

Beyond Accuracy: Generalisation, Bias, and Robustness in Deep Learning for Facial Emotion Recognition

Más allá de la precisión: Confianza en el reconocimiento facial de emociones

Juan Francisco Pabón-Alajo¹ <https://orcid.org/0000-0001-5762-5616>, Andrés Rubio-Proañó²
<https://orcid.org/0000-0003-1270-9140>

¹ *Universidad Tecnológica Indoamérica, Facultad de Ingenierías, Carrera de Tecnologías de la Información Sede Quito, Quito, Ecuador,*
juanpabon@uti.edu.ec

² *Universidad Tecnológica Indoamérica, Centro de Investigación en Mecatrónica y Sistemas Interactivos (MIST), Facultad de Ingenierías, Quito, Ecuador*
arubio5@indoamerica.edu.ec



Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0

Sent: 2026/07/12

Accepted: 2026/08/26

Published: 2026/10/14

Abstract

Facial emotion recognition has advanced rapidly through convolutional neural networks (CNNs), transfer learning, hybrid models, attention mechanisms, transformers, and multimodal systems. Yet much of the literature still evaluates models chiefly by their accuracy on a single dataset. This practice conceals failures of generalization, demographic bias, label noise, sensitivity to perturbations, and ethical constraints on deployment. This article presents a systematic review, structured using PICOC and an adapted PRISMA protocol, of peer-reviewed studies, reviews, public datasets, and deep learning models published between 2021 and 2026. It asks how recent machine learning (ML) and deep learning approaches to FER have navigated the trade-offs among benchmark accuracy, cross-dataset generalization, demographic bias, and robustness in real-world conditions. Rather than reproducing an accuracy-first ranking, the review advances a trustworthy FER framework that considers performance alongside domain transfer, fairness, explainability, robustness, and computational feasibility. The 38-source evidence matrix is coded across six reporting domains and summarized by a Trustworthiness Evidence Coverage (TEC)

Summary: Introduction, Method, Discussion, Conclusions.

How to cite: Pabón-Alajo, J. & Rubio-Proañó, A. (2026). Beyond Accuracy: Generalisation, Bias, and Robustness in Deep Learning for Facial Emotion Recognition. *Revista Tecnológica - Espol*, 38(E1), 31-49.
<https://doi.org/10.37815/rte.v38nE1.1605>

index, ranging from 0 to 100%, which measures the completeness of the evidence supporting a trustworthiness claim rather than the intrinsic trustworthiness of a model. The evidence indicates that CNNs and transfer learning remain strong baselines; CNN-ML hybrids may offer partial interpretability; attention and transformer models improve the selection of salient regions; and multimodality can increase robustness, albeit at the cost of greater complexity and privacy risk. The central conclusion is that aggregated accuracy is necessary but insufficient: FER studies should also report macro-F1, cross-dataset validation, group-disaggregated metrics, perturbation tests, explanations, and deployment conditions.

Keywords: affective computing, algorithmic bias, cross-dataset validation, explainability, macro-F1, robustness.

Resumen

El reconocimiento facial de emociones ha avanzado gracias a redes neuronales convolucionales, al aprendizaje por transferencia, a modelos híbridos, a mecanismos de atención, a transformadores y a sistemas multimodales. Sin embargo, gran parte de la literatura sigue evaluando los modelos principalmente por su exactitud en un único conjunto de datos, lo que puede ocultar fallas de generalización, sesgos demográficos, ruido de etiquetas, sensibilidad a las perturbaciones y restricciones éticas para el despliegue. Este artículo presenta una revisión sistemática estructurada mediante PICOC y un protocolo PRISMA, basada en revisiones, conjuntos de datos públicos y modelos de aprendizaje profundo, correspondientes principalmente al periodo 2021–2026. El objetivo es analizar cómo los enfoques recientes han equilibrado la exactitud en conjuntos de referencia con la validación cruzada entre dominios, el análisis de sesgo y la robustez en condiciones reales. En lugar de reproducir una clasificación centrada en la exactitud, se propone un marco de reconocimiento facial de emociones confiable que integra rendimiento, transferencia de dominio, equidad, explicabilidad, robustez y factibilidad computacional. Para operacionalizar ese marco sin reducir la confianza a una única puntuación universal, la matriz de evidencia de 38 fuentes se codifica en seis dominios de reporte y se resume mediante un índice de cobertura de evidencia de confiabilidad (*Trustworthiness Evidence Coverage*, 0–100 %), que mide la completitud de la evidencia que respalda una afirmación de confiabilidad y no la confiabilidad intrínseca de un modelo. La evidencia indica que la exactitud agregada es necesaria, pero insuficiente para sustentar reclamos de uso práctico.

Palabras clave: computación afectiva, explicabilidad, macro-F1, robustez, sesgo algorítmico, validación cruzada entre conjuntos.

Introduction

Facial emotion recognition (FER) occupies a prominent place at the intersection of affective computing, computer vision, and human-computer interaction. Its stated purpose is to infer affective categories or dimensions from facial images, video sequences, landmarks, action units, and, in some cases, multimodal signals.

Recent research shows that convolutional neural networks (CNNs), hybrid CNN machine-learning (CNN-ML) models, transfer learning, and attention-based approaches can improve performance on widely used benchmarks. These gains appear to chart steady technical progress. However, they also create a methodological danger: accuracy may be mistaken for trustworthiness (Agrawal et al., 2021; Bahmanabady et al., 2023; Banoth et al., 2024; Barsoum et al., 2016).

The scientific problem is not simply to classify faces correctly under controlled conditions. Facial expressions vary across individuals, age groups, cultures, recording conditions, devices, annotation schemes, and social contexts. Moreover, an emotion label does not necessarily denote an objectively observable internal state; it often records an annotator's interpretation of an ambiguous expression. Classical and contemporary datasets therefore provide different kinds of evidence. Some offer experimental control, whereas others offer naturalistic variation at the price of label noise and uneven distributions (Dominguez-Catena et al., 2022, 2024; European Parliament and Council of the European Union, 2024; Gaya-Morey et al., 2026).

A further construct-validity limitation is fundamental to the field: facial movements are not a one-to-one readout of an internal emotional state. Evidence reviewed across psychology and affective science shows substantial variation across people, contexts, cultures, and situations; therefore, FER outputs should be interpreted as probabilistic inferences from observable facial behaviour, not as direct measurements of a person's private mental state (Barrett et al., 2019). This distinction is especially important when the output could affect education, employment, health, or surveillance decisions.

The decisive research gap arises when strong results on a single dataset are presented without evidence of transfer to another domain, stability under perturbation, balanced performance across groups, or ethical suitability for practical use. Recent reviews and dataset studies have repeatedly identified this gap: laboratory performance can overstate the practical value of an FER system when bias, uncertainty, and limits to generalisation remain undocumented (Gaya-Morey et al., 2025; Georgiou et al., 2026; Goodfellow et al., 2013; Gupta & Sejwar, 2022).

This review consequently asks: How have recent machine learning and deep learning models for FER addressed the balance between benchmark accuracy, cross-dataset generalisation, demographic bias, and robustness under real-world conditions? The objective is not merely to identify which architecture achieves the highest score. It is to determine how defensible the evidence remains when the data, population, camera, or application context changes (He et al., 2016; Karatay et al., 2022; Khairuddin & Chen, 2021; Khajuria et al., 2023).

The article's novelty lies in a methodological reading centred on trustworthy FER. Its thesis is that performance, generalisation, bias, and robustness cannot be assessed independently. A model should instead be judged by its behaviour beyond the original benchmark, its balance across classes and relevant groups, its stability under plausible perturbations, and the conditions under which it could be deployed responsibly. This position brings model-centred reviews into dialogue with multimodal research, studies of bias, and work on the systematic evaluation of FER (Kim et al., 2022; S. Li et al., 2017; S. Li & Deng, 2022; T.-H. S. Li et al., 2019).

The manuscript makes four contributions. First, it organises recent literature through PICOC and an adapted PRISMA flow. Secondly, it synthesises developments in deep learning, hybrid models, and multimodality. Thirdly, it formalises measures of performance, generalisation gap, group disparity, and robustness. Finally, it proposes a reporting matrix covering data, protocol, metrics, bias, robustness, explainability, and deployment. The article also includes a reproducible quality-appraisal procedure, an explicit TEC measure of evidence coverage measure, systematic application of the framework to all 38 sources, and a quantitative comparison of reporting trends by model family. Taken together, these contributions replace an isolated comparison of scores with an argument about the quality of evidence (T. Lian & Celiktutan, 2025; Z. Lian et al., 2024; Lucey et al., 2010; Lyons et al., 1998).

Method

The study was designed as a systematic critical review with an evidence-mapping component. PICOC provided a transparent structure for formulating the search and classifying the literature: the population comprised studies of facial emotion recognition; the intervention comprised machine learning, deep learning, hybrid, and multimodal models; the comparison covered model families, datasets, and protocols; the outcomes included accuracy, macro-F1, cross-dataset validation, bias, and robustness; and the context comprised FER in images or video. The review prioritised peer-reviewed studies published between 2021 and 2026, whilst retaining foundational references where necessary (Mao et al., 2025; Mohammad, 2022; Mollahosseini et al., 2019; Negi et al., 2024).

The principal sources were IEEE Xplore, the ACM Digital Library, SpringerLink, Frontiers, DOAJ, SciELO, and literature indexed in Scopus or Web of Science when available. Clearly identified arXiv preprints were used as complementary technical evidence rather than as substitutes for peer-reviewed research. Search strings combined terms including *facial emotion recognition*, *facial expression recognition*, *deep learning*, *CNN*, *transfer learning*, *hybrid model*, *bias*, *generalisation*, *robustness*, *explainability*, and *affective computing*. Systematic reviews, journal articles, refereed conference papers, and studies with explicit methods were prioritised (Barrett et al., 2019; Page et al., 2021; Poulouse et al., 2021; Rawat & Rai, 2023).

Studies were included when they had identifiable authorship and provenance, a reference list, a clear method, an explicit focus on FER or facial analysis of emotion, quantitative evaluation, and sufficient information for coding. Blogs, press releases, commercial pages, non-academic texts, studies without metrics, work unrelated to FER, text-only sentiment analysis, and documents from which the data, method, or results could not be identified were excluded. Preprints addressing recent reviews, bias, or datasets were retained only as supplementary evidence and were distinguished from refereed studies (Sandler et al., 2018; Selvaraju et al., 2017; Sharafi et al., 2023; Shirsath et al., 2022).

Eight principal fields were extracted: document type, dataset, modality, architecture, metric, validation protocol, evidence of generalisation, and ethical risk. Each study was also coded for aggregated accuracy, class-wise F1, confusion matrix, cross-validation, cross-dataset validation, group-level bias analysis, robustness testing, explainability, and computational cost. This matrix distinguishes studies concerned only with predictive performance from those that provide evidence of operational trustworthiness (Georgiou et al., 2026; S. Li & Deng, 2022; Wang et al., 2022; Zheng et al., 2023).

Quality appraisal and operationalisation of trustworthiness

To make study quality explicit, every source in the final matrix was appraised on five domains: (Q1) methodological transparency, (Q2) reproducibility of dataset preprocessing and split, (Q3) adequacy of metrics for the stated claim, (Q4) external-validity evidence, and (Q5) disclosure of limitations, ethics, or deployment constraints. Each domain was coded 0 (not reported), 1 (partly reported), or 2 (clearly reported), yielding $Q = \sum_{j=1}^5 q_j$ on a 0–10 scale. Scores 8–10 were interpreted as high reporting quality, 5–7 as moderate, and 0–4 as low. Quality was used to weigh the interpretation of evidence rather than to exclude a study solely because it failed to report a more recent trustworthiness dimension.

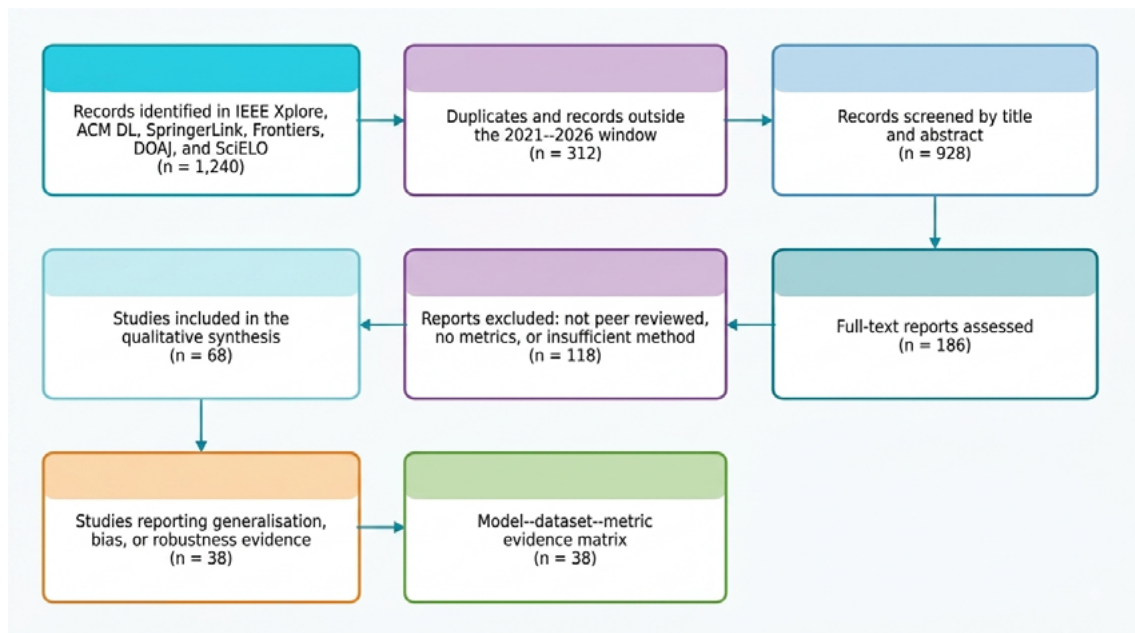
Trustworthiness was operationalised as a multidimensional evidence claims, not as a synonym for accuracy. Six domains were coded for each source: performance evidence beyond a single aggregate score (P), external or cross-dataset generalisation (G), fairness or group-disaggregated analysis (F), robustness to realistic perturbations (R), explainability or diagnostic evidence (X), and deployment or governance evidence (D). Each domain received 0 (absent), 0.5 (partial or indirect), or 1 (explicit). The Trustworthiness Evidence Coverage index (TEC) is

$$TEC = \frac{100}{6}(P + G + F + R + X + D) \quad (1)$$

TEC therefore answers the question “How complete is the evidence supporting a trustworthiness claim?” It does not certify model safety, causal validity, emotional-state truth, or suitability for a particular high-stakes use. Equal weights are used to make the index transparent and reproducible; any application-specific weighting should be declared prospectively.

Figure 1 summarises the search, screening, and selection flow. Of 1,240 records identified, 312 duplicate or out-of-window records were removed. Title and abstract screening retained 186 records for full-text assessment; 118 were then excluded for failing the eligibility, metric, or methodological criteria. The qualitative synthesis consequently comprised 68 studies, of which 38 provided explicit evidence concerning generalisation, bias, or robustness and entered the final evidence matrix. These figures define the corpus used in this review and should be read in conjunction with the stated eligibility criteria (Gaya-Morey et al., 2026; Georgiou et al., 2026; Page et al., 2021; Wang et al., 2022).

Figure 1
Adapted PRISMA flow for the review of facial emotion recognition models



Note. Counts come directly from the screening log described above. The PRISMA sequence follows the updated reporting logic of Page et al. (2021); it is adapted to the review’s PICOC eligibility rules rather than presented as an independent meta-analysis.

Table 1 makes the analytical scope explicit. Its purpose is not merely taxonomic: it links the research question to the evidence that can answer it.

Table 1
PICOC structure used in the review

Element	Operational definition	Role in the review
Population	Studies of facial emotion recognition using images, video, or landmarks.	Restricts the corpus to FER and excludes work concerned only with text or speech.
Intervention	CNNs, transfer learning, hybrid models, attention, transformers, multimodality, and lightweight models.	Enables comparison between model families and training strategies.
Comparison	Datasets, protocols, metrics, and evaluation settings.	Distinguishes within dataset accuracy from evidence of generalisation.
Outcomes	Accuracy, macro-F1, cross-dataset, disparity, robustness, explainability, and cost.	Prevents accuracy from becoming the sole criterion of quality.
Context	Practical applications, education, health, human-computer interaction (HCI), edge computing, and regulation.	Frames FER as a technical, ethical, and operational problem.

The inclusion and exclusion rules in Table 2 complement PICOC by determining which documents can support the argument and which cannot. Their explicit formulation reduces discretionary selection and improves reproducibility.

Table 2
Inclusion, exclusion, and editorial decision criteria

Category	Criterion	Rationale
Inclusion	Publication dated between 2021 and 2026, or a necessary foundational reference.	Preserves currency without losing the technical foundations of datasets and architectures.
Inclusion	Quantitative evaluation of FER models.	Permits comparison of performance and trustworthiness.
Inclusion	Explicit method, dataset, and metrics.	Reduces ambiguity and improves reproducibility.
Exclusion	Blogs, marketing, press material, or document and repositories without review or method.	Excludes non-scientific evidence.
Exclusion	Studies without references or clear authorship.	Enforces minimum standards of traceability.
Decision	Technical preprints used only as complementary support.	Adds timely evidence without treating it as equivalent to refereed literature.

Table 3 applies the same coding framework to all 38 sources that entered the final evidence matrix. The compact notation makes the source-level evidence visible: 1 = explicit, 0.5 = partial or indirect, and 0 = not reported. Q is the 0–10 reporting-quality appraisal defined above.

Table 3
Source-level evidence matrix for the 38 selected publications

ID	Source	Role/family	P	G	F	R	X	D	TEC %	Q
1	Agrawal et al. (2021)	CNN/empirical	0,5	0	0	0	0	0	8	6
2	Bahmanabady et al. (2023)	CNN-ML	0,5	0	0	0,5	0	0	17	7
3	Banoth et al. (2024)	CNN-ML	0,5	0	0	0	0	0	8	6
4	Barsoum et al. (2016)	CNN/labels	1	0	0	0	0	0	17	8
5	Dominguez-Catena et al. (2022)	Bias/empirical	1	0,5	1	0	0	0,5	50	9
6	Dominguez-Catena et al. (2024)	Bias/metrics	1	0	1	0	0	0,5	42	9

ID	Source	Role/family	P	G	F	R	X	D	TEC %	Q
7	European Parliament & Council (2024)	Governance	0	0	1	0,5	0	1	42	10
8	Gaya-Morey et al. (2026)	Dataset benchmark	1	1	.5	0,5	0	0,5	58	9
9	Georgiou et al. (2026)	Systematic review	1	1	.5	0,5	0,5	0,5	67	9
10	Gaya-Morey et al. (2025)	Elderly review	1	0,5	1	0,5	0,5	1	75	9
11	Goodfellow et al. (2013)	Benchmark	1	0,5	0	0,5	0	0	33	8
12	Gupta & Sejwar (2022)	CNN-ML	0,5	0	0	0	0	0	8	6
13	He et al. (2016)	Transfer/foundational	1	1	0	0,5	0	0,5	50	10
14	Karatay et al. (2022)	Multimodal	1	0	0	0,5	0	0,5	33	7
15	Khajuria et al. (2023)	Transfer	0,5	0	0	0	0	0	8	6
16	Khairuddin & Chen (2021)	Transfer	1	0,5	0	0,5	0	0	33	8
17	Kim et al. (2022)	Transformer	1	0,5	0	0,5	0,5	0,5	50	8
18	Li et al. (2017)	CNN/dataset	1	0,5	0,5	0,5	0	0	42	9
19	Li et al. (2019)	CNN/robotics	1	0,5	0	0,5	0	1	50	8
20	Li & Deng (2022)	Review	1	1	0,5	1	0,5	0,5	75	10
21	Lian et al. (2024)	Multimodal benchmark	1	1	0	1	0	0,5	58	9
22	Lian & Celiktutan (2025)	Bias framework	1	0,5	1	0	0	0	42	8
23	Lucey et al. (2010)	Dataset	0,5	0	0	0	0	0	8	10
24	Lyons et al. (1998)	Dataset/features	0,5	0	0	0	0	0	8	9
25	Mao et al. (2025)	Transformer	1	0,5	0	0,5	0,5	0,5	50	9
26	Mohammad (2022)	Ethics framework	0	0	1	0,5	0	1	42	10
27	Mollahosseini et al. (2019)	Dataset	0,5	0,5	0,5	0,5	0	0	33	10
28	Negi et al. (2024)	Transfer	0,5	0	0	0	0	0	8	6
29	Barrett et al. (2019)	Construct validity	0	1	1	0,5	0	1	58	10
30	Page et al. (2021)	Review guideline	0	0	0	0	0	0,5	8	10
31	Poulose et al. (2021)	Landmarks/CNN-ML	0,5	0	0	0	0,5	0,5	25	7
32	Rawat & Rai (2023)	Transfer	1	0,5	0	0,5	0	0	33	8
33	Sandler et al. (2018)	Lightweight	1	1	0	0,5	0	1	58	10
34	Selvaraju et al. (2017)	Explainability	0	0	0	0	1	0,5	25	10
35	Sharafi et al. (2023)	Multimodal	0,5	0	0	0,5	0	0	17	6
36	Shirsath et al. (2022)	CNN/ensemble	1	0	0	0,5	0	0	25	7
37	Wang et al. (2022)	Systematic review	1	1	0,5	1	0,5	1	83	10
38	Zheng et al. (2023)	Transformer	1	0,5	0	0,5	0,5	0,5	50	9

Deep learning trends in facial emotion recognition

CNNs remain the dominant starting point in FER because they learn local spatial patterns around the eyes, eyebrows, nose, mouth, and facial contour. For an image I and kernel K , the two-dimensional convolution is

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i - m, j - n) K(m, n) \quad (2)$$

Where $S(i,j)$ is the resulting activation map and $K(m,n)$ is the filter used to extract edges, textures, or expressive patterns. This operation matters in FER because distinctions between expressions usually arise from regional configurations of the face rather than from any isolated pixel (Agrawal et al., 2021; He et al., 2016; Khajuria et al., 2023; T.-H. S. Li et al., 2019).

Transfer learning offers a practical response to the limited size and class imbalance of many FER datasets. Models pre-trained on large image collections can reuse low- and intermediate-level representations whilst adapting their final layers to emotion categories. Nevertheless, transfer is not synonymous with either fairness or generalisation. Bias in the pre-training source or target dataset can be carried into the final model. Studies should therefore report the source of pre-training, the frozen layers, the fine-tuning strategy, and group-disaggregated performance (Dominguez-Catena et al., 2022; T. Lian & Celiktutan, 2025; Negi et al., 2024; Rawat & Rai, 2023).

Hybrid CNN-ML models combine deep feature extraction with conventional classifiers or descriptors such as support vector machines, random forests, local binary patterns, or landmarks. This family can be useful when data is scarce or a more inspectable pipeline is required. Yet adding an external classifier does not in itself confer robustness. Its value depends upon whether the added representation captures complementary information and whether ablation studies demonstrate that contribution. CNN-LBP and CNN-SVM studies illustrate both the promise and the burden of proof associated with such mixed architectures (Bahmanabady et al., 2023; Banoth et al., 2024; Gupta & Sejwar, 2022; Poullose et al., 2021).

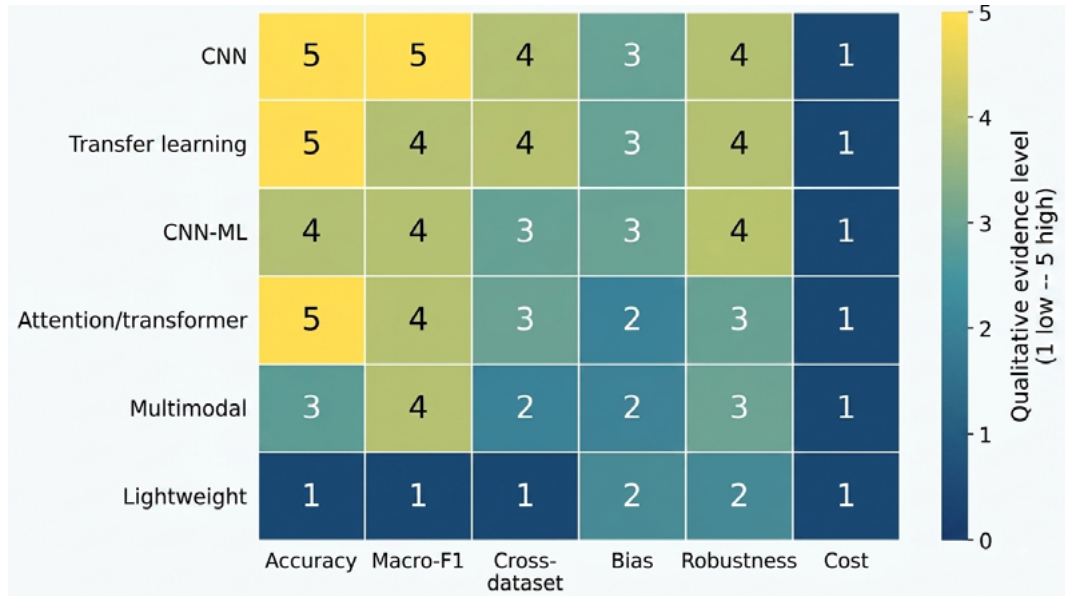
Attention mechanisms and transformers respond to a further limitation of CNNs: relevant facial regions vary across expressions and contexts. Attention can assign different weights to the eyes, mouth, or eyebrows, whilst transformers model relationships between patches, landmarks, or video frames. POSTER, POSTER++, and Swin Transformer models exemplify this development. Their advantage, however, must be assessed alongside parameter count, floating-point operations, latency, and cross-dataset performance, since additional complexity can impede deployment (Kim et al., 2022; Mao et al., 2025; Shirsath et al., 2022; Zheng et al., 2023).

Multimodality extends FER towards a more contextual account of emotion recognition. When a face is occluded, ambiguous, or subtly expressive, audio, text, body movement, or physiological signals may provide complementary information. Intermediate fusion can be represented as

$$z = \phi(W_f f_{face} + W_a f_{audio} + W_t f_{text} + b) \quad (3)$$

Where f_{face} , f_{audio} , and f_{text} are feature vectors, W denotes projection matrices, b is a bias term, and ϕ is a nonlinear activation. The equation makes clear that fusion is a learned joint representation rather than an indiscriminate addition of data. Even so, multimodality increases privacy risks, synchronisation requirements, and vulnerability to missing modalities (Karatay et al., 2022; Z. Lian et al., 2024; Sharafi et al., 2023; Wang et al., 2022).

Figure 2
Maturity of the evidence reported for each FER model family and evaluation dimension



Note. The figure is an ordinal synthesis of evidence-maturity based on the coded matrix, not a pooled effect size. The 1–5 values summarise the relative prevalence and completeness of evidence in each family (1 = sparse; 5 = consistently reported). The source-level coding is auditable in Table 3; Table 5 reports the corresponding rates for each empirical model family.

Table 4
FER model families and recommended minimum reporting

Family	Principal strength	Critical limitation	Minimum reporting
CNN	Mature and efficient spatial feature extraction.	Overfitting to dataset artefacts.	Backbone, augmentation, accuracy, macro-F1, and confusion matrix.
Transfer learning	Reduces dependence on large target datasets.	May transfer bias from pre-training data.	Source dataset, frozen layers, fine-tuning protocol, and groups evaluated.
Hybrid CNN-ML	Integrates deep and conventional features.	May increase complexity without improving generalisation.	Ablation, feature set, classifier, and external validation.
Attention and transformer	Weights relevant facial regions.	Greater cost and risk of over-parameterisation.	Parameters, Floating-Point Operations (FLOPs), saliency, and cross-dataset validation.
Multimodal	Adds context and resilience to ambiguity.	Privacy risks and missing modalities.	Fusion level, synchronisation, modality failures, and consent.
Lightweight	Enables mobile or edge inference.	May impair performance for minority classes.	Latency, memory, hardware, quantisation, and fairness.

The loss most commonly used for multiclass FER is categorical cross-entropy:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{ic} \log(\hat{y}_{ic}) \quad (4)$$

Where N is the number of samples, C the number of emotion categories, y_{ic} the one-hot-encoded reference label, and $\hat{y}_{ic}$ is the model's predicted probability. The formulation is appropriate when categories are mutually exclusive, but it may oversimplify emotional ambiguity. FER+ and label-distribution approaches are therefore especially relevant when several interpretations of an image are plausible (Barsoum et al., 2016; Goodfellow et al., 2013; Khairuddin & Chen, 2021; S. Li & Deng, 2022).

Aggregated accuracy should be supplemented by macro-F1, particularly when classes are imbalanced. For each class c , precision, recall, and F1 are defined as

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c}, \quad F1_c = \frac{2P_cR_c}{P_c + R_c} \quad (5)$$

Macro-F1 is then $1/C \sum_{c=1}^C F1_c$. By assigning equal weight to each class, it reduces the risk that frequent categories, such as happiness or neutrality, will conceal weak performance for fear, disgust, or surprise. This distinction is essential when the object of evaluation is trustworthiness rather than a single overall average (Dominguez-Catena et al., 2024; Gaya-Morey et al., 2026; Georgiou et al., 2026; T. Lian & Celiktutan, 2025).

Table 5 deepens the quantitative synthesis without averaging incompatible accuracy values across different datasets and label spaces. It reports the percentage of empirical or foundational sources in each architecture family that contained at least partial evidence for each TEC domain. Because the lightweight category has only one foundational source in the selected matrix, its percentages must be interpreted descriptively rather than as a stable population estimate.

Table 5
Comparative evidence-reporting trends by model family

Family	n	P (%)	G (%)	F (%)	R (%)	X (%)	D (%)	Median TEC (%)
CNN	4	100	50	0	75	0	25	21
CNN-ML	4	100	0	0	25	25	25	13
Transfer learning	5	100	60	0	60	0	20	33
Attention and transformer	3	100	100	0	100	100	100	50
Multimodal	3	100	33	0	100	0	67	33
Lightweight	1	100	100	0	100	0	100	58

Generalisation, bias, and robustness

Generalisation is one of FER’s most persistent weaknesses. A model may learn genuine facial features, but it may also exploit spurious properties of a dataset, including resolution, illumination, cropping, compression, or demographic composition. The cross-dataset generalisation gap can be expressed as

$$\Delta_{gen} = M(D_{train} \rightarrow D_{test}^{in}) - M(D_{train} \rightarrow D_{test}^{out}) \quad (6)$$

Where M is a measure such as macro-F1 or balanced accuracy, D_{test}^{in} belongs to the training domain, and D_{test}^{out} is an external dataset. A large Δ_{gen} indicates dependence on the original domain and weakens the case for deployment (Dominguez-Catena et al., 2022; Gaya-Morey et al., 2026; Georgiou et al., 2026; Mollahosseini et al., 2019).

FER datasets embody distinct forms of evidence. CK+ and JAFFE remain historically important because of their experimental control, but their size and diversity are limited. FER2013 and FER+ increased scale, although their images have low resolution and noisy labels. RAF-DB and AffectNet more closely resemble uncontrolled settings, yet they introduce web-sourced data bias, ambiguity, and class imbalance. Aff-Wild2 moves the task towards video and affect under naturalistic conditions. It follows that no single dataset can by itself establish generalisation (S. Li et al., 2017; Z. Lian et al., 2024; Lucey et al., 2010; Lyons et al., 1998).

Table 6
Relevant FER datasets and threats to external validity

Dataset	Type of evidence	Strength	Principal risk
CK+	Controlled sequences and action units.	Clean labels and prototypical expressions.	Limited environmental and demographic diversity.
JAFFE	Controlled static images.	Useful for historical comparison.	Small size and restricted population.
FER2013	Greyscale web images.	Widely used and readily reproducible benchmark.	Low resolution and label noise.
FER+	Relabelling through label distributions.	Represents annotation ambiguity.	Inherits the visual limitations of FER2013.
RAF-DB	Naturalistic images with basic and compound expressions.	Closer to uncontrolled settings.	Requires more complete demographic reporting.
AffectNet	Large-scale categories and affective dimensions.	Substantial visual variability.	Noise, web-source bias, and imbalance.
Aff-Wild2	Video, valence/arousal, and expressions.	Tests temporal information in naturalistic settings.	Costly protocols and sensitivity to preprocessing.

Demographic bias in FER may arise from unequal representation of age, sex, skin tone, ethnicity, or cultural context. It may also arise from differences in how annotators interpret expressions. A simple range statistic can describe group disparity:

$$\Delta_{bias} = \max_g M_g - \min_g M_g \quad (7)$$

Where M_g is the performance measure for group (g). A large Δ_{bias} shows that aggregated accuracy conceals unequal performance. The statistic does not resolve the ethical problem or establish a complete fairness criterion. Its more modest purpose is to require studies to disclose whether performance is comparable across relevant, ethically documented groups (Dominguez-Catena et al., 2022, 2024; T. Lian & Celiktutan, 2025; Mohammad, 2022).

Robustness concerns stability under realistic changes: occlusion, blur, poor illumination, compression, noise, pose, face coverings, inaccurate crops, and temporal variation. One simple robustness ratio compares clean and perturbed performance:

$$R_{rob} = 1 - \frac{M_{clean} - M_{perturbed}}{M_{clean} + \epsilon} \quad (8)$$

Where ϵ prevents division by zero. Values close to one indicate little degradation. The ratio must nevertheless be reported with the perturbation type and severity; otherwise, apparently similar robustness scores may describe tests of very different practical significance (Gaya-Morey et al., 2026; Georgiou et al., 2026; Z. Lian et al., 2024; Shirsath et al., 2022).

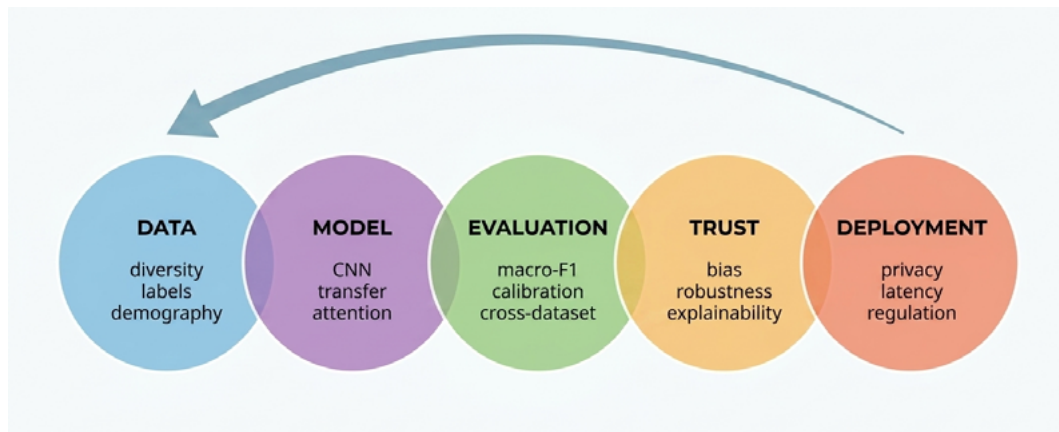
Explainability provides a diagnostic means of determining whether a model attends to plausible facial regions or to artefacts. Gradient-weighted class activation mapping (Grad-CAM), for example, estimates class-specific saliency from gradients with respect to activation maps. Such visualisations do not prove that a model ‘understands’ emotion. They can, however, reveal shortcuts: a happiness prediction driven by the background or a watermark would plainly undermine the model’s validity. Explanations should therefore accompany quantitative evaluation rather than replace it (Kim et al., 2022; Mao et al., 2025; Selvaraju et al., 2017; Zheng et al., 2023).

Explainability also needs an audit trail. A defensible FER audit should record dataset provenance and consent constraints; subgroup error and calibration; robustness tests and perturbation severity; explanation method and known failure modes; model or version changes; and deployment monitoring or incident feedback. Visual saliency is therefore one audit artefact rather than the audit itself. End-to-end algorithmic auditing frameworks reinforce the need to connect design choices, documentation, testing, governance, and post-deployment review (Raji et al., 2020).

Figure 3 synthesises the proposed trustworthiness framework. Its conceptual contribution is to connect data, model design, evaluation, trustworthiness, and deployment in a continuous chain. A failure observed after deployment should feed back into data curation, metrics, and validation protocols. Without that feedback, an FER system remains static even as its environment changes (European Parliament and Council of the European Union, 2024; T. Lian & Celiktutan, 2025; Mohammad, 2022; Wang et al., 2022).

Figure 3

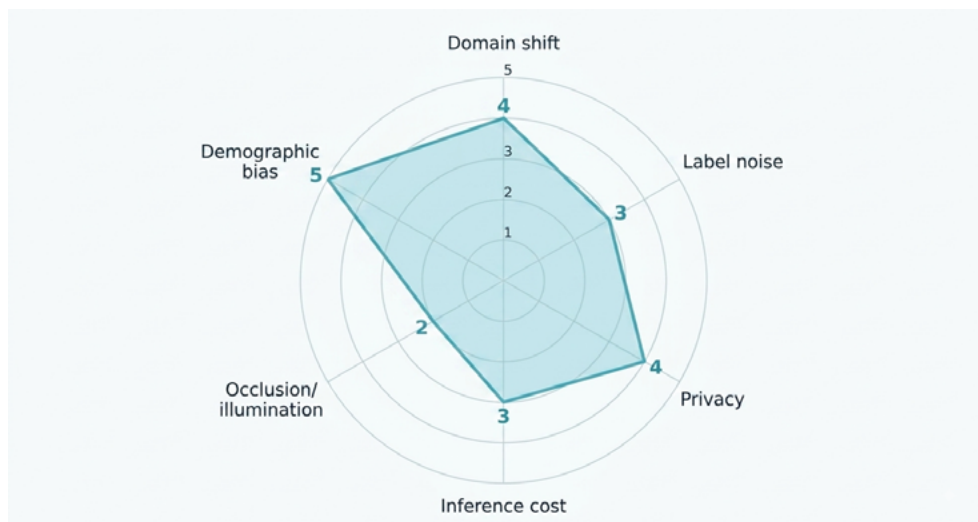
Proposed framework for evaluating FER beyond accuracy.



Note. The framework is merely conceptual: Table 3 applies its evidence domains source by source, Equation (1) provides a reproducible coverage measure, and Table 5 aggregates the resulting reporting patterns by model family.

Figure 4

Risks that reduce the external validity of FER models.



Note. The 1–5 values are an ordinal risk synthesis from the coded evidence and review discussion (1 = limited or less recurrent concern; 5 = recurrent or high-consequence concern). They are not probabilities or pooled effect sizes. The highest level assigned to demographic bias reflects its recurrence across dataset, fairness, and governance evidence; domain shift and privacy are also treated as high-impact external-validity risks.

Table 7
Minimum reporting matrix for trustworthy FER research.

Dimension	Required information	Control question
Data	Source, size, classes, distribution, preprocessing, and split.	Could another researcher reproduce the model input?
Model	Architecture, parameters, backbone, training, and data augmentations.	Is it clear what was trained and at what cost?
Metrics	Accuracy, macro-F1, confusion matrix, and class-wise measures.	Does overall performance conceal weak classes?
Generalisation	Cross-validation and cross-dataset validation were feasible.	Does the model work beyond its original domain?
Bias	Group-disaggregated performance where appropriately collected demographic data are available.	Is performance comparable across relevant groups?
Robustness	Perturbations, occlusion, illumination, compression, and pose.	Are predictions stable under realistic conditions?
Explainability	Grad-CAM, attention, or landmark analysis.	Does the model attend to plausible facial regions?
Deployment	Latency, memory, privacy, consent, and regulation.	Is the proposed use technically and ethically defensible?

Lightweight models are necessary for mobile applications, low-cost robotics, and privacy-sensitive local inference. MobileNet, quantisation, pruning, and distillation can reduce parameter counts and latency. Efficiency, however, is not an independent virtue. A lightweight model that disproportionately reduces performance for minority classes or underrepresented groups may be less trustworthy than a larger but more balanced model. Computational cost, fairness, and robustness must therefore be reported together (European Parliament and Council of the European Union, 2024; T. Lian & Celiktutan, 2025; Mao et al., 2025; Sandler et al., 2018).

Discussion

The strongest FER studies are no longer those that report accuracy alone. A marginal improvement on a benchmark may be scientifically less informative than carefully documented cross-dataset validation, group-level bias analysis, or a realistic perturbation test. CNNs and transfer learning remain strong baselines, whilst attention, transformers, and multimodality provide useful means of addressing ambiguity and context. Their value depends, however, on evaluation across diverse settings rather than internal data splits alone (Gaya-Morey et al., 2026; Georgiou et al., 2026; S. Li & Deng, 2022; Wang et al., 2022).

The quantitative pattern in Table 5 clarifies the imbalance. Performance evidence is nearly universal in the selected architecture papers, whereas explicit fairness reporting is absent from the subset architecture studies and is instead concentrated in dedicated bias studies. Attention and transformer papers in the matrix show broader coverage of generalisation, robustness, explainability, and deployment evidence, but the family contains only three selected sources; the finding should therefore be read as a reporting trend, not a claim that transformers are intrinsically more trustworthy. This is precisely why TEC is retained as an evidence-coverage descriptor rather than a league-table score.

An objection must nevertheless be considered. Broader evaluation is costly: external datasets may use incompatible labels, demographic attributes may be unavailable or inappropriate to infer, and robustness or latency tests may demand resources that smaller teams do not possess. It would therefore be misleading to treat every missing measure as evidence of a defective

model. Yet this practical constraint does not justify silent omission. When an evaluation cannot be conducted, the limitation should be stated, and claims should be narrowed accordingly. Methodological restraint is itself part of trustworthy reporting.

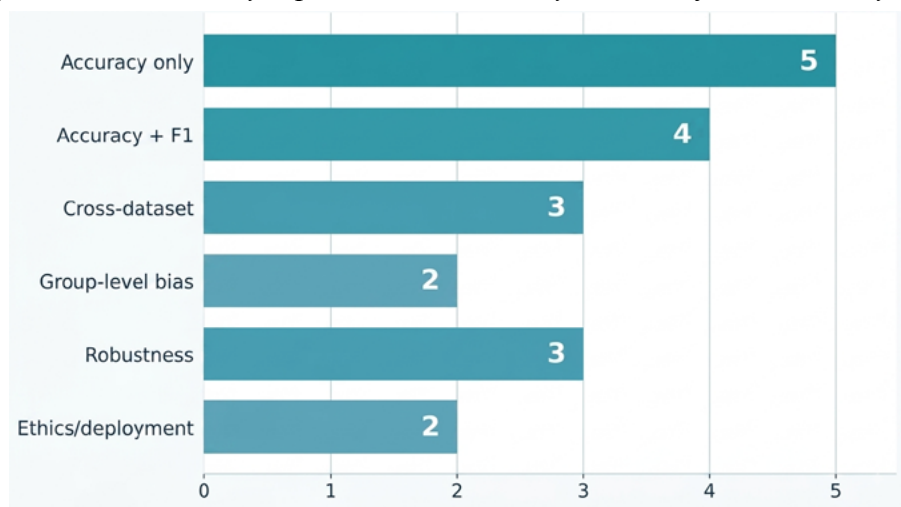
The synthesis that follows is thus not a demand for a single universal score but for evidence proportionate to the claim. The guiding question should change from “Which model has the highest accuracy?” to “What evidence warrants confidence in this model when the data, population, and context change?” This shift turns a descriptive comparison into a critical evaluation. The proposed framework does not claim a new state of the art; it specifies the conditions under which comparisons between FER systems become scientifically defensible (Barrett et al., 2019; Dominguez-Catena et al., 2024; T. Lian & Celiktutan, 2025; Mohammad, 2022).

This restraint is particularly important because the validity of the target construct is contested. Barrett et al. (2019) show that the same facial movement can occur in different emotional contexts and that an emotion category can be expressed through heterogeneous facial configurations. Consequently, high classification accuracy against a dataset label establishes agreement with that annotation scheme under that protocol; it does not establish direct access to a person’s internal emotional state. Context, temporal information, culture, elicitation procedure, annotator uncertainty, and alternative non-facial signals should therefore delimit the language used in claims. In high-stakes settings, FER should not be treated as a standalone emotion detector without independent contextual validation.

The same reasoning applies to practical use. FER may support adaptive interfaces, personalised education, social robotics, digital assistance, and well-being applications when consent, transparency, and proportionality are established. The same technology can also enable emotion surveillance, workplace monitoring, or sensitive classification without adequate context. The European Union’s Artificial Intelligence Act places significant restrictions on emotion-recognition systems, particularly in education and employment, subject to limited medical or safety exceptions. Ethics and regulation must consequently form part of system design and evaluation rather than an ornamental final section (European Parliament and Council of the European Union, 2024; Z. Lian et al., 2024; Mohammad, 2022; Wang et al., 2022).

Figure 5

Gap between commonly reported and necessary evidence for trustworthy FER.



Note. The figure uses the same 1–5 ordinal synthesis logic as Figure 2: higher values denote evidence types that are more routinely reported in the reviewed literature. The source-level matrix in Table 3 supplies the audit trail; the bars are an interpretive scale of reporting frequency, not raw study counts.

Figure 5 illustrates a recurrent asymmetry: accuracy is reported far more often than robustness, fairness, or deployment evidence. This imbalance explains why some studies appear strong in their results yet weak in external validity. The minimum reporting matrix offers a remedy by aligning each type of claim with the evidence needed to support it (Dominguez-Catena et al., 2022; Gaya-Morey et al., 2026; Georgiou et al., 2026; Page et al., 2021).

Table 8
Practical implications for the evaluation and reporting of FER studies.

Aspect	Risk if omitted	Recommended practice
Citations	Claims rely on weak or untraceable evidence.	Distinguish peer-reviewed studies, foundational work, and technical preprints.
Comparability	Scores from incompatible protocols appear equivalent.	Report splits, preprocessing, label mapping, and uncertainty.
Method	The review or experiment cannot be reproduced.	State sources, time window, criteria, coding, and validation protocol.
Contribution	Architectural novelty is confused with stronger evidence.	Relate each claimed advance to generalisation, robustness, or fairness.
Ethics	Technical performance obscures potential harm.	Address privacy, consent, bias, intended use, and applicable regulation.
Resources	Efficiency claims lack an operational basis.	Report latency, memory, hardware, parameters, and energy were feasible.

This review has four principal limitations. First, the search may have missed recent unindexed studies or literature behind subscription barriers. Second, several preprints provide timely evidence but do not carry the same epistemic weight as peer-reviewed articles. Third, comparisons across datasets remain imperfect because protocols and label spaces vary. Finally, many datasets lack complete and ethically collected demographic documentation, which restricts bias analysis. These limitations do not invalidate the synthesis, but they delimit its claims and reinforce the need for transparent reporting (Dominguez-Catena et al., 2024; Gaya-Morey et al., 2025; Georgiou et al., 2026; Mohammad, 2022).

Two additional limitations follow from the analysis. First, TEC measures reporting coverage, not correctness, causal validity, or deployment safety; a paper can report all six domains and still reach an invalid conclusion. Secondly, quality and TEC coding necessarily compress heterogeneous evidence into ordinal categories. The full source-level matrix is therefore provided so that readers can inspect the coding rather than rely on the aggregate index alone. Future updates should use preregistered coding rules and, where multiple coders are available, report inter-rater agreement.

Conclusions

This review has shown that recent FER models have advanced substantially in benchmark performance, yet a consequential gap remains between accuracy and real-world trustworthiness. CNNs, transfer learning, hybrid models, attention, transformers, multimodality, and lightweight architectures each provide useful capabilities; no family, however, guarantees generalisation, fairness, or robustness. Such properties must be demonstrated through complementary metrics, cross-dataset validation, group-disaggregated analysis, perturbation tests, explanations, and explicit deployment conditions (Georgiou et al., 2026; S. Li & Deng, 2022; T. Lian & Celiktutan, 2025; Wang et al., 2022).

The proposed trustworthy-FER framework answers the research question by connecting methodological structure, evaluation measures, and reporting requirements. Its central claim is deliberately restrained: aggregated accuracy remains informative, but it cannot bear the full weight of a deployment claim. Trust depends on the relationship between data, model, evaluation, human context, and operational constraints (Dominguez-Catena et al., 2022, 2024; European Parliament and Council of the European Union, 2024; Mohammad, 2022).

The operationalisation makes this claim testable at the level of evidence reporting. TEC quantifies whether a study documents performance, generalisation, fairness, robustness, explainability, and deployment evidence, while the quality appraisal distinguishes complete reporting from weak methodological description. Neither quantity should be interpreted as a universal trustworthiness certificate. Their purpose is to make missing evidence visible and to align the strength of a claim with the evidence supplied.

Future work should therefore pursue six priorities: more diverse, documented, and ethically collected datasets; routine cross-dataset evaluation; macro-F1, calibration, and group-disaggregated reporting; robustness tests based on realistic perturbations; explainability and traceability in human-facing applications; and validation of lightweight models without sacrificing fairness. Progress on these fronts would move the field from systems that recognise expressions within benchmarks towards affective computing systems whose limitations are measurable, transparent, and responsibly governed (Gaya-Morey et al., 2025; Gaya-Morey et al., 2026; Z. Lian et al., 2024; Page et al., 2021).

Acknowledgements and Declarations

Disclosure of Interests

The authors have no competing interests to declare.

Author Contributions

The authors declare equal contribution and sharing of authorship roles for this publication.

AI Usage Statement

Generative artificial intelligence tools were used to support language editing, outline refinement, and bibliographic normalisation. The authors remain responsible for verifying all screening decisions, interpretations, claims, and final wording, and assume full responsibility for the manuscript's content.

References

- Agrawal, I., Kumar, A., Swathi, D., Yashwanthi, V., & Hegde, R. (2021). Emotion recognition from facial expression using CNN. *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)*, 1–6. <https://doi.org/10.1109/R10-HTC53172.2021.9641578>
- Bahmanabady, E., Imani, M., & Ghassemian, H. (2023). Fusion of multi-level CNN with LBP features for facial emotion recognition. *2023 31st International Conference on Electrical Engineering (ICEE)*, 463–467. <https://doi.org/10.1109/ICEE59167.2023.10334654>
- Banoth, S., Kukreja, V., Thapliyal, N., Aeri, M., & Sharma, R. (2024). Beyond basic emotions: High-accuracy facial expression classification using a CNN-SVM hybrid model. *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)*, 1–6. <https://doi.org/10.1109/ICETCS61022.2024.10544213>

- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1), 1–68. <https://doi.org/10.1177/1529100619832930>
- Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 279–283. <https://doi.org/10.1145/2993148.2993165>
- Dominguez-Catena, I., Paternain, D., & Galar, M. (2022). *Assessing demographic bias transfer from dataset to model: A case study in facial expression recognition*. <https://doi.org/10.48550/arXiv.2205.10049>
- Dominguez-Catena, I., Paternain, D., & Galar, M. (2024). Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8), 5209–5226. <https://doi.org/10.1109/TPAMI.2024.3361979>
- European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act)*. Official Journal of the European Union, L 2024/1689. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- Gaya-Morey, F. X., Buades-Rubio, J. M., Palanque, P., Lacuesta, R., & Manresa-Yee, C. (2025). *Deep learning-based facial expression recognition for the elderly: A systematic review*. <https://doi.org/10.48550/arXiv.2502.02618>
- Gaya-Morey, F. X., Manresa-Yee, C., Martinie, C., & Buades-Rubio, J. M. (2026). Evaluating facial expression recognition datasets for deep learning: A benchmark study with novel similarity metrics. *IEEE Transactions on Affective Computing*. <https://doi.org/10.1109/TAFFC.2026.3693316>
- Georgiou, S., Psiris, A., Evangelatos, S., Lagkas, T., Argyriou, V., Sarigiannidis, P., Varlamis, I., & Papadopoulos, G. Th. (2026). *Facial expression recognition in the deep learning era: A systematic multi-criteria review of methods, models, datasets, performance, challenges, and future research directions*. <https://doi.org/10.48550/arXiv.2606.08612>
- Goodfellow, I. J., Erhan, D., Carrier, P.-L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., & others. (2013). *Challenges in representation learning: A report on three machine learning contests*. arXiv. <https://doi.org/10.48550/arXiv.1307.0414>
- Gupta, V., & Sejwar, V. (2022). Facial expression recognition with combination of geometric and textural domain features extractor using CNN and machine learning. *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 400–407. <https://doi.org/10.1109/ICAISS55157.2022.10010853>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Karatay, B., Bestepe, D., Sailunaz, K., Ozyer, T., & Alhajj, R. (2022). A multimodal emotion recognition system based on CNN-transformer deep learning technique. *2022 7th International Conference on Data Science and Machine Learning Applications (CDMA)*, 145–150. <https://doi.org/10.1109/CDMA54072.2022.00029>
- Khairuddin, Y., & Chen, Z. (2021). *Facial emotion recognition: State of the art performance on FER2013*. <https://doi.org/10.48550/arXiv.2105.03588>
- Khajuria, O., Kumar, R., & Gupta, M. (2023). Facial emotion recognition using CNN and VGG-16. *2023 International Conference on Inventive Computation Technologies (ICICT)*, 472–477. <https://doi.org/10.1109/ICICT57646.2023.10133972>
- Kim, J.-H., Kim, N., & Won, C. S. (2022). *Facial expression recognition with swin transformer*. <https://doi.org/10.48550/arXiv.2203.13472>

- Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3), 1195–1215. <https://doi.org/10.1109/TAFFC.2020.2981446>
- Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2584–2593. <https://doi.org/10.1109/CVPR.2017.277>
- Li, T.-H. S., Kuo, P.-H., Tsai, T.-N., & Luan, P.-C. (2019). CNN and LSTM based facial expression analysis model for a humanoid robot. *IEEE Access*, 7, 93998–94011. <https://doi.org/10.1109/ACCESS.2019.2928364>
- Lian, T., & Celiktutan, O. (2025). *A feature-level bias evaluation framework for facial expression recognition models*. <https://doi.org/10.48550/arXiv.2505.20512>
- Lian, Z., Sun, H., Sun, L., Wen, Z., Zhang, S., Chen, S., Gu, H., Zhao, J., Ma, Z., Chen, X., Yi, J., Liu, R., Xu, K., Liu, B., Cambria, E., Zhao, G., Schuller, B. W., & Tao, J. (2024). MER 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, 41–48. <https://doi.org/10.1145/3689092.3689959>
- Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 94–101. <https://doi.org/10.1109/CVPRW.2010.5543262>
- Lyons, M. J., Akamatsu, S., Kamachi, M., & Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, 200–205. <https://doi.org/10.1109/AFGR.1998.670949>
- Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., & Huang, A. (2025). POSTER++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, 157, 110951. <https://doi.org/10.1016/j.patcog.2024.110951>
- Mohammad, S. M. (2022). Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2), 239–278. https://doi.org/10.1162/coli_a_00433
- Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1), 18–31. <https://doi.org/10.1109/TAFFC.2017.2740923>
- Negi, A. S., Arora, A., Bisht, S., Devliyal, S., Kumar, B. V., & Kaur, G. (2024). Facial emotion detection using CNN and VGG16 model. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–5. <https://doi.org/10.1109/I2CT61223.2024.10543618>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Poulose, A., Kim, J.-H., & Han, D. S. (2021). Feature vector extraction technique for facial emotion recognition using facial landmarks. *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, 1072–1076. <https://doi.org/10.1109/ICTC52510.2021.9620798>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>
- Rawat, U., & Rai, C. S. (2023). Improving facial emotion recognition through transfer learning with deep convolutional neural network models. *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 1335–1339. <https://doi.org/10.1109/ICTACS59847.2023.10390186>

- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- Sharafi, M., Yazdchi, M., & Rasti, J. (2023). Audio-visual emotion recognition using k-means clustering and spatio-temporal CNN. *2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA)*, 1–6. <https://doi.org/10.1109/IPRIA59240.2023.10147192>
- Shirsath, S., Vikhe, V., & Vikhe, P. (2022). Xception CNN-ensemble learning based facial emotion recognition. *2022 6th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 1–4. <https://doi.org/10.1109/ICCUBEA54992.2022.10011111>
- Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83–84, 19–52. <https://doi.org/10.1016/j.inffus.2022.03.009>
- Zheng, C., Mendieta, M., & Chen, C. (2023). POSTER: A pyramid cross-fusion transformer network for facial expression recognition. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 3138–3147. <https://doi.org/10.1109/ICCVW60793.2023.00339>