JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CNN-Based Visual and Facial Deception Detection: A Structured Review of Facial Expression, Micro-Expression, and Behavioral Cue Analysis

Priya R. Ghongade, Prof.Vijaykumar M.P
Department of Computer Science and Engineering, Shreeyash College of Engineering & Technology, Chh. Sambhajinagar, India
Priyaghongade12@gmail.com

Abstract—Automated deception detection seeks to infer the veracity of human communication from observable behavioural signals rather than from invasive physiological instrumentation. Conventional polygraphy, while historically dominant, is confounded by its reliance on autonomic arousal, its intrusiveness, and its contested scientific validity. The emergence of deep learning has enabled non-invasive analysis of facial expressions and micro-expressions using convolutional ne

ural networks (CNNs), motivating sustained research interest in visual, facial-video-based deception analysis. This review synthesizes the state of the art in CNN-based visual and facial deception detection, examining spatial feature extraction, temporal extensions such as CNN-LSTM architectures for facial expression sequences, and the wider behavioural-science literature on facial action units, micro-expressions, and eye gaze/blink dynamics that underpins this line of research. Widely used facial-emotion resources such as FER2013 and CK+ are discussed and explicitly distinguished from purpose-built deception corpora with a visual/facial channel, including the Real-Life Trial dataset. The review develops a taxonomy of CNN-based visual/facial approaches, synthesizes findings from the literature into a comparative table, and identifies persistent research gaps concerning dataset scarcity, cross-domain generalization, the loss of temporal information in static-frame CNN analysis, explainability, demographic and cultural bias, and real-time deployment. A recurring theme is that facial and behavioural arousal are correlates of stress or cognitive load rather than direct markers of falsehood, and this distinction is treated as a central methodological caution throughout. A conceptual, literature-grounded CNN-based visual pipeline is proposed as a direction for future research rather than as a validated system. The review concludes that CNN-based facial analysis is best positioned as a probabilistic, explainable, human-in-the-loop decision-support approach for deception-related behavioural cues, not as a definitive lie detector.

Keywords—*Deception detection; convolutional neural networks; facial expression recognition; facial micro-expressions; Facial Action Units; visual behavioural analysis; deep learning; explainable artificial intelligence*

I. INTRODUCTION

Deception, defined in behavioural science as the deliberate act of creating a false belief in another party without their consent, has long been of interest to security agencies, forensic investigators, human-resource practitioners, and designers of online trust systems [7], [8]. Computational deception detection reframes this problem as a pattern-recognition task: given observable facial and visual signals produced during communication, infer the probability that the communicated content is false. This reframing is attractive because it promises scalability and non-intrusiveness, but it inherits a difficulty that predates computing — there is no single, reliable physiological or behavioural signature of lying [8], [9].

Traditional approaches to instrumented lie detection, most notably the polygraph, infer deception indirectly from autonomic arousal indices such as skin conductance, respiration, and cardiovascular activity. Authoritative reviews of this literature have repeatedly concluded that polygraph accuracy is highly variable, sensitive to examiner protocol, and vulnerable to countermeasures, which limits its scientific and legal standing [10]. Because arousal-based measurement also requires attached sensors and a controlled setting, it is unsuitable for naturalistic, remote, or large-scale screening scenarios.

The growth of deep learning over the past decade has enabled a parallel, non-invasive research direction: inferring behavioural correlates of deception directly from facial video and imagery using convolutional neural networks (CNNs), with temporal extensions such as CNN-LSTM architectures used to capture how an expression evolves across a sequence of frames [16]. These approaches draw on decades of behavioural research into facial action units, micro-expressions, and gaze and blink dynamics [11] – [13], and re-implement the extraction of such cues as learned representations rather than hand-engineered rules.

Although facial, vocal, linguistic, and physiological behavioural channels have each been investigated as evidentiary bases for automated deception analysis, and although prior work has explored combining several of them, visual and facial analysis constitutes one of the most extensively studied and most mature branches of this literature, owing to its long-standing behavioural-science foundation in facial action coding and micro-expression research, and to the wide availability of facial-emotion datasets such as FER2013 and CK+ [14], [15]. This review accordingly narrows its focus specifically to CNN-based visual and facial deception detection, treating vocal, linguistic, physiological, and combined-modality analysis as related but out-of-scope research directions that are noted only where necessary for context.

A second and equally important theme running through this body of work is a conceptual caution rather than a technical one: nervousness, stress, and cognitive load are not equivalent to deception. A truthful respondent under scrutiny may exhibit elevated facial arousal, while a practiced deceiver may remain behaviorally composed [8], [9], [12]. Any system that treats facial or behavioural arousal as direct proof of lying risks scientifically unsound and ethically consequential conclusions. This review treats that distinction as a first-class research concern rather than a footnote.

Scope and Contributions of This Review

This paper presents a structured, narrative review of CNN-based visual and facial deception detection. Its contributions are: (i) a taxonomy that organizes CNN-based visual/facial approaches by the specific facial cue exploited — expression, micro-expression, Facial Action Unit, and gaze/blink-based analysis — and by architectural variant (static-frame CNN versus temporal CNN-LSTM extension); (ii) a critical, literature-grounded review of CNN architectures for facial deception-relevant analysis, including their reported strengths and limitations, alongside a brief comparative note on the Vision Transformer (ViT) as an alternative visual architecture; (iii) a clear separation of emotion-recognition datasets (e.g., FER2013, CK+) from purpose-built deception corpora with a usable visual/facial channel; (iv) a synthesis of research gaps spanning dataset scarcity, generalization, the loss of temporal information in frame-level analysis, explainability, bias, real-time deployment, and ethics; and (v) a conceptual, literature-derived CNN-based visual pipeline proposed as a direction for future empirical validation rather than as a completed or benchmarked system. Consistent with its designation as a review, this paper does not report new experimental results; where a source presentation reported specific accuracy figures, those figures are explicitly attributed to that source rather than treated as verified benchmarks.

II. Review Methodology

This work is a structured narrative review rather than a formal systematic review conducted under a pre-registered protocol; accordingly, no exact count of screened records is reported, consistent with the absence of a registered systematic search. The review was organized around the central theme of non-invasive, CNN-based analysis of facial and visual correlates of deception.

Candidate literature was identified through the digital libraries and indexing services of IEEE Xplore, SpringerLink, ScienceDirect, the ACM Digital Library, arXiv, and general Scopus-indexed and Web-of-Science-indexed venues, using combinations of the keywords: deception detection, lie detection, facial micro-expressions, facial action units, facial expression recognition, convolutional neural network, CNN-LSTM, vision transformer, and explainable AI.

Inclusion favored peer-reviewed conference and journal papers, systematic reviews, and widely cited preprints published between 2009 and 2024, together with foundational works establishing datasets (e.g., FER2013 [14], CK+ [15]) or architectures (e.g., CNN [21], LSTM [16]) that remain in active use for visual/facial analysis. Studies were excluded where their primary contribution concerned audio-only, text-only, or physiological-only deception detection without a facial/visual component, or where bibliographic details could not be reasonably corroborated; such studies are referenced only where necessary for behavioural-science background. The literature is classified along two dimensions used throughout the remainder of the paper: (a) the specific facial/visual cue exploited, and (b) the CNN-based architectural variant used to model it, which together define the taxonomy presented in Section III.

III. Taxonomy of CNN-Based Visual/Facial Deception Detection

As illustrated in Fig. 1, CNN-based visual/facial deception detection approaches can be organised into five closely related sub-areas based on the specific facial cue exploited, together with a common architectural distinction between static (frame-level) and temporal (sequence-level) CNN models.

A. Facial Expression Recognition

Facial expression recognition uses a CNN to classify coarse emotion categories from static facial imagery, and serves as a foundation on which more deception-specific analysis is built. FER2013 is the dataset most widely used to pre-train or benchmark CNNs for this purpose [14].

B. Micro-Expression Analysis

Micro-expressions are brief, largely involuntary facial expressions lasting a fraction of a second, and are of particular interest in deception research because they are considered more difficult to consciously suppress than a posed expression. Deep CNN approaches have been reported to improve the detection of such transient facial cues relative to classical machine-learning baselines.

C. Facial Action Units (FACS)-Based Analysis

Facial Action Unit analysis decomposes an expression into anatomically grounded, individually coded muscle movements defined by the Facial Action Coding System [13], and CNNs have been applied to localise and classify these units directly from facial imagery as a more granular alternative to whole-expression classification.

D. Eye Gaze and Blink Dynamics

Blink frequency and gaze dynamics have separately been studied as liveness, attention, and behavioural indicators extracted from facial image sequences [20], and are a natural complement to CNN-based expression and action-unit analysis within a purely visual pipeline.

E. Temporal Extension: CNN-LSTM for Facial Sequences

Where a facial video sequence rather than a single frame is available, a CNN front end can be paired with an LSTM temporal head to model the onset, apex, and offset of an expression over time, an architecture historically trained on sequence datasets such as CK+ [15], [16]. This temporal extension is discussed further as an architectural variant in Section IV-B.

Fig. 1. Taxonomy of CNN-Based Visual/Facial Deception Detection Approaches

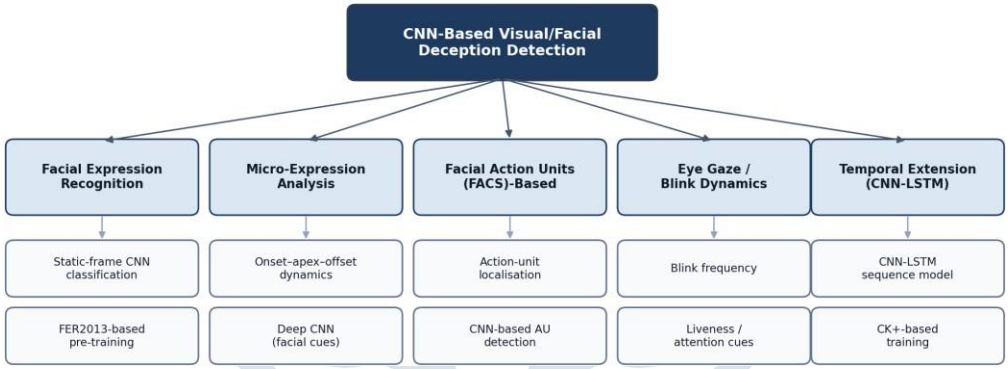

Fig. 1. Taxonomy of CNN-based visual/facial deception detection approaches, organised by facial cue and representative techniques.

IV. CNN Architectures for Visual/Facial Deception Detection

A. Convolutional Neural Networks

CNNs remain the dominant architecture for spatial feature extraction from facial imagery, learning hierarchical filters for edges, facial parts, and expression-relevant configurations without hand-crafted features [14], [21]. Their principal strength is translation-invariant spatial feature learning at moderate computational cost; their principal limitation for deception analysis is that a CNN applied to individual frames cannot, by itself, capture the temporal transitions between expressions that carry much of the discriminative signal in micro-expression research.

B. Temporal Extension: CNN-LSTM for Facial Sequences

LSTM and related gated recurrent architectures address the temporal limitation of frame-wise CNNs by modelling sequential dependencies across a sequence of facial descriptors, and are frequently paired with a CNN front end in a CNN-LSTM configuration capable of learning both spatial facial features and temporal emotional transitions, as in CK+-based pipelines [15], [16]. The advantage of this configuration is explicit modelling of an expression's onset, apex, and offset; its limitation is greater computational and data requirements than a static CNN alone, together with difficulty scaling to long sequences.

C. Note on an Alternative Visual Architecture: Vision Transformer (ViT)

Beyond CNN and CNN-LSTM pipelines, Vision Transformer (ViT) architectures have also been applied to facial imagery, replacing convolutional filters with patch-based self-attention to capture global facial context [17]. Work applying ViT to facial data for deception-relevant classification has reported potential accuracy gains over CNN baselines, though such models typically require larger annotated facial datasets and greater computational resources than CNN or CNN-LSTM alternatives. Because the focus of this review is CNN-based analysis, ViT is discussed here only as a comparative point of reference (see Table I) rather than as a core architecture of this paper.

V. Datasets and Benchmarks

A recurring and scientifically important distinction in this literature is between datasets constructed for general facial-emotion recognition and datasets constructed specifically to study deception. The former are useful for pre-training or transfer learning but do not themselves contain ground-truth veracity labels.

A. Facial Emotion-Recognition Datasets

FER2013 comprises approximately 35,000 grayscale, 48×48-pixel facial images labelled with one of seven basic emotion categories, collected via web image search and widely used to pre-train or benchmark facial-emotion CNNs [14]. CK+ extends the original Cohn-Kanade corpus with posed and spontaneous expression sequences from multiple subjects, annotated with Facial Action Units and peak emotion labels, and is correspondingly used to train temporal CNN-LSTM models of expression transition [15]. Neither dataset contains labels indicating whether a subject was being truthful or deceptive; their relevance to deception detection is limited to providing transferable facial-representation features, and any accuracy figure obtained on these corpora describes emotion classification, not lie detection.

B. Deception-Specific Corpora with a Visual/Facial Channel

Purpose-built deception datasets with a visual/facial channel are comparatively scarce and smaller in scale. The Real-Life Trial dataset comprises video clips of witnesses and defendants in real courtroom proceedings, manually labelled as truthful or deceptive based on trial outcomes, and is one of the few corpora with ecologically valid, high-stakes deception captured on video [2]. Bag-of-Lies additionally provides a video and gaze channel, collected under a controlled visual description task with an experimentally induced ground truth, alongside other physiological channels that fall outside a purely visual analysis [4]. Box-of-Lies, derived from a televised social deduction game, likewise includes a video channel usable for facial analysis, although it was designed and is more commonly used as a multimodal dialogue corpus rather than a dedicated visual benchmark [5]. These corpora are markedly smaller than mainstream computer-vision benchmarks, and most were collected in a single cultural, linguistic, and recording context, which materially limits the external validity of models trained on any single one of them.

VI. Critical Literature Review

Table I summarizes a representative cross-section of the reviewed literature that has a usable visual or facial component, spanning survey work, static and temporal CNN architectures, and a comparative ViT-based study. Because published evaluation protocols, splits, and metrics vary substantially across these studies, quantitative accuracy values are reported only where directly and reliably attributable to a specific cited source; qualitative findings are reported elsewhere.

TABLE I. Summary of Representative Literature on CNN-Based Visual/Facial Deception Detection (rows marked "—" correspond to studies cited in the source presentation and could not be independently re-verified in this review; treat with appropriate caution)

Ref.	Author/Year	Visual Cue / Focus	Dataset	Method	Major Finding	Limitation
[1]	Constâncio et al., 2023	Survey (general ML/DL)	N/A (review)	ML/DL survey — statistical analysis	Synthesises ML techniques for deception detection; underscores the value of behavioural-cue analysis, including facial/visual cues	Survey only; no new experiments
[2]	Pérez-Rosas et al. 2015	Facial video (courtroom)	Real-Life Trial dataset	Classical + deep visual features from courtroom video	Provides high-stakes, real, ecologically valid deception video with usable facial signal	Small sample; single legal/cultural context
[3]	Wu et al., 2018	Facial/video	Real-world video clips	Deep video-based model (visual features central)	Reports strong contribution of visual/facial cues to deception classification on real videos	Limited dataset diversity and scale
[4]	Gupta et al., 2019	Facial video + gaze	Bag-of-Lies	Deep visual/gaze feature analysis	Introduces a corpus with a usable video and gaze channel for visual deception analysis	Controlled task-based elicitation; limited ecological validity

Ref.	Author/Year	Visual Cue / Focus	Dataset	Method	Major Finding	Limitation
[6]	Mathur & Matarić 2020	Facial affect	Interview video data	CNN-based facial-affect representation	Facial affect representations derived from CNN features add discriminative value	Facial affect ≠ deception; correlational cue only
—	Dewan et al., 2021 (per source PPT)	Facial micro-expression	Facial micro-expression data	Deep CNN	Reported improvement over classical ML for micro-expression-based cues	Sensitive to lighting and facial occlusion; unverified externally
—	Yang et al., 2022 (per source PPT)	Facial image (alternative arch.)	Facial image data	Vision Transformer (ViT)	Reported improvement in spatial facial feature capture over CNN baselines	Requires large annotated datasets; unverified externally
[20]	Kollreider et al., 2009	Eye gaze / blink	Face image sequences	Non-intrusive liveness/gaze analysis	Blink frequency and gaze dynamics provide behavioural indicators usable alongside CNN facial analysis	Degraded by eyewear and rapid movement

As Table I shows, the reviewed visual/facial literature separates broadly into three groups: (i) survey work that synthesises existing findings without new experimentation [1]; (ii) CNN-based systems operating on facial video or images, whether drawn from courtroom, real-world, or task-elicited settings [2]–[4], [6]; and (iii) studies of complementary visual cues, such as gaze and blink dynamics, or of an alternative visual architecture such as ViT, that inform or contextualise CNN-based analysis without replacing it as the review's central focus [20].

Critical Synthesis

Static frame-level CNN versus CNN-LSTM. Purely spatial CNN models perform adequately on static, frame-level facial classification but are structurally unable to capture the emergent temporal pattern of a micro-expression, i.e., its onset, apex, and offset; CNN-LSTM hybrids close part of this gap at the cost of additional data and computation [15], [16]. CNN versus ViT. Where sufficient annotated facial data are available, Vision Transformer models have been reported to outperform CNN-based baselines on facial classification tasks by modelling global context through self-attention, but this advantage is contingent on a dataset scale that deception-specific facial corpora, as noted in Section V, generally do not provide [17].

Static images versus video sequences. Deception-relevant information such as micro-expression transitions is inherently temporal; static-image approaches necessarily discard this information, which is a key reason the literature has moved from static facial classifiers toward CNN-LSTM sequence models for expression-transition analysis [15], [16].

Controlled versus real-world data, and why high controlled-setting accuracy may not transfer. A substantial share of the reviewed accuracy figures were obtained on data collected under controlled laboratory conditions with scripted or experimentally induced deception, frontal camera angles, and a single demographic and cultural population. Real-world deception — in interviews, courtrooms, or online interaction — involves spontaneous, high-stakes lying, variable pose and lighting, occlusion, and substantial demographic diversity. Because deep models trained under one regime tend to overfit to regime-specific artefacts, a high accuracy figure obtained in a constrained laboratory setting is not evidence that the same CNN model will generalize to naturalistic deployment; cross-corpus evaluation is required, and remains uncommon [2], [4], [5]. Accuracy versus generalisation and computational cost. Larger and more expressive visual architectures (e.g., ViT) can improve in-distribution accuracy but typically increase computational cost, reduce interpretability, and, absent explicit regularisation and cross-corpus validation, do not automatically improve out-of-distribution generalisation.

VII. Research Gaps and Challenges

A. Dataset Gap

Deception-specific datasets with a usable visual/facial channel remain small, narrow in demographic and cultural coverage, and often collected under a single experimental protocol, in contrast to the scale of general-purpose vision

corpora [2], [4], [5]. This scarcity constrains both model capacity and the statistical reliability of reported accuracy figures.

B. Generalisation Gap

CNN models trained on one facial dataset frequently show reduced performance when evaluated on an independently collected dataset, reflecting overfitting to protocol-specific and population-specific artefacts rather than to generalizable facial deception cues.

C. Temporal Information Loss Gap

A substantial share of CNN-based deception studies analyses single static frames or short clips, which discards the onset–apex–offset dynamics that behavioural research identifies as informative for micro-expression analysis; the CNN-LSTM extension discussed in Section IV-B addresses this only partially, and at additional computational and data cost.

D. Explainability Gap

Deep CNN classifiers are largely opaque; a deception-probability output without accompanying visual evidence is difficult to scrutinize, contest, or trust in a high-stakes application, which motivates the explainability discussion in Section IX.

E. Bias Gap

Demographic factors (age, gender, skin tone), cultural display rules for emotion, and recording conditions (camera, compression, lighting) can all systematically shift CNN model behaviour, and most existing visual deception corpora are too narrow to characterize, let alone correct for, such effects [22].

F. Real-Time Gap

CNN and CNN-LSTM pipelines applied to high-resolution facial video incur meaningful inference latency and computational load, which is in tension with real-time or edge-deployment requirements.

G. Ethical Gap

Automated facial deception scoring raises questions of consent, surveillance, privacy of facial biometric data, and the risk of consequential false accusations; these are treated in depth in Section IX.

H. Deception Is Not Emotion: A Necessary Distinction

The behavioural-science literature underlying this field is explicit that facial expression, stress, and cognitive load are correlates of arousal, not direct markers of falsehood [8], [9], [12]. A truthful person questioned under suspicion may show fear, frustration, or nervous arousal indistinguishable, at the level of a facial action unit, from the arousal of a person who is lying; conversely, a practiced or low-anxiety deceiver may remain behaviourally composed and evade detection by cue-based systems entirely [8]. Meta-analytic behavioural work has found that even trained human observers perform only marginally better than chance at distinguishing truth from lies from behavioural cues alone, and that few, if any, cues are reliably and uniquely diagnostic of deception across contexts [8], [12]. Any CNN-based system in this space should therefore be described, and evaluated, as detecting facial or behavioural correlates associated with deception in a given dataset and protocol — a probabilistic signal — rather than as directly detecting the truth or falsehood of a statement. This distinction is treated in this review as a necessary constraint on both the framing of results and the design of any deployed system.

VIII. PROPOSED CONCEPTUAL CNN-BASED VISUAL FRAMEWORK

Building on the taxonomy in Section III and the architectural review in Section IV, this section synthesizes a conceptual, literature-grounded CNN-based visual pipeline intended to guide future empirical work. It is presented as a research

direction, consistent with modules described at a conceptual level in the source presentation underlying this review, and not as a system that has been implemented or benchmarked here.

As shown in Fig. 2, facial video or image input is first subjected to face detection and alignment, followed by preprocessing such as normalization and illumination correction, before CNN-based spatial feature extraction, from which facial-expression and micro-expression-relevant features are derived. Where a facial sequence rather than a single frame is available, an optional temporal facial analysis stage (CNN-LSTM) can be applied to capture the onset–apex–offset dynamics of an expression, consistent with the CK+-based pipelines discussed in Section IV-B. An explainability layer, discussed further in Section IX, is positioned between feature extraction and the final output, most naturally realized through Grad-CAM-based saliency visualization over facial regions, so that the resulting deception-probability score is accompanied by attributable visual evidence. Critically, the pipeline terminates not in an automated verdict but in a human-in-the-loop decision-support interface, reflecting the position argued in Section VII-H that the output of such a system should be interpreted as a probabilistic decision-support signal rather than as definitive proof of deception.

Fig. 2. Conceptual Architecture of a CNN-Based Visual/Facial Deception Analysis Pipeline

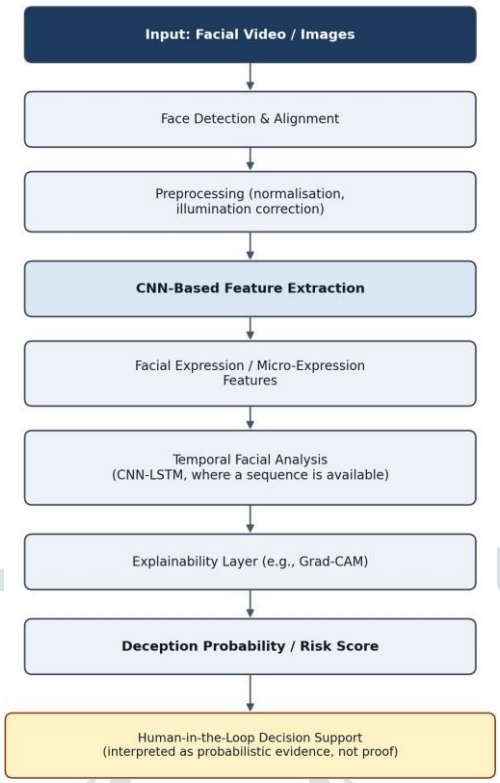

Fig. 2. Conceptual architecture of a CNN-based visual/facial deception analysis pipeline, synthesized from the reviewed literature.

IX. Ethical, Explainability, and Deployment Considerations

A. *Explainable AI*

Given the explainability gap identified in Section VII, candidate techniques include Grad-CAM-based saliency visualization over facial regions to indicate which facial areas drove a visual prediction [18]; SHAP-based feature attribution for CNN embedding-level features to quantify the contribution of specific facial descriptors to a decision score
[19]; and facial-region contribution analysis that reports which part of the face (e.g., eyes, mouth, brow) most influenced
a given classification. Explainability is especially important in this domain because an unexplained deception score carries direct consequences for an individual's legal, professional, or personal standing, and because, per Section VII-H, the underlying cues are inherently probabilistic rather than definitive.

B. *Real-World Deployment Considerations*

Deploying a CNN-based facial-analysis pipeline outside controlled data-collection conditions raises practical constraints absent from most reported evaluations: real-time video processing under limited on-device or edge compute, GPU acceleration or model compression (quantization, distillation, pruning) to meet latency budgets, and robustness to facial occlusion (e.g., masks, spectacles, hands), pose variation, and low-light imaging. A practically important design question is how the system should behave when the face is partially or fully occluded or otherwise unreliable, which argues for architectures with graceful degradation (e.g., explicit occlusion-awareness or confidence-weighted output) rather than architectures that assume a clean, fully visible frontal face is always available.

C. *Ethical and Legal Considerations*

Automated facial deception analysis processes highly sensitive biometric and behavioural data and therefore raises questions of informed consent, the scope and duration of data retention, and the potential for covert surveillance use. The consequences of system error are asymmetric and serious in both directions: a false positive may lead to an unwarranted accusation, reputational harm, or denial of opportunity, while a false negative may permit genuine deception to go undetected in a security-sensitive context. Because reported CNN model behaviour can vary across demographic and cultural groups (Section VII-E), deployment without fairness auditing risks disparate impact. Consistent with the position argued throughout this review, the outputs of such systems should not be treated as legally dispositive evidence of deception, should remain subject to meaningful human oversight, and any jurisdiction-

specific claim about admissibility or legal standing is outside the scope of this review and should be verified against authoritative legal sources rather than inferred from the technical literature.

X. Future Research Directions

1. Construction of larger-scale, naturalistically collected visual/facial deception datasets spanning diverse languages, cultures, and recording conditions.
2. Systematic cross-dataset and cross-corpus evaluation protocols for CNN-based facial models to assess true generalisation rather than in-distribution accuracy alone.
3. Deeper integration of temporal CNN-LSTM (and related sequence) modelling to better capture micro-expression onset–apex–offset dynamics.
4. Self-supervised and transfer-learning approaches to reduce dependence on scarce labelled facial deception data.
5. Robust handling of facial occlusion and degraded visual conditions so that systems degrade gracefully rather than failing silently.
6. Explainable CNN-based analysis that attributes a deception-probability score to specific, inspectable facial evidence (e.g., Grad-CAM regions, action units).
7. Fairness-aware CNN model development and auditing across demographic and cultural subgroups.
8. Privacy-preserving learning (e.g., federated or on-device processing) for sensitive facial biometric data.
9. Edge-optimized and real-time CNN inference pipelines suitable for practical latency budgets.
10. Human-AI collaborative decision frameworks in which CNN-based facial analysis augments, rather than replaces, trained human judgement.
11. Standardized, transparently reported evaluation protocols (datasets, splits, metrics) to enable meaningful comparison across CNN-based facial deception studies.

As shown in Fig. 3, these directions are not independent: dataset limitations propagate into generalisation challenges, static frame-level analysis into temporal information loss, black-box CNN models into explainability deficits, sensitive facial biometric data into privacy and ethical risk, and high computational demand into deployment barriers. Progress on any one axis in isolation is unlikely to be sufficient; the literature synthesized in this review suggests that dataset scale, temporal modelling, explainability, and ethical safeguards need to advance jointly for CNN-based visual deception detection to mature into a scientifically defensible and responsibly deployable technology.

Fig. 3. Research Challenges in CNN-Based Visual/Facial Deception Detection and Their Downstream Consequences

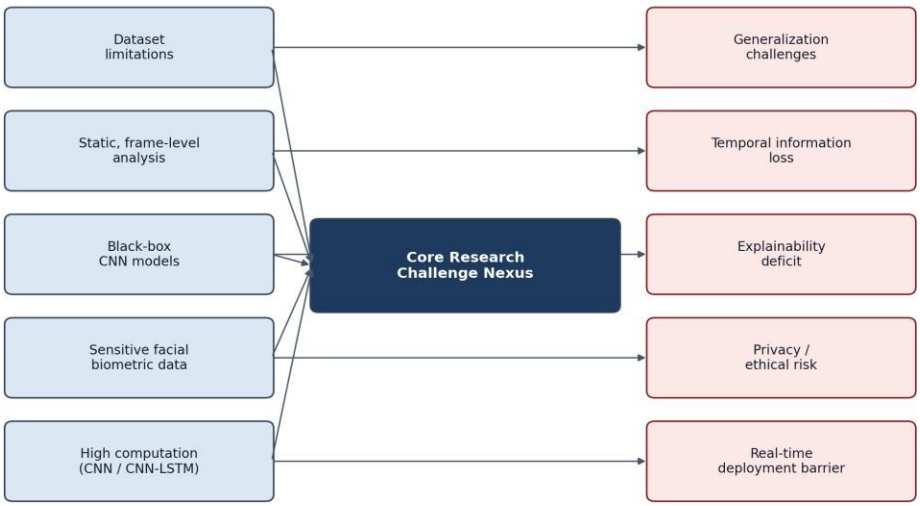

Fig. 3. Research challenges in CNN-based visual/facial deception detection and their downstream consequences.

XI. Conclusion

This review has synthesized the state of CNN-based visual and facial deception detection, examining static and temporal (CNN-LSTM) architectures used to model facial expressions, micro-expressions, and related behavioural cues such as gaze and blink dynamics. Reported findings across the literature are consistent in indicating that temporal extensions of CNN architectures capture expression dynamics that static frame-level CNN analysis cannot, and that FER2013 and CK+, while valuable for representation learning, are emotion-recognition rather than deception-specific resources; genuinely deception-labelled visual corpora, such as the Real-Life Trial dataset, remain comparatively scarce. At the same time, this review has identified substantial and unresolved gaps in dataset scale and diversity, cross-corpus generalisation, the loss of temporal information in static analysis, explainability, demographic and cultural bias, real-time deploy ability, and ethical governance. Most importantly, the behavioural-science foundation of this field establishes that

facial arousal, stress, and cognitive load are not equivalent to deception, and that no facial or behavioural cue currently known is a reliable, unique signature of lying. Current evidence therefore does not justify treating automated CNN-based visual systems as definitive lie detectors. Rather, this review concludes that the appropriate and scientifically defensible framing for this technology is as a probabilistic, explainable, human-in-the-loop decision-support approach for deception-related facial and behavioural cues — one whose future value will depend on larger and more representative visual datasets, more principled temporal modelling, genuine cross-domain generalisation, transparent and auditable explainability, and deployment under clear ethical and legal safeguards.

REFERENCES

[1] A. S. Constâncio, D. F. Tsunoda, H. F. N. Silva, J. M. Silveira, and D. R. Carvalho, "Deception detection with machine learning: A systematic review and statistical analysis," PLoS ONE, vol. 18, no. 2, p. e0281323, 2023.
[2] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in Proc. ACM Int. Conf. Multimodal Interaction (ICMI), 2015, pp. 59–66.
[3] Z. Wu, B. Singh, L. Davis, and V. Subrahmanian, "Deception detection in videos," in Proc. AAAI Conf. Artificial Intelligence, 2018, pp. 1695–1702.
[4] V. Gupta, M. Agarwal, M. Arora, T. Chakraborty, R. Singh, and M. Vatsa, "Bag-of-Lies: A multimodal dataset for deception detection," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2019, pp. 83–90.
[5] F. Soldner, V. Perez-Rosas, and R. Mihalcea, "Box of Lies: Multimodal deception detection in dialogues," in Proc. Conf. North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019, pp. 1768–1777.
[6] L. Mathur and M. J. Matarić, "Introducing representations of facial affect in automated multimodal deception detection," in Proc. ACM Int. Conf. Multimodal Interaction (ICMI), 2020, pp. 305–314.
[7] P. Ekman, Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage, rev. ed. New York, NY, USA: W. W. Norton & Company, 2009.
[8] B. M. DePaulo et al., "Cues to deception," Psychological Bulletin, vol. 129, no. 1, pp. 74–118, 2003.
[9] A. Vrij, Detecting Lies and Deceit: Pitfalls and Opportunities, 2nd ed. Chichester, U.K.: Wiley, 2008.
[10] National Research Council, The Polygraph and Lie Detection. Washington, DC, USA: National Academies Press, 2003.
[11] M. Zuckerman, B. M. DePaulo, and R. Rosenthal, "Verbal and nonverbal communication of deception," Advances in Experimental Social Psychology, vol. 14, pp. 1–59, 1981.
[12] P. Ekman and M. O'Sullivan, "Who can catch a liar?" American Psychologist, vol. 46, no. 9, pp. 913–920, 1991.
[13] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.
[14] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in Neural Information Processing (ICONIP), 2013, pp. 117–124.
[15] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 94–101.
[16] S. Hochreiter and J. Schmid Huber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[17] A. Dudovskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626.
[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
[20] K. Kollreider, H. Fronthaler, and J. Bigun, "Non-intrusive liveness detection by face images," Image and Vision Computing, vol. 27, no. 3, pp. 233–244, 2009.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 1097–1105.
[22] J. Van Der Zee, S. Poppe, R. Taylor, and R. Anderson, "To freeze or not to freeze: A culture-sensitive motion capture approach to deception detection in interrogation contexts," in Proc. Hawaii Int. Conf. System Sciences (HICSS), 2015, pp. 762–771.