4–0.6
DAiSEE [55]	C3D; Inception + LSTM; TCN	55–63%

Table 13. Micro-Expression, AU-Based, and 3D/4D FER

Dataset(s) / Ref	Techniques / Methods	Performance
CASME II [36]	3D-CNN; Optical flow; GNN	75–88%
DISFA [38]	AU detection; Structural SVM; ResNet	F1: 0.60–0.75
4DFAB [44]	4D mesh analysis; Dynamic 3D fitting	70–80%

4.8. The Authenticity Classification Gap

The results presented above highlight a clear gap between conventional facial emotion recognition and emotion authenticity recognition. Figure 10 summarizes this difference in terms of the reported classification accuracies.

The reviewed studies consistently report high recognition performance for basic emotion classification on controlled benchmark datasets. Among the extracted study-level averages, the emotion category "happiness" achieved the highest mean recognition accuracy of **95.40%**, followed by "surprise" with a mean accuracy of **93.60%**. These results indicate that contemporary FER systems are generally effective at identifying the emotional category being expressed under controlled evaluation conditions.

A noticeable reduction in performance was observed when the task shifted from conventional emotion recognition to authenticity assessment (Fig. 10). Based on the extracted study-level results, genuine facial expressions achieved a mean classification accuracy of approximately **78.50%**, whereas posed (fake) expressions achieved a mean accuracy of approximately **76.20%**. Compared with the high recognition accuracy typically reported for basic emotion classification, authenticity recognition exhibits an absolute reduction of approximately 17 percentage points in the reviewed studies.

This performance difference suggests that recognizing the emotional category and determining whether an expression is genuine are related but fundamentally different learning tasks. While existing FER algorithms achieve high performance for emotion classification, distinguishing genuine from posed facial expressions remains substantially more challenging because the differences are often manifested through subtle spatial, temporal, and physiological facial cues. Consequently, relying solely on static facial appearance may be insufficient for robust authenticity recognition. The reviewed evidence suggests that incorporating temporal facial dynamics, facial muscle behavior, and multimodal information can improve the robustness of authenticity-aware FER systems, although further validation using larger and more diverse datasets is still required.

5. Limitations and Open Challenges

Despite substantial advances in facial emotion recognition (FER), reliably distinguishing between genuine and posed facial expressions remains a significant research challenge. The evidence synthesized throughout this review 5 and Figures 1 indicates that contemporary FER models generally achieve high performance for conventional emotion classification under benchmark evaluation protocols, whereas authenticity recognition remains considerably more challenging. The reviewed literature suggests that

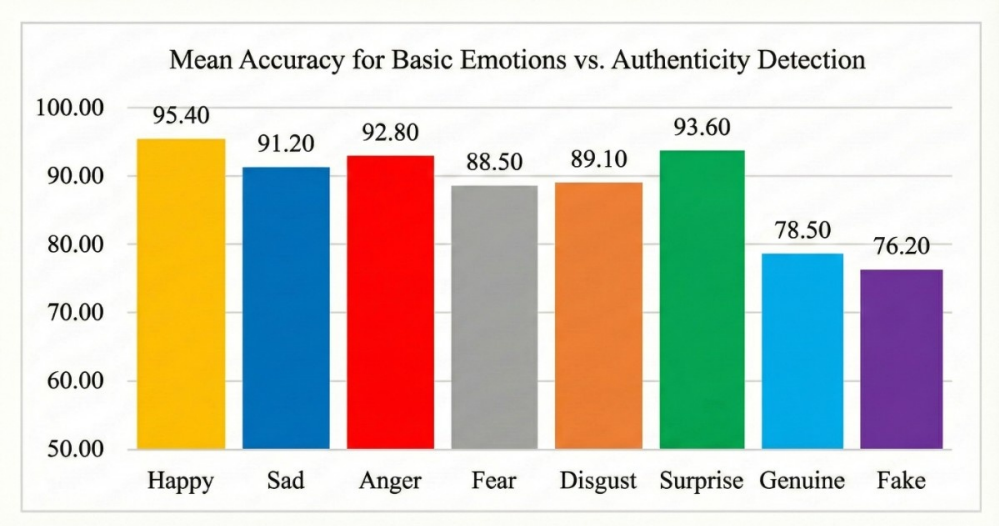


Figure 10. Classification accuracy of basic emotion classes compared to genuinity detection.

these limitations arise from a combination of dataset characteristics, annotation uncertainty, evaluation protocols, model generalization, computational requirements and ethical considerations. The following subsections summarize the principal research challenges identified in the reviewed studies.

5.1. Dataset Bias and the Authenticity Gap

The availability of large-scale datasets with reliable authenticity annotations remains a major challenge for authenticity-aware FER. As discussed in Tables 5, widely used benchmark datasets such as CK+, JAFFE, FER2013, RAF-DB, AffectNet, and Aff-Wild2 have substantially contributed to the development of conventional emotion recognition methods. However, many of these datasets primarily contain acted or experimentally elicited facial expressions and therefore provide limited information regarding the subtle facial dynamics associated with spontaneous emotional behavior.

Authenticity-oriented datasets are generally smaller and frequently exhibit class imbalances, demographic biases, heterogeneous acquisition conditions, and inconsistent annotation protocols. Moreover, authenticity labels are inherently difficult to establish because an individual's internal emotional state cannot be directly observed. Existing datasets commonly rely on controlled experimental protocols, self-reported emotions, or expert annotations, each of which introduces potential sources of uncertainty. Consequently, the development of larger and more diverse datasets with standardized acquisition and annotation protocols remains an important research priority.

5.2. Generalization Across Real-World Conditions

Limited generalization to real-world environments remains another important challenge. The comparative analyses presented in Tables 11–5 indicate that the recognition performance generally decreases as the evaluation conditions become increasingly unconstrained. Models developed using laboratory datasets often exhibit reduced performance when evaluated on spontaneous facial expressions collected under varying illumination, head pose, occlusion, demographic diversity, and cultural differences.

These findings suggest that many existing models may partially learn dataset-specific characteristics in addition to generalized representations of facial behavior. Therefore, improving robustness across different acquisition conditions is likely to require advances in domain adaptation, cross-dataset evaluation, continual learning, and bias-aware training strategies.

5.3. Temporal Dynamics and Fine-Grained Facial Behaviour

Although convolutional neural networks have demonstrated strong performance in learning spatial facial representations, the reviewed literature indicates that authenticity recognition also depends on temporal facial behavior. Genuine and posed facial expressions may exhibit differences in their onset, apex, and offset dynamics; however, these temporal characteristicmodelingross individuals and emotional contexts. Consequently, static image analysis alone may be insufficient for robust authenticity recognition in several practical scenarios.

Micro-expression analysis further increases the complexity of this problem. These brief and subtle facial movements are often difficult to capture using conventional imaging systems and static feature representations. As summarized in Table 13, integrating temporal sequence modelling, facial action units, dynamic facial geometry, and complementary multimodal information represents a promising direction for improving authenticity-aware FER, although additional validation under realistic conditions remains necessary.

5.4. Computational Constraints for Real-Time Deployment

Deploying authenticity-aware FER systems in real-world applications introduces additional computational challenges. Reliable analysis of subtle facial dynamics may benefit from high-frame-rate imaging systems capable of capturing rapid facial movements. However, such hardware is not always available in consumer-grade devices, potentially limiting deployment in applications such as mobile healthcare, remote patient monitoring, intelligent tutoring systems, and human–computer interaction.

Furthermore, many recent transformer-based and multimodal architectures require considerable computational resources for both training and inference. Developing lightweight, computationally efficient, and hardware-aware models remains an important direction for enabling practical real-time deployment on edge devices.

5.5. Explainability, Fairness, and Ethical Considerations

As authenticity-aware FER expands into healthcare, education, security, digital forensics, and human–computer interaction, explainability and fairness become increasingly important. Many deep learning models provide limited insight into the facial regions or temporal patterns contributing to their predictions. Explainable artificial intelligence (XAI) techniques, including attention visualization, Grad-CAM, SHAP, and network dissection, can improve model interpretability by providing more transparent explanations of decision-making processes.

Ethical considerations also remain central to the responsible deployment of authenticity recognition systems. Because these systems process sensitive biometric information, issues related to privacy, informed consent, demographic fairness, potential algorithmic bias, and misuse for surveillance or behavioral profiling require careful attention. Future research should therefore emphasize transparent evaluation protocols, fairness-aware model development, standardized reporting practices, and privacy-preserving learning frameworks to support the trustworthy deployment of authenticity-aware FER systems.

6. Conclusion

This systematic review presents a comprehensive synthesis of recent advances in facial emotion recognition (FER), with particular emphasis on the emerging problem of distinguishing genuine facial expressions from posed facial expressions. The reviewed literature indicates that deep learning approaches, including convolutional neural networks, graph neural networks, transformers, multimodal fusion strategies, and temporal sequence modelling, have substantially improved the performance of conventional FER across a wide range of benchmark datasets. However, the evidence synthesized in this review suggests that

authenticity recognition remains considerably more challenging than conventional emotion classification methods.

The comparative analyses presented in Tables 11–5 and Figures 1 indicate that recognition performance generally decreases as the evaluation conditions become increasingly realistic, progressing from controlled laboratory datasets to unconstrained environments, video-based analysis, micro-expression recognition, and authenticity recognition. Although direct quantitative comparisons across studies are limited by differences in datasets, evaluation protocols, and performance metrics, the reviewed literature consistently suggests that authenticity recognition requires models capable of exploiting complementary spatial, temporal, and contextual information rather than relying solely on static facial appearance.

This review also highlights several important research challenges identified across the included studies, including the limited availability of large-scale authenticity-oriented datasets, uncertainty in authenticity annotations, restricted cross-dataset generalization, computational requirements for real-time deployment, and the growing need for explainable, fair, and privacy-aware artificial intelligence. Addressing these challenges requires continued progress in dataset development, benchmark standardization, model design, and rigorous evaluation under realistic deployment conditions.

Overall, authenticity-aware FER represents a rapidly evolving research area with potential applications in healthcare, affective computing, education, security, digital forensics, and intelligent human–computer interaction. Continued advances in multimodal learning, explainable artificial intelligence, trustworthy machine learning, and standardized evaluation protocols are expected to further improve the robustness, interpretability, and practical applicability of future authenticity-aware FER systems.

7. Future Scope

The findings of this review suggest several promising directions for future research. First, there is a clear need to develop large-scale authenticity-oriented facial expression datasets containing spontaneous emotional expressions, standardized annotation protocols, balanced demographic representation, and diverse acquisition conditions. Such datasets would facilitate more reliable benchmarking and improve the evaluation of model generalization across different populations and real-world environments.

A second important direction is the integration of multimodal behavioral information. Combining facial expressions with speech, body posture, eye gaze, physiological signals, and contextual information may provide complementary cues for distinguishing genuine and posed emotions more reliably than facial appearance alone. Future studies should also investigate effective strategies for multimodal feature fusion and cross-modal representation learning.

Recent developments in foundation models, vision transformers, graph neural networks, state-space models, and self-supervised learning have provided new opportunities for authenticity-aware FER. These approaches have the potential to improve representation learning while reducing the dependence on manually annotated data. In addition, domain adaptation, continual learning, and cross-dataset evaluation should receive greater attention to improve robustness under diverse acquisition conditions.

Future research should also focus on developing lightweight and computationally efficient models suitable for deployment on edge devices, mobile platforms, and wearable systems. Integrating explainable artificial intelligence, uncertainty estimation, fairness-aware learning, and privacy-preserving techniques will be important for supporting trustworthy deployment in practical applications, including healthcare, education, driver monitoring, human–computer interaction, and digital forensics.

Finally, the reviewed literature highlights the need for standardized evaluation protocols specifically designed for authenticity recognition. Beyond conventional accuracy-based evaluation, future benchmarks should consider subject-independent testing, cross-dataset validation, calibration analysis, robustness assessment, fairness evaluation, interpretability, and reproducibility. Establishing such standardized evaluation frameworks would facilitate more reliable comparison between competing approaches and

support the development of scientifically rigorous, transparent, and practically deployable authenticity-aware FER systems.

Data Availability

No new datasets were generated or analyzed during the current study. This article is a comprehensive review based on previously published research, publicly available facial expression recognition and deception/veracity analysis datasets, and related scholarly literature. Information regarding the datasets discussed in this review, including their sources and accessibility, is provided within the manuscript and corresponding cited references. All referenced datasets are publicly accessible through their respective repositories or original publications and adhere, where applicable, to FAIR (Findable, Accessible, Interoperable, and Reusable) data principles.

REFERENCES

1. Mohamed Abdeldayem, Wael Badawy, Hesham FA Hamed, and Amr M Nagy. Improving facial expression recognition in real-world environments. *Statistics, Optimization & Information Computing*, 14(6):3546–3564, 2025.
2. Alexander Romanishyn, Olena Malytska, and Vitaliy Goncharuk. Ai-driven disinformation: policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence*, 8:1569115, 2025.
3. Matteo Cardaioli, Alessio Miolla, Mauro Conti, Giuseppe Sartori, Merylin Monaro, Cristina Scarpazza, and Nicolò Navarin. Face the truth: interpretable emotion genuineness detection. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2022.
4. Mohamed A Saleh, Alan Ting Yong, N Marbukhari, YM Yussoff, N Nabila Mohamed, Ali Abd Almisreb, and Habibah Hashim. Facial expression recognition: A new dataset and a review of the literature. *Turkish Online Journal of Qualitative Inquiry*, 12(6), 2021.
5. Ramcharan Ramanaharan, Deepani B Guruge, and Johnson I Agbinya. Deepfake video detection: Insights into model generalisation—a systematic review. *Data and Information Management*, page 100099, 2025.
6. Sehmus Yakut, Yusuf Taha Tuten, Eren Caglar, and Mehmet S. Aktas. A multimodal transformer-based framework for emotion analysis in multilingual video content. *Computers*, 15(2):77, 2026.
7. Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)*, pages 46–53. IEEE, 2000.
8. Mostafa M Moussa, Usman Tariq, Fares Al-Shargie, and Hasan Al-Nashash. Discriminating fake and real smiles using electroencephalogram signals with convolutional neural networks. *IEEE Access*, 10:81020–81030, 2022.
9. Minh Tu Nguyen, Quoc Khanh Nguyen, Kotani Kazunori, and Prarinya Siritanawan. Towards recognizing facial expressions at deeper level: Discriminating genuine and fake smiles from a sequence of images. In *2019 6th NAFOSTED Conference on information and computer science (NICS)*, pages 387–392. IEEE, 2019.
10. Alwin Poulose, Jung Hwan Kim, and Dong Seog Han. Feature vector extraction technique for facial emotion recognition using facial landmarks. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1072–1076. IEEE, 2021.
11. Ji-Hae Kim, Byung-Gyu Kim, Partha Pratim Roy, and Da-Mi Jeong. Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. *IEEE access*, 7:41273–41285, 2019.
12. Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.
13. AN Arun, P Maheswaravenkatesh, and T Jayasankar. Facial micro emotion detection and classification using swarm intelligence based modified convolutional network. *Expert Systems with Applications*, 233:120947, 2023.
14. Dhvanil Bhagat, Abhi Vakil, Rajeev Kumar Gupta, and Abhijit Kumar. Facial emotion recognition (fer) using convolutional neural network (cnn). *Procedia Computer Science*, 235:2079–2089, 2024.
15. Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Affective Computing*, 2019. Introduces the Real-world Affective Face Database (RAF-DB).
16. GA Rajesh Kumar, Ravi Kant Kumar, and Goutam Sanyal. Discriminating real from fake smile using convolution neural network. In *2017 international conference on computational intelligence in data science (ICCIDS)*, pages 1–6. IEEE, 2017.
17. Xuan-Phung Huynh and Yong-Guk Kim. Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3065–3072, 2017.
18. Yasmeen ELsayed, Ashraf ELSayed, and Mohamed A Abdou. An automatic improved facial expression recognition for masked faces. *Neural Computing and Applications*, 35(20):14963–14972, 2023.

19. Neelesh Bhakt, Pankaj Joshi, and Piyush Dhyani. A novel framework for real and fake smile detection from videos. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*, pages 1327–1330. IEEE, 2018.
20. Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010.
21. Michael Lyons, Miyuki Kamachi, and Jiro Gyoba. The japanese female facial expression (jaffe) dataset, 1997.
22. Hamid Golzadeh, Diego R Faria, Luis J Manso, Anikó Ekárt, and Christopher D Buckingham. Emotion recognition using spatiotemporal features from facial expression landmarks. In *2018 International Conference on Intelligent Systems (IS)*, pages 789–794. IEEE, 2018.
23. Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010.
24. Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(1):149–161, 2007.
25. Ninad Mehendale. Facial emotion recognition using convolutional neural networks (ferc). *SN Applied Sciences*, 2(3):446, 2020.
26. Wenjie Niu, Yuankun Zhao, Zhiyan Yu, Yu Liu, and Yu Gong. Research on a face recognition algorithm based on 3d face data and 2d face image matching. *Journal of Visual Communication and Image Representation*, 91:103757, 2023.
27. Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016.
28. Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.
29. Abhinav Dhall, Roland Göcke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2106–2112, 2011.
30. Carlos F. Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570, 2016.
31. Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 5683–5692, 2019.
32. Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.
33. Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855*, 2019.
34. Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.
35. Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 881–888, 2013.
36. Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014.
37. Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*, pages 1–6. IEEE, 2013.
38. S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.
39. Jun Wan, Sergio Escalera, Gholamreza Anbarjafari, Hugo Jair Escalante, Xavier Baró, Isabelle Guyon, Meysam Madadi, Juri Allik, Jelena Gorbova, Chi Lin, et al. Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3189–3197, 2017.
40. Sarah Tasnim Diya, Most Jannatul Ferdos, and Md. Mizanur Rahman. Genuine and Fake Facial Emotion Dataset (GFFD-2025), 2025.
41. Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006.
42. Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–6. IEEE, 2013.
43. Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.
44. Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4dfab: A large scale 4d facial expression database for biometric applications. *arXiv preprint arXiv:1712.01443*, 2017.

45. J. Rannow Budke and M. Da Costa-Abreu. Using neural and distance-based machine learning techniques in order to identify genuine and acted emotions from facial expressions. In *11th International Conference of Pattern Recognition Systems (ICPRS 2021)*, volume 2021, pages 121–126, 2021.
46. Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 2024.
47. Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2018.
48. Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011.
49. Hung Nguyen, Nha Tran, Hien D Nguyen, Loan Nguyen, and Kazunori Kotani. Ktfev2: Multimodal facial emotion database and its analysis. *IEEE Access*, 11:17811–17822, 2023.
50. Nancy Gelvez, Carlos Motenegro, and Paulo Gaona. Advanced emotion recognition: A heuristic approach applied to eeg signals using machine learning. *Statistics, Optimization & Information Computing*, 14(2):904–922, 2025.
51. Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. Emotiw 2016: video and group-level emotion recognition challenges, 2016.
52. Abhinav Dhall, Monisha Singh, Roland Goecke, Tom Gedeon, Donghuo Zeng, Yanan Wang, and Kazushi Ikeda. Emotiw 2023: Emotion recognition in the wild challenge, 2023.
53. Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E. Hughes, and Louis-Philippe Morency. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 137–144, New York, NY, USA, 2016. Association for Computing Machinery.
54. Fuzail Khan. Facial expression recognition using facial landmark detection and feature extraction via neural networks. *arXiv preprint arXiv:1812.04510*, 2018.
55. Abhay Gupta, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*, 2016.
56. Bo Yang, Jianming Wu, and Gen Hattori. Facial expression recognition with the advent of human beings all behind face masks, 2020.
57. Elisa S. Dan-Glauser and Klaus R. Scherer. The geneva affective picture database (gaped): A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43(2):468–477, June 2011.
58. Ning Yan, Kunhee Han, and Seungsoo Shin. A comprehensive deepfake detection framework based on multimodal liveness detection and deep learning integration. *KSII Transactions on Internet and Information Systems*, 20(1):354, 2026.
59. Shaoqian Yu, Xingyu Chen, Yuzhe Sheng, Han Zhang, Xinlong Li, and Sijia Yu. Ace-net: A fine-grained deepfake detection model with multimodal emotional consistency. *Electronics*, 14(22):4420, 2025.
60. Ayod Bhad et al. Multi-modal deep learning approaches for enhanced audio-visual deepfake detection. *International Journal of Advanced Computer Science and Applications*, 2025.
61. Muhammad Aleem, Muhammad Umair, Muhammad Zubair, Rozeena Ibrahim, Muhammad Tahir Naseem, Muhammad Mohsin Raza, Muhammad Nadeem Ali, and Byung-Seo Kim. Seeing through the fake: Explainable ai with multiple cnns for deepfake detection. *IEEE Access*, 14:131–162, 2026.
62. Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.
63. Rohit Pathar, Abhishek Adivarekar, Arti Mishra, and Anushree Deshmukh. Human emotion recognition using convolutional neural network in real time. In *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, pages 1–7. IEEE, 2019.
64. Deepak Kumar Jain, Pourya Shamsolmoali, and Paramjit Sehdev. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120:69–74, 2019.
65. Mudit Agarwal, Mradul Mittal, Milind Gautam, and Deva Nand. Real-time facial emotion classification using deep convolution neural network. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–6. IEEE, 2022.
66. Ansamma John, MC Abhishek, Ananthu S Ajayan, S Sanoop, and Vishnu R Kumar. Real-time facial emotion recognition system with improved preprocessing and feature extraction. In *2020 Third international conference on smart systems and inventive technology (ICSSIT)*, pages 1328–1333. IEEE, 2020.
67. Gonzalo Fernández Florio, Marıa Elena Buemi, Daniel Acevedo, and Pablo Negri. Attribute classification for the analysis of genuineness of facial expressions. In *11th International Conference of Pattern Recognition Systems (ICPRS 2021)*, volume 2021, pages 109–114. IET, 2021.
68. Teerapong Winyangkun, Noparut Vanitchanant, Varin Chouvatut, and Benjamas Panyangam. Real-time detection and classification of facial emotions. In *2023 15th International Conference on Knowledge and Smart Technology (KST)*, pages 1–6. IEEE, 2023.