Komal Khatak, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Hybrid Deep Learning for Deepfake Detection A Systematic Survey

Komal Khatak[1], Sonal Beniwal[2]
[12]Department of Computer Science and Engineering,
Bhagat Phool Singh Mahila Vishwavidyalaya, Khanpur Kalan, Sonipat, Haryana, India

Abstract- Advancements in generative AI, especially in generative adversarial networks (GANs) and diffusion models, have enabled greater accessibility for users to produce hyper-realistic synthetic media. This has made deep fake detection tools that work with a single modality more vulnerable to adverse real-world environments. To address this, we have implemented a structured survey of hybrid deep learning frameworks that integrate different types of networks, data modalities and representations of the domain. We present a 6-category taxonomy which covers the following areas: (i) CNN architectural hybrids, (ii) CNN-CNN temporal models, (iii) cross-audio-visual modalities, (iv) spatial-frequency hybrid systems, (v) hybrid systems with a forensics perspective, and (vi) adversarial hybrid systems that integrate explainability and adversarial robustness (separation of the model) systems. For the aforementioned areas, we have analyzed the rationale of the design, the fused strategies, and the performance of the systems in relation to the current benchmarks as well as challenges that still persist. Through the analysis of the 45 studies we have examined, we have determined that hybrid models consistently outperform single stream models, especially under compression, domain shifting, and adversarial attacks. Lastly, we have identified challenges that need to be addressed including the generalization gap, the absence of a benchmarking framework, and poor interpretability and we outline systematic and important methods to direct future research with these challenges.

Keywords: deepfake detection; hybrid deep learning; adversarial robustness; explainable AI (XAI); CNN-RNN; spatial-frequency analysis.

I. INTRODUCTION

The advancement of generative AI tools has enabled even non-expert users to produce synthetic media that is nearly indistinguishable from real recordings of people's faces. Modern face manipulation techniques such as identity swapping, expression reenactment, and full-synthesis have matured to the point where the imperfections that mark manipulations and virtual artifacts are undetectable even to untrained observers. This places an increasing burden on automated detection systems (Rana et al., 2022). The secondary risks posed by these systems are significant. The presence of synthetic media has been recorded in electoral disinformation campaigns, non-consensual distribution of intimate media, financial crime, and the systematic breakdown of trust in evidence in legal and journalistic domains.

The earliest forensic analysis techniques, such as sensor pattern noise analysis, error-level analysis

(ELA), and a focus on the detection of physiological cues, offered a reasonable chance of detection with first-generation manipulations. These techniques have, however, seen a dramatic drop in performance with the advent of new generative techniques that have learned to suppress artifacts (Stroebel et al., 2023). The current leading edge has been the use of deep learning, and in particular, Convolutional Neural Networks (CNNs) combined with attention mechanisms, temporal modeling, and cross-domain feature fusion. However, a fundamental problem remains, namely that a detector built for a specific manipulation and a particular dataset performs poorly when assessing unseen systems, be they real-world generators or post processing techniques (Uma Maheshwari and Paulchamy, 2024).

Integrating diverse modalities into a single framework has shown to be a promising approach

against this kind of brittleness. Designed to incorporate spatial, temporal, frequency, and multiple streams of data, hybrid models, have shown to capture a wider range of manipulation artifacts and have a much better cross-dataset generalization compared to the models that operate on a single stream (Heidari et al., 2024). The purpose of this paper, is to provide a systematic and taxonomy-based survey of different hybrid approaches, covering the analysis of their model architecture, evaluation against benchmarks, how robust they are, and how interpretable these models are. The rest of the paper is structured as follows: Research questions driving the review are covered in Section II. Section III contains the methods used followed by provided background in Section IV. Section V encompasses hybrid methodologies and benchmark datasets with evaluation metrics are covered in Section VI. Section VII provides the current state of the art challenges and Section VIII discusses approaches that should be followed to address those challenges. Finally, Section IX provides the concluding statement of the paper.

II. RESEARCH QUESTIONS GUIDING THE REVIEW

This systematic review is structured around five research questions (Table I), developed following a preliminary analysis of recent surveys on deepfake detection (Saeed et al., 2024) and an identification of underexplored areas in hybrid-based deep learning approaches. These questions provide a unifying framework for the analysis of architectures, evaluation practices, and forward-looking research directions.

Table I. Research Questions Guiding The Systematic Review

No.	Research Question	Purpose
RQ1	What hybrid deep learning architectures have been proposed for deepfake detection, and how do they integrate spatial, temporal, frequency, and multimodal information?	Develop a structured taxonomy of hybrid models (CNN-RNN, spatial-frequency, audio-visual, forensics-informed) and analyze their architectural and fusion design choices.
RQ2	How do hybrid approaches compare with single-modality detectors across major deepfake benchmarks?	Assess whether hybrid models yield measurable improvements in accuracy, AUC, and cross-dataset generalization over standard CNN baselines.
RQ3	How do Explainable AI (XAI) methods and adversarial robustness training enhance the forensic trustworthiness of hybrid detectors?	Examine how XAI tools (SHAP, LIME, saliency maps) and adversarial training strategies (FGSM, PGD) are embedded in hybrid pipelines and their effect on interpretability.

RQ4	How do dataset characteristics and evaluation metrics influence the reported performance of hybrid detection systems?	Investigate how quality factors (PSNR, BRISQUE, MS-SSIM) and dataset bias shape benchmark outcomes and whether current protocols reflect genuine model robustness.
RQ5	What are the primary technical and methodological challenges in hybrid deepfake detection, and which research directions hold the most promise?	Synthesize open problems such as the generalization gap, resistance to diffusion-model outputs, and computational constraints, and identify high-priority future directions.

III. METHODOLOGY

This review synthesizes hybrid deep learning approaches to deepfake detection. A structured protocol is used to promote transparency, reproducibility, and the absence of selection bias.

A. Search Strategy

Between January 2018 and March 2025, searches across three major databases - IEEE Xplore, the ACM Digital Library, and Scopus - followed a structured method. Because interest in hybrid detection systems grew during this period, the timeframe aligns with that shift. Search strategies combined two groups of keywords: one targeting techniques like deepfake analysis, face alteration identification, or digital media investigation; another focusing on integrated models such as CNN-RNN setups, audio-visual integration, or spatial-frequency methods. Only articles published in English, found in scholarly journals or conferences, made it into the review pool.

B. Inclusion and Exclusion Criteria

For consideration in this analysis: (a) Studies needed to introduce or assess hybrid deep learning models aimed at spotting deepfakes or altered media. Here, 'hybrid' means combining clearly defined elements - two structural parts or data types - at minimum. Work had to appear in published form and report measurable results using at least one widely accessible benchmark dataset; evaluation criteria required inclusion of common measures like accuracy, AUC, precision, recall, or F1-score. (b) Research got left out when centered on creating fake content instead of detecting it. Papers relying solely on one-type detection systems without added supportive mechanisms did not qualify. Exclusion applied equally to studies addressing only sound or written material if images played no role. Also filtered out were contributions lacking empirical grounding - such as commentaries, dissertations, or informal online posts.

C. Study Selection Process

Starting things off came a preliminary stage. At this point, each found study got examined just by looking at its title and abstract, setting aside ones clearly unrelated - like research into making media or applications beyond forensic settings. If an abstract seemed unclear, judgments were held back until further examination. Those questionable entries stayed in the pool, preventing early elimination. Full texts of what was left then became the focus during what followed. One by one, every item underwent review guided by fixed criteria for inclusion or exclusion. At the outset, duplicates and older versions of manuscripts were removed - only updated, complete submissions remained. The initial retrieval yielded 387 results; once 78 overlapping records disappeared, 309 advanced to preliminary screening. Reading summaries drastically reduced the pool: 231 failed relevance checks, leaving 78 eligible for closer analysis. Examining full texts eliminated a further 33 - for diverse causes: 18 lacked hybrid methodology, 9 omitted statistical data, 4 dealt exclusively with non-visual perception, 2 appeared in foreign languages - and finally, 45 studies survived.

D. Data Extraction and Synthesis

Right away, a single approach set the path for pulling data from each study - all 45 followed an identical format (Table II). Since model designs spanned many forms, origins of information changed across cases, also evaluation techniques lacked uniformity, merging figures into one overall calculation made little sense. Rather than force aggregation, insights took shape through narrative grouping, avoiding statistical blending. As this unfolded, repeated features in setup started becoming visible; at the same time, changes in outcomes across test types grew clearer. With continued review, similarities in framework emerged more fully. When tested against unseen data, performance gaps emerged across systems. Toward the close, concepts connecting stronger resistance to attacks with more transparent decision patterns began forming.

Table II Data Extraction Categories And Elements

No.	Extraction Category	Elements Collected
1	Study Identification	Authors, publication year, venue, DOI
2	Architecture Type	CNN-based, CNN-RNN temporal, audio-visual, spatial-frequency, forensics-informed, XAI-integrated
3	Components & Fusion	Specific networks combined (e.g., VGG16 + LSTM); fusion strategy (early, late, or feature-level)
4	Data Modalities	Visual (RGB, DCT, frequency domain), audio, compressed domain, multimodal
5	Datasets Used	Training and test datasets; evaluation protocol (intra-dataset, cross-dataset)
6	Performance Metrics	Accuracy, AUC, precision, recall, F1-score, EER
7	Benchmark Results	Results on FF++, Celeb-DF, DFDC, DeeperForensics, DeepfakeBench
8	Robustness Analysis	Performance under compression, noise, blur, resolution changes, adversarial attacks
9	Explainability	XAI techniques employed (saliency maps, SHAP, LIME, attention visualization)
10	Computational Efficiency	FLOPs, parameter count, inference time (when reported)

11	Key Findings & Limitations	Main contributions, novelty claims, acknowledged limitations, suggested future work

IV. BACKGROUND

A. Generative Technologies Underlying Deepfake Media

Today's fake videos mostly stem from just two kinds of machine learning setups. Though introduced back in 2014, GANs still play a major role - they run with dual components opposing one another, where one generates visuals while its counterpart detects flaws, gradually sharpening the realism of synthetic faces (Heidari et al., 2024). In contrast to adversarial training, variational autoencoders operate quietly by distilling facial structures into compact representations before reapplying them onto new individuals, a technique widely seen in datasets such as DeeperForensics-1.0. Recently though, diffusion-based systems have drawn increasing notice; starting from random noise, these models form images slowly through iterative refinement, producing outputs that feel both natural and diverse. Because their creation process skips traditional artifacts common in earlier tech, distinguishing authenticity grows more complex (Uma Maheshwari and Paulchamy, 2024).

B. Limitations of Conventional Detection Methods

Once, techniques such as analyzing sensor noise, examining ELA artifacts, or noticing unnatural facial movements - like missing blinks - relied on clear signs from primitive photo edits. As advanced AI systems evolved, these signals weakened quickly, reducing older approaches to near uselessness beyond lab settings (Stroebel et al., 2023). Instead, models based on convolutional networks emerged, learning faint indicators from data instead of preset logic; still, their success tends to waver when resolution changes, sources differ, or different generator versions are tested (Saeed et al., 2024).

C. The Case for Hybrid and Explainable Architectures

One step beyond single-axis models reveals limits in spotting subtle tampering - merging location data with timing cues sharpens detection (Uma Maheshwari and Paulchamy, 2024). Where movement traits mix with irregular rhythms inside unified systems, blind spots fade. Trust builds not just through accuracy, but by showing how conclusions form. Legal settings or media checks demand evidence that stands alongside right answers. Suspicion traces back to exact facial zones - or shifts across video clips - marked plainly behind every outcome (Heidari et al.,2024).

D. The Role of Standardized Benchmarking

Despite advances, uneven testing methods slowed deepfake detection progress - differing training setups, mismatched data divisions, and shifting quality compression played a role. Unified frameworks like DeepfakeBench (Yan et al., 2023) responded by aligning processing steps and scoring rules on diverse collections. Still, as noted in Section VI, challenges endure when assessing realistic fakes or models applied beyond their original context.

V. HYBRID DEEP LEARNING METHODOLOGIES

Though one model alone often fails, combining distinct networks helps tackle challenges too complex for isolated systems. Because each architecture highlights different patterns, fusion approaches capture traces missed elsewhere - texture flaws here, timing issues there. Some rely on spatial features, others on spectral traits, yet none cover every clue on its own. Merging modalities widens detection reach beyond what any singular method achieves. Six such combined strategies appear grouped in Table III. Each brings unique structure, with results unpacked ahead.

Table Iii. Taxonomy Of Hybrid Deep Learning Methodologies For Deepfake Detection

No.	Hybrid Type	Core Idea	Representative Example	Main Components	Reported Benefit
1	CNN Architectural Hybrid	Pre-trained backbone augmented with custom CNN and dense layers for image-level detection	DFP model: VGG16 + additional convolutional and FC layers (Raza et al., 2022)	VGG16 extractor, extra conv. layers, dense classifier	94% accuracy, 95% precision; outperforms Xception, NASNet, and plain VGG16
2	Temporal Hybrid (CNN-RNN)	Sequence modeling layered over per-frame spatial features to capture inter-frame artifacts	Video detectors reviewed in (Heidari et al., 2024); CNN + LSTM/GRU pipelines	CNN for spatial features; RNN/LSTM/GRU for temporal dynamics	Superior accuracy and AUC vs. frame-only CNNs on FF++ and DFDC, especially at high compression
3	Audio-Visual Cross-Modal Hybrid	Joint analysis of video and speech streams to identify temporal inconsistencies between modalities	Multi-modal detectors in (Heidari et al., 2024); audio-visual fusion networks	Visual CNN, audio feature extractor, attention/transformer fusion module	Stronger detection on datasets with diverse speakers; uncovers artifacts invisible in visual-only analysis
4	Spatial-Frequency Feature Hybrid	Parallel processing of RGB content and frequency-domain representations for complementary artifact capture	DCT/residual-augmented CNN branches in (Heidari et al., 2024)	RGB input + DCT/residual maps; multi-branch CNN with mid-level fusion	Maintains high accuracy under heavy JPEG compression and resizing where spatial-only models degrade
5	Forensics-Informed Hybrid	Handcrafted forensic features injected into deep feature spaces to reduce reliance on spurious cues	LBP/sensor-noise + CNN hybrids in (Yadav and Kumar, 2025)	Handcrafted descriptors fused with CNN feature maps	Improved cross-dataset stability; model learns higher-level manipulation fingerprints rather than dataset artifacts

6	XAI-Adversarial Robustness Hybrid (XAI-ART)	Adversarial training (FGSM/PGD) combined with explainability tools in a single detection pipeline	XAI-ART framework (Uma Maheshwari and Paulchamy, 2024)	CNN backbone, adversarial training module, SHAP/LIME explanation layer	~97.5% accuracy; minimal performance degradation under adversarial perturbations; human-interpretable region-wise justifications

A. CNN Architectural Hybrids

From VGG16 - originally built to spot real human faces - researchers attach specialized layers tuned to detect subtle markers of synthetic visuals. Rather than depend on bulky models such as Xception or NASNet, this method favors lighter frameworks without sacrificing performance. With custom convolution stages linked through tightly packed networks, detection sharpens for traces left in manipulated footage. Performance peaks at 94% accuracy; simultaneously, precision reaches 95% when tested on varied sets blending authentic and artificial images. What stands out is how small changes beat complicated designs - shown clearly in Raza's 2022 work on DFP. Better results come not from more resources, yet from smarter choices. Because of this shift, lighter models detect just as well using fewer calculations. Instead of heavy layers, compact convolutions handle tasks smoothly. Lower-resolution pooling steps reduce workload at key points. Even dropout settings get adjusted carefully, removing excess quietly. As a result, running models grows feasible even on basic devices. Possibility spreads beyond high-end systems into everyday tools (Raza et al., 2022).

B. CNN-RNN Temporal Hybrids

While deepfake often evade static image detector, temporal inconsistencies in motion expose manipulation artifacts- blinks seem stiff, lip sync drags behind sound, heads shift too abruptly. It's rhythm that exposes flaws; motion builds sequences tough to mimic perfectly. First cues arrive through space-based clues, teased apart per frame by filters tuned to facial structure. Information moves forward without loss, fed into networks designed to retain what came before. Loops return again and again, studying tiny shifts: tilt changes, pauses lasting just instants, glides between positions. First things shape how later events make sense. When examined closely, flaws become clear - looking at entire sequences works better than quick judgments. Evidence from recent studies (Heidari et al., 2024) supports this: the way something unfolds matters more than isolated details alone. Understanding improves further under tough conditions like high video compression, where small clues in single images vanish. In tests including FaceForensics++ and DFDC, methods tracking movement over time consistently reach stronger AUC scores.

C. Audio-Visual Cross-Modal Hybrids

Even when visuals mislead, pairing audio rhythms with motion signals boosts how well systems detect fakes. From unprocessed sound, algorithms extract voice traits through MFCCs or spectral views rather than depending only on pictures. While one part tracks shifting facial dynamics frame by frame using convolutions, another pathway handles auditory input. Instead of analyzing each stream independently, alignment happens early - guided by attention mechanisms or transformer units before conclusions form. As artificial faces grow indistinguishable from humans, subtle lags between speech output and lip motion turn into key indicators. Though individual frames may appear realistic, mismatched timing between speech and lip movements often reveals altered videos. These errors vanish if audio or video runs independently (Heidari et al., 2024).

Spatial-Frequency Feature Hybrids

Though imperceptible to the eye, patterns in the frequency realm often betray a photo's synthetic origin - especially muted highs and repeating spectral spikes. One analysis stream handles conventional pixel values. Another examines transformed versions like DCT grids, leftover noise layers, or error-level visuals. Midway through processing, signals from both paths merge into shared understanding ahead of labeling. Evidence gathered by Heidari and colleagues in 2024, together with findings from Yadav and Kumar a year later, shows these dual-path methods outperform image-only rivals when files undergo heavy JPEG treatment or scale shifts - steps that erase fine details crucial to spatial inspection alone. Occasionally, engineers embed predefined markers such as binary texture codes or device-specific irregularities so earlier detection logic supports learning instead of being discarded.

Forensics-Informed Hybrids

Though built on neural networks, some hybrid designs begin by weaving insights from traditional forensics into their core structure. Instead of relying solely on raw data, they embed manually designed signals tied to how cameras capture images - patterns like sensor-specific noise, traces left by color filter grids, or distortions from JPEG compression. Because these clues come from actual imaging physics, the system leans less on superficial trends within datasets. Rather than wasting resources picking up irrelevant statistical quirks, it focuses on meaningful signs of tampering. As a result, what emerges performs better when tested on unseen data collections. When labeled examples are few, such setups gain advantage - the hardcoded forensic traits act as guiding hints, shaping the search for solutions more effectively.

XAI-Adversarial Robustness Hybrids

A new type of hybrid system stands out due to its practical impact - combining defenses against manipulated inputs with built-in transparency features. Instead of working separately, these components operate inside one unified structure focused on detecting deceptive media. Take the design by Uma Maheshwari and Paulchamy from 2024: it learns from altered data created through FGSM along with PGD techniques, strengthening resistance when faced with intentional tricks meant to fool recognition. At the very same time, explanation methods like SHAP appear alongside LIME visuals, highlighting specific areas that influenced outcomes. Performance remains strong - accuracy reaches about 97.5% under normal conditions, barely dropping even during attack scenarios. What makes interpretation non-negotiable? Professionals such as lawyers or reporters need more than yes-or-no answers without context. For trust and review purposes, knowing exactly which face parts or motion sequences triggered a result becomes essential according to the authors.

Benchmark Datasets And Evaluation Metrics

Progress in detecting deepfakes depends on shared tools for measurement. Without clear standards, comparing mixed approaches becomes messy. Consistency matters just as much as data quality when testing systems. Five main collections of media shape most research today. Each brings its own measuring techniques into play. These form the backbone of current assessment practices.

A. Principal Benchmark Datasets

Even now, many continue using FaceForensics++ (FF++) (Yan et al., 2023). It builds on 1,000 authentic videos pulled from YouTube. Alongside these stand modified versions made via four distinct techniques: DeepFakes - using autoencoders - Face2Face, FaceSwap, and NeuralTextures. Instead of just one format, the data comes in three compression levels: raw, c23, and c40. That variation lets analysts track changes in detection success when visuals degrade. Since each version maintains consistent alterations, studies can zoom into individual forgery styles or stretch out toward cross-method contrasts.

Better visuals mark the beginning of Celeb-DF (v2), going beyond FF++ by refining its face-swapping technique through 590 genuine celebrity interview clips. Over 5,000 manipulated videos emerge from this approach - each displaying subtler inconsistencies. Though detection tools handle FF++ with strong results, performance drops noticeably

once applied to this collection. As a result, adaptation across data sources becomes clearer when measured against Celeb-DF. Researchers now lean on it as a standard gauge precisely because earlier benchmarks fail under similar pressure.

Despite aiming for realism, the DFDC collection from (Lin et al., 2024) mixes about 19,000 real video segments with 10,000 altered versions, recorded in irregular conditions using diverse participants. Since assessment uses a hidden partition paired with skewed label representation, it mirrors patterns seen in organic online sharing. Such an arrangement brings performance testing nearer to real-world usage compared to earlier attempts. Notable here is how well it captures lopsided, chaotic engagement common across internet spaces.

What sets DeeperForensics-1.0 apart begins with its reliance on authentic video manipulations - blurring, noise injection, compression - not artificially generated ones. Though built from ten thousand autoencoder-produced fakes, the key difference lies in how distortions were tuned: human judgment shaped them. Where earlier datasets used synthetic degradation, this one leaned on subjective realism ratings. As perception drove preprocessing choices, models face tougher benchmarks mirroring daily visual conditions. So performance here reflects robustness beyond clean laboratory settings.

DeepfakeBench (Yan et al., 2023) begins by tackling uneven benchmarking practices. Different data sources - FF++ variants, Celeb-DF, DFDC, DFD, and DeeperForensics-1.0 - are unified under a single preprocessing pipeline. With consistent splits between training and evaluation sets, plus standardized labels, comparisons gain stability. Thanks to this setup, fifteen prominent detection approaches now face fairer head-to-head analysis. As months pass, its role grows within research circles evaluating novel techniques.

Table Iv. Overview Of Principal Deepfake Benchmark Datasets

No.	Dataset	Scale (approx.)	Manipulation Types	Notable Characteristics	Common Metrics
1	FaceForensics++ (FF++)	1,000 real videos; several thousand manipulated clips	DeepFakes (AE), Face2Face, FaceSwap, NeuralTextures	Multiple compression levels (raw, c23, c40); supports intra- and cross-manipulation evaluation; widely used training benchmark (Yan et al., 2023)	Frame/video accuracy, AUC, precision, recall, F1, EER
2	Celeb-DF (v2)	590 real + 5,000+ fake celebrity videos	High-quality face swapping	Improved face-swap pipeline yields fewer visible artifacts; standard cross-dataset hard test set (Kumar et al., 2020)	AUC, accuracy, F1-score, EER
3	DFDC (Kaggle Challenge)	19,000+ real + 10,000+ fake videos	Multiple internal face-swap and reenactment pipelines	Large subject diversity, varied capture conditions, hidden test sets, class imbalance reflecting real platform traffic (Yan et al., 2023; Lin et al., 2024)	AUC, accuracy, precision-recall, weighted PR

4	DeeperForensics-1.0	10,000 manipulated videos	Autoencoder-based face swapping (DF-VAE)	Realistic post-processing perturbations (blur, noise, compression); human-validated perceptual realism; designed for robustness benchmarking (Kim et al., 2024)	AUC, accuracy under perturbations, cross-dataset tests
5	DeepfakeBench Corpus	Integrates FF variants, Celeb-DF, DFDC, DFD, DeeperForensics-1.0	Mixed (AE, GAN, graphics-based)	Unified preprocessing, standardized splits and metadata; supports fair comparison of 15 representative detectors (Yan et al., 2023)	Frame-level AUC, accuracy, AP, EER (standardized pipeline)

Dataset Quality and Its Detection Impact

A new study by Kim and colleagues in 2024 examined how visual quality is spread across key deepfake collections, drawing on both reference-based and blind assessment tools like PSNR, MS-SSIM, MSE, BRISQUE, LPIPS, and CLIP-IQA. While comparing these sets, noticeable differences emerged - not just between them, but also inside each one - suggesting overall AUC values might hide gaps in detection linked to image clarity. As fake images grow more realistic, systems such as Xception, SPSL, and UCF tend to become less confident and make more errors. Because of this pattern, relying solely on average results risks overlooking weaknesses when handling well-crafted fakes.

C. Core Detection Metrics

Most times, deepfake detection gets treated like a yes-or-no sorting task, so evaluation follows that structure. Overall correctness is captured by accuracy - the share of right predictions - which stays easy to calculate yet reacts strongly when real and fake examples aren't balanced. Instead of picking one threshold, AUC sums up how well a model tells classes apart at every possible level, smoothing out issues caused by uneven data distribution, which explains its popularity in thorough tests like DeepfakeBench (Yan et al., 2023). What precision shows is how many of the flagged clips truly were manipulated, whereas recall reveals how many actual manipulations got caught. One overlooks false alarms; the other misses gaps in catching fraud. A different way to summarize performance is through the F1-score, using a harmonic average. Where mistakes in accepting and rejecting balance out defines the Equal Error Rate (EER), especially meaningful when analyzing forensic or biometric systems (Altuncu et al., 2024).

D. Robustness and Efficiency Metrics

Basic measures of correctness fall short when judging how systems perform outside labs. When tested, detection tools face deliberate changes like compressed images, soft edges from blur, added static patterns, reduced image size - also attempts to trick them without catching a person's eye. Certain tests go further by using carefully picked fake samples so subtle they puzzle trained reviewers too (Lin et al., 2024). Speed and resource demands emerge from counting math steps, model size in storage terms, plus processing duration per frame under uniform graphics processors. Looking closely at 92 designs across half a dozen aspects - precision, adaptability, stability under distortion, defense against manipulation, space needs, quickness - one truth stands: none lead everywhere at once. Each use case shapes which compromises matter most (Lin et al., 2024).

V. CHALLENGES AND LIMITATIONS

A. Generalization and Dataset Bias

Most deepfake detection models struggle with structural flaws that persist regardless of category.

Instead of identifying consistent signs of tampering, they often fixate on subtle patterns tied to their training data. High scores emerge during testing within familiar datasets. Yet accuracy drops sharply when faced with new forgery methods, better fakes, or altered capture settings. Evidence points toward reliance on cues like compression noise, uneven lighting, or backdrop textures - side effects of how samples were made. These shortcuts fail once real-world variety enters the picture. Widely used test sets including FF++ and Celeb-DF offer narrow scenarios. As a result, performance figures may paint an overly optimistic view of actual field utility.

B. Robustness to Perturbations and Adversarial Attacks

Though real videos often undergo changes like resizing or compression, these steps alter fine details that detection tools need to spot fakes. When platforms reprocess footage, even advanced models struggle - accuracy drops appear across recent studies (Rana et al., 2022; Heidari et al., 2024; Lin et al., 2024). Worse still, attackers may introduce hidden tweaks that people cannot see but which confuse automated systems. Such manipulations specifically weaken detectors built around one design. Because of this, putting trust in just one type of model opens doors to deliberate bypassing, especially where security matters most.

Computational and Deployment Constraints

Though accurate detection methods demand heavy computation, they usually need large GPU memory alongside delays exceeding seconds for each video segment - making them unsuitable for real-time analysis of streaming footage. Simpler models are available, yet tend to struggle with subtle, well-crafted fake content, losing vital precision. Such a balance between speed and effectiveness continues to limit practical use in areas like handheld devices or instant broadcast review (Heidari et al., 2024; Lin et al., 2024)

Evaluation Inconsistency

Lacking shared testing rules made study comparisons hard for years. Though training and test data vary, so do image compression methods, category spreads, and performance metrics - differences in model outcomes often reflect these design choices rather than real advances. Under uniform evaluations such as DeepfakeBench (Yan et al., 2023), previously strong models show weaker gains when tested fairly. Earlier reports of major breakthroughs now face doubt due to these more controlled assessments (Rana et al., 2022; Heidari et al., 2024; Lin et al., 2024; Kim et al., 2024).

E. Interpretability and Coverage Gaps

Now working behind closed doors, deepfake detectors deliver answers while keeping their reasoning out of sight. Slight irregularities in blinking rhythms or speech timing guide these judgments - yet users never view such signals. Over days, trust fades when people cannot inspect the signs shaping outcomes. In courtrooms, transparency matters; rulings must tie directly to visible proof. Because rulings often face challenges, clear explanations based on visible flaws become essential. Still, most current approaches do not offer this kind of open, logical backing. Despite progress in understanding, identifying fake audio or misleading text remains weak. Beyond these shortcomings, larger solutions are seldom discussed - methods such as following hidden digital footprints, protecting personal information while examining content, or matching detection tools to real-world platform rules stay ignored (Rana et al., 2022; Heidari et al., 2024).

VI. FUTURE DIRECTIONS

A. More Realistic and Diverse Benchmark Design

Moving ahead involves adding diverse ages, tongues, and everyday background sounds into testing - blending computer-generated audio from diffusion methods with artificial clips where visuals and voice shift in sync. Success should center on wide-ranging detection skills rather than slotting them as minor metrics. Once prioritized, systems start chasing underlying signs of tampering instead of reacting to how data is arranged. Past efforts back this move (Rana et al., 2022; Lin et al., 2024), revealing benefits when ignoring shallow traits.

Generalizable and Robustness-Oriented Architectures

Most progress happens once rigid templates fade into adaptable systems meant for wide change. Not

limited to tagged samples, pulling insights from endless unlabeled footage through self-guided methods can deepen core comprehension. Learning how to learn - shaped by principles like meta-training - often sharpens machine reactions to unknown physical challenges. Relying only on expected dangers holds growth back; strength builds when practice covers everyday image flaws alongside digital breaches. Performance drops fast when clean labs meet messy reality; aligning invisible patterns across settings using domain shifts eases the mismatch (Heidari et al., 2024; Lin et al., 2024)

C. Efficient Multi-Modal Deployment

Fast operation on mobile devices comes naturally to compact models, yet shrinking them too much can backfire. Efficiency matters - so does precision when spotting deception. Though speed wins points, trust demands correctness just as much. Combining vision, audio, and spoken cues closes holes single-sense systems leave open. Desynchronization between lip movement and audio serves as a primary indicator of synthetic manipulation. Hidden flaws sometimes surface through timing hiccups or shifts in pitch. Perfect at first glance, synthetic visuals carry subtle distortions. Artificial voices reach close resemblance to real ones. Still, mismatches in audio flow and mouth motion give them away. Occasionally, it is these tiny gaps that betray the illusion. Errors often appear when false signals align across sources. Though small, these glitches may boost our ability to spot sophisticated fakes - according to new research (Heidari et al., 2024; Lin et al., 2024)

D. Provenance and Ecosystem-Level Defenses

Starting with traceability matters just as much as later scrutiny when spotting deception. Instead of waiting to catch alterations, methods like lasting visual markers, edit-transparent digital seals, or shared time records let users verify authenticity firsthand. Over time, these pieces could align - not operating separately but reinforcing one another - tying responsibility, source tracking, and evaluation into a unified system (Rana et al., 2022; Heidari et al., 2024).

E. Explainable and Trustworthy AI for Forensics

Deepfake detection tools need clear reasoning behind their decisions if courts, newsrooms, or regulators are to rely on them - opaque models fall short here. Moving forward, one key task involves crafting explanation techniques tied directly to time and space elements within videos, tailored for forensic analysis. Setting benchmarks for what counts as a high-quality explanation in such cases also matters greatly. Another concern emerges when explanations clash with model defenses: how much clarity can persist under attack? Combining explainable AI with training that anticipates adversarial inputs offers one path ahead - the XAI-ART approach illustrates this possibility well (Uma Maheshwari and Paulchamy, 2024).

VII. CONCLUSION

Despite growing pressure from advanced fake media, detection methods now lean on mixed system designs - ones combining visual patterns, timing cues, spectral details, and multiple data types into cohesive frameworks. Evidence gathered here shows these blended setups beat simpler, one-track models on key tests, especially when files are heavily compressed, tested beyond their original datasets, or subtly altered by attackers. Drawing together distinct traces of manipulation gives such systems an edge, making them today's leading approach for spotting synthetic content.

Even so, major obstacles stand in the way of dependable deployment in serious real-world applications. What often happens is that detection models built on lab-style data falter when faced with new kinds of tampering or messy everyday environments - the core issue holding progress back. On top of that, inconsistent testing methods have, over time, made results look better than they are. Since most current designs work like closed systems without clear reasoning trails, courts and investigators cannot rely on them where transparency matters.

Progress ahead depends on linked breakthroughs in several areas. Building benchmarks that better reflect

real-world conditions comes first. Next, designing systems able to learn stable patterns despite changes in input. Merging trustworthy, interpretable artificial intelligence into one coherent structure follows close behind. Adding tracking methods that work at the level of entire information ecosystems matters just as much. Success hinges not on small upgrades in isolated parts, but on combining these elements effectively. Only through such integration can digital media remain reliable as synthetic tools grow stronger.

REFERENCES

1. Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS).
2. Ahmed, A. A., et al. (2024). Review on hybrid deep learning models for enhancing encryption techniques against side channel attacks. IEEE Access, 12, 188435–188453. https://doi.org/10.1109/ACCESS.2024.3431218
3. Altuncu, E., Franqueira, V. N. L., and Li, S. (2024). Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. Frontiers in Big Data, 7, 1400024. https://doi.org/10.3389/fdata.2024.1400024
4. Chaudhary, U. (2025). Deepfake detection using convolutional and recurrent neural network. International Journal for Research in Applied Science and Engineering Technology, 13(4), 2572–2575.
5. Cozzolino, D., et al. (2019). ForensicTransfer: Weakly-supervised domain adaptation for forgery detection. arXiv:1812.02510.
6. Guera, D., and Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS).
7. Heidari, A., Jafari Navimipour, N., Dag, H., and Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. WIREs Data Mining and Knowledge Discovery, 14(2), e1520. https://doi.org/10.1002/widm.1520
8. Kim, H., Lee, J., Park, L. H., and Kwon, T. (2024). On the correlation between deepfake detection performance and image quality metrics. In Proceedings of the 3rd ACM Workshop on Security Implications of Deepfakes and Cheapfakes (pp. 14–19). https://doi.org/10.1145/3660354.3660358
9. Kumar, A., Bhavsar, A., and Verma, R. (2020). Detecting deepfakes with metric learning. In Proceedings of the IEEE 8th International Workshop on Information Forensics and Security (WIFS) (pp. 1–6). https://doi.org/10.1109/WIFS49906.2020.9360901
10. Li, Y., and Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. In Proceedings of the IEEE/CVF CVPR Workshops.
11. Li, Y., Chang, M.-C., and Lyu, S. (2018). In Ictu Oculi: Exposing AI-created fake videos by detecting eye blinking. In Proceedings of the IEEE WIFS.
12. Lin, C., Shen, C., Deng, J., Wang, Q., Hu, P., and Li, Q. (2024). Towards benchmarking and evaluating deepfake detection. IEEE Transactions on Dependable and Secure Computing, 21(6), 5112–5127. https://doi.org/10.1109/TDSC.2024.3369711
13. Mirsky, Y., and Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys, 54(1), 1–41.
14. Monteiro, S., Wanzeller, C., and Caldeira, F. (2024). Performance analysis on deep fake detection. Communications of the IBIMA, 2024, 457767. https://doi.org/10.5171/2024.457767
15. Nguyen, H. H., Yamagishi, J., and Echizen, I. (2019). Capsule-Forensics: Using capsule networks to detect forged images and videos. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).
16. Prabakar, D., et al. (2022). Hybrid deep learning model for copy move image forgery detection. In Proceedings of the 6th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (pp. 1023–1028). https://doi.org/10.1109/I-SMAC55078.2022.9987319

17. Rana, M. S., Nobi, M. N., Murali, B., and Sung, A. H. (2022). Deepfake detection: A systematic literature review. IEEE Access, 10, 25494–25514. https://doi.org/10.1109/ACCESS.2022.3154404
18. Raza, A., Munir, K., and Almutairi, M. (2022). A novel deep learning approach for deepfake image detection. Applied Sciences, 12(19), 9820. https://doi.org/10.3390/app12199820
19. Rössler, A., et al. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).
20. Saeed, N., Mumtaz, G., Yaqub, M., and Ahmad, M. H. (2024). Improving deepfake detection: A comprehensive review of adversarial robustness, real-time processing and evaluation metrics. Journal of Computer Biomedical Informatics, 7(2). https://doi.org/10.5697/97022024
21. Safwat, S., Marzouk, A., Eldesoky, I., and Ali, F. (2024). Hybrid deep learning model based on GAN and ResNet for detecting fake faces. IEEE Access, PP, 1–1. https://doi.org/10.1109/ACCESS.2024.3416910
22. Saxena, A., et al. (2023). Detecting deepfakes: A novel framework employing XceptionNet-based convolutional neural networks. Traitement du Signal, 40(3), 835–846. https://doi.org/10.18280/ts.400301
23. Stroebel, L., Llewellyn, M., Hartley, T., Ip, T. S., and Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. Journal of Cyber Security Technology, 7(2), 83–113. https://doi.org/10.1080/23742917.2023.2192888
24. Sun, Z., Han, Y., Hua, Z., Ruan, N., and Jia, W. (2021). Improving the efficiency and robustness of deepfakes detection through precise geometric features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3609–3618).
25. Tolosana, R., et al. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131–148.
26. Uma Maheshwari, R., and Paulchamy, B. (2024). Securing online integrity: A hybrid approach to deepfake detection and removal using explainable AI and adversarial robustness training. Automatika, 65(4), 1517–1532. https://doi.org/10.1080/00051144.2024.2400640
27. Yadav, S., and Kumar, A. (2025). Enhancing fake image detection with a hybrid approach of deep learning and image forensics. International Journal of Science and Technology (IJSAT), 16(2), 1–5.
28. Yan, Z., Zhang, Y., Yuan, X., Lyu, S., and Wu, B. (2023). DeepfakeBench: A comprehensive benchmark of deepfake detection. In Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks. arXiv:2307.01426.
29. Yang, X., Li, Y., and Lyu, S. (2019). Exposing deepfakes using inconsistent head poses. In Proceedings of the IEEE ICASSP.