Vol.4, No.7, Page 2379- 2389, 2026 https://doi.org/10.59628/jast.v4i7.3069 ISSN 2958-9568 JAST

SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا

Sana'a University Journal of Applied Sciences and Technology

https://journals.su.edu.ye/index.php/jast/

# Lightweight Hybrid Deep Learning Models for Real-Time Deepfake Video Detection: A Comprehensive Survey

**Azhar Abdulmughni * and Nagi AL-Shaibany**

**Department of Information Technology, Faculty of Computer Sciences and IT, Sana'a University, Sana'a, Yemen**

*Corresponding author: azhar.abdalmgni@su.edu.ye

## Abstract

The rapid advancement of Generative Artificial Intelligence (GAI) has led to the proliferation of deepfake media, posing significant threats to digital security, privacy, and information integrity. To overcome this challenge, substantial research efforts have been directed toward developing automated detection techniques using deep learning methodologies. This study presents a comprehensive survey of deep learning-based deepfake detection methods, emphasizing lightweight and hybrid architectures designed for real-time deployment. The survey systematically categorizes the landscape of deepfake generation techniques and evaluates state-of-the-art detection frameworks, including: classical CNNs, efficient backbone architectures (MobileNet, EfficientNet), and spatiotemporal models (CNN-LSTM/GRU). Furthermore, this study examines model compression techniques — including pruning and quantization — essential for resource-constrained deployment, and provides a structured analysis of benchmark datasets, major detection architecture categories, and persistent research gaps. By critically examining the trade-off between detection accuracy and computational latency, this paper identifies key open challenges and concludes by highlighting a research gap for probabilistically robust, lightweight frameworks, offering a roadmap for future research toward reliable, real-time deepfake forensics in unconstrained environments.

## Article Info

**Keywords:**
Deepfake detection, Deep Learning, Lightweight models, Transfer Learning, Hybrid architectures, Real-time detection.

**Article History:**
**Received:** 22-April-2026,
**Revised:** 02-May-2026,
**Accepted:** 12-June-2026,
**Published:** 28 July 2026.

## 1. INTRODUCTION

Recently, the ability to produce and alter synthetic media has become increasingly accessible. The term "deepfake" — a portmanteau of "deep learning" and "fake" — emerged as a direct consequence of rapid advancements in artificial intelligence and deep learning. These technologies can produce or modify audio, video, and image content with a degree of realism that challenges human perception. [1]

Deepfakes can replace, reenact, and alter human faces and voices. The most common manipulation modalities include face swapping, audio-visual lip-syncing, and attribute altering (e.g., age, expression). These manipulations can be applied to static images or temporally coherent video sequences; the latter presents additional forensic difficulties because temporal consistency is required. Deepfakes generally fall into three types: photo deepfake, voice deepfake, and video deepfake [2]. Although there are legitimate creative and entertainment applications for deepfake technology, its use for identity fraud, disinformation, and non-consensual material poses significant societal risks that call for strong responses.

Deepfake detection remains challenging, despite significant research activity. Detection models frequently fail to generalize to novel manipulation techniques or adapt to previously unseen datasets. Moreover, detection skills are consistently surpassed by the rapid development of generative techniques, from generative adversarial network (GANs)-based face swaps to diffusion model-synthesized landscapes. However, there is a fundamental conflict between detection accuracy and computing

**Table 1.** Overview of related surveys

| (Year/Author) | Main Focus of the Survey |
|---|---|
| **(2024) [5]** | DL content detection progress: spatial vs. temporal analysis |
| **(2025) [6]** | Adversarial robust detection systems |
| **(2024) [7]** | Media-modality fusion and advanced ML |
| **(2024) [8]** | Architectural variants, datasets, and meta-literature review |
| **(2023) [9]** | Impact analysis and detection in all formats |
| **(2022) [10]** | SLR of methodologies and performance analysis |
| **(2025) [11]** | Compression and transfer learning techniques for deepfake detection |
| **(2023) [1]** | Systematic DL strategies for hybrid detection |

viability because real-world deployment scenarios, such as real-time social media tracking, require lightweight models that work under tight latency and resource constraints. [3, 4].

**Table 1** presents an overview of recent surveys related to deepfake detection.

Although several surveys have covered deepfake detection from many perspectives, a substantial gap remains in the literature. Current surveys either target efficiency without a systematic analysis of hybrid spatiotemporal architectures or concentrate on detection in general without differentiating between computing resource requirements. No previous review offers an organized evaluation of model compression techniques (pruning and quantization) in the context of deepfake detection, nor does it specifically address the trade-off between lightweight model design and real-time detection accuracy.

## 2. METHODOLOGY

This survey systematically analyzed 75 peer-reviewed studies published between 2020 and 2026. Records were identified through searches of six major academic databases: IEEE Xplore, ScienceDirect, SpringerLink, Wiley Online Library, ACM Digital Library and Google Scholar. The following Boolean search string was applied across all databases: (“deepfake detection” OR “face forgery” OR “face manipulation”) AND (“deep learning” OR “neural network”) AND (“lightweight” OR “real-time” OR “edge deployment”). The inclusion criteria were as follows: (1) peer-reviewed publication, (2) direct relevance to video deepfake detection using deep learning architectures, and (3) reporting of quantitative performance metrics on recognized benchmark datasets. Studies lacking experimental validation or not directly related to deep learning-based detection of deepfakes were excluded. To ensure methodological rigor, each included study was evaluated against three quality assessment criteria: (1) methodological rigor — the study clearly describes its model architecture, training procedure, and evaluation protocol; (2) performance reporting — the study reports quantitative metrics such as accuracy, AUC, or F1-score on recognized benchmarks; and (3) reproducibility — the study provides sufficient implementation details to allow replication of results. Studies that failed to satisfy at least two of these three criteria were excluded from the final review.

The remainder of this paper is organized as follows: Section 2 presents the research methodology. Section 3 introduces a taxonomy of deepfake types. Section 4 reviews the deepfake generation techniques. Section 5 discusses the key detection challenges. Section 6 presents the benchmark dataset. Section 7 surveys deep learning-based detection approaches, including lightweight architectures, hybrid models, model compression strategies and transfer learning techniques. Section 8 provides a comparative analysis of recent state-of-the-art (SOTA) methods. Section 9 discusses the research gaps and future directions. Finally, Section 10 concludes the paper.

Figure 1: shows the PRISMA flow diagram of the study selection process.

## 3. DEEPFAKE TYPES:

Deepfakes can be categorized into three primary manipulation types, each presenting separate challenges**. First , Face Swap** entails the full substitution of a source person’s face with that of a target, generally accomplished through encoder-decoder frameworks or Generative Adversarial Networks (GANs) (e.g., DeepFaceLab), where the identity is changed while maintaining the target’s authentic expressions [12–14]. **Second is Face Reenactment** (or puppet mastery), which focuses on manipulating the target’s facial expressions, mouth movements, or head poses in agreement based on the source video. This technique is often used when the identity of the source remains the same, but the message is altered [15–17]. **Finally, Attribute Manipulation** fine-tunes the target’s appearance by modifying specific facial features, such as hair color, age, and skin tone, and adding or re-

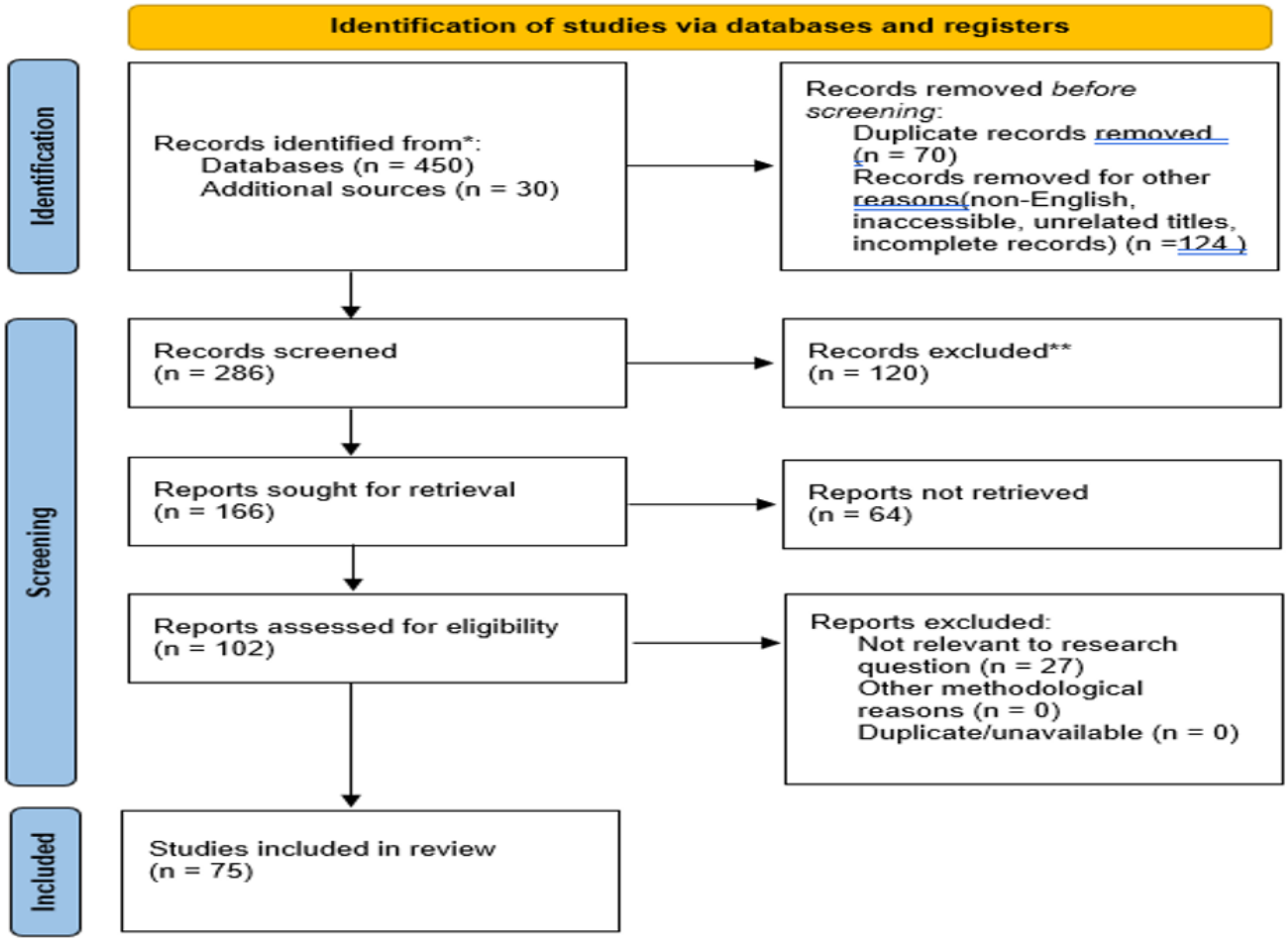


**Figure 1.** PRISMA flow diagram of the study selection process

moving extra details, such as glasses and makeup. This technique primarily utilizes image-to-image translation models, such as StarGAN. [18] While these manipulations are the primary types of deepfakes, the field of deepfake is evolving so quickly that many researchers have identified additional categories of deepfake videos. like Entire Face Synthesis which involves generating fully new, non-existent facial characters from the ground up, usually by GANs, raising severe identity-misuse concerns [19–21] another types is Audio Deepfakes which is produced by AI-synthesized voices nearly indistinguishable from authentic ones, representing security risks mainly for visually impaired people [22, 23] Finally, Textual Deepfakes, It uses NLP models to generate highly deceptive text that reads like it was written by a human, the challenge is that the ability to catch these fake is falling behind,especially for low-resource languages [24].

## 4. DEEPFAKE GENERATION TECHNIQUES

Four principal architectures dominate the deepfake generation. (GANs) use a generator-discriminator framework to produce highly realistic synthetic media. Variants such as StyleGAN, CycleGAN, and Conditional GANs have achieved remarkable success in image and video generation, making them foundational to modern deepfake creation and detection research. [25–28] Variational Autoencoders (VAEs) learn latent representations to generate facial images, offering stable training at the cost of lower resolution and occasional blending artifacts. While VAEs are commonly utilized in picture production and creative applications, they have limitations such as skewed likelihood estimations and high processing demands. [29–31] Diffusion Models (DMs) have recently surpassed GANs in photorealism, generating highly diverse and controllable imagery, although at a substantially higher computational cost and with unique spectral fingerprints. The rapid evolution of diffusion-based generation continues to challenge the existing deepfake detection frameworks. [32–34] **Transformers** are the foundation of models such as GPT, and their strength lies in their capacity to simulate long-range dependencies and context across sequential data, making them indispensable for tasks such as producing coherent video sequences, complex text-to-video situations, and high-quality audio synthesis. [35]

**Table 2** compares the four primary deepfake generation architectures, highlighting the key technical distinctions that define their roles in the current generative AI landscape.

**Table 2.** Comparative Summary of Deepfake Generation Techniques

| Technique | Strengths | Weaknesses | Key Models | Detecting Fingerprints |
|---|---|---|---|---|
| AE / VAEs [29–31] | Stable training, accessible tools | Lower resolution, blending seams | DeepFaceLab and SimSwap | Visible boundary seams |
| GANs [25–28] | Fast inference, high resolution | Training instability, flickering | StyleGAN, CycleGAN, StarGAN | GAN fingerprints, texture artifacts |
| Diffusion Models | Superior realism, controllability | Very slow inference, huge compute | Stable Diffusion, Midjourney | Frequency-domain artifacts |
| Transformers | Long-range coherence, multimodal | Massive computational overhead | GPT-4V, Sora-like models | Temporal inconsistencies |

## 5. KEY CHALLENGES IN DEEPFAKE DETECTION:

Deepfake detection remains a difficult research challenge because of several interlocking technical and practical factors. [36]

### 5.1. Rapid Advancement of Deepfake Generation Techniques

Deepfake generation models, especially those based on Generative Adversarial Networks (GANs), are evolving rapidly. New methods produce highly realistic manipulated content that is difficult for both humans and machines to detect. The rapid development of tools such as StyleGAN, DeepFaceLab, and FaceSwap has significantly outpaced the progress of detection mechanisms.[37]

### 5.2. Generalization Across Datasets and Techniques

Many deepfake detectors perform well on specific datasets but fail when tested on new or unseen manipulation types. This is due to overfitting on a particular dataset or the lack of diverse and representative training samples. Generalization is one of the most frequently cited limitations in the recent literature.[38]

Most benchmark datasets (e.g., FaceForensics++, Celeb-DF, and DFDC) are outdated or limited in terms of ethnicities, backgrounds, and manipulation methods. This lack of data restricts the training of more adaptable and generalizable models. [39]

### 5.3. Data Bias and Fairness

Biases in training data, such as the overrepresentation of certain demographics or environments, can lead to detection systems that are unfair or ineffective when applied to real-world scenarios. For example, detectors may perform poorly on videos of underrepresented ethnic groups. [40]

### 5.4. Computational Efficiency and Real-Time Constraints

In real-time applications, such as social media content monitoring and security screening, detection systems must offer both high accuracy and low latency. However, deep learning models that achieve high detection rates, such as complex CNNs or transformer-based models, are often computationally expensive and unsuitable for lightweight or real-time deployment. [41]

### 5.5. Robustness to Real-World Distortions

In real-world scenarios, media often experiences compression, noise, resizing, or platform-specific "laundering" that obscures the forensic artifacts that models depend on and degrades their performance. Real-world degradation compression, noise, and resizing obscure forensic artifacts.

### 5.6. Lack of Interpretability

Deepfake detection models, especially those based on deep learning, often act as "black boxes," providing limited explanations for their decisions. This lack of interpretability reduces trust in their output and hinders their adoption in sensitive domains, such as legal proceedings. [40]

### 5.7. Adversarial Attacks

Adversarial attacks represent a critical yet underexplored vulnerability in deepfake detection. Gradient-based perturbations, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), can significantly degrade detection accuracy while remaining imperceptible to human observers. Lightweight architectures are particularly susceptible because their reduced parameter capacity limits their robustness against adversarial noise. This accuracy-robustness trade-off remains largely unresolved in the current literature, representing a

**Table 3.** Key Benchmark Datasets for Deepfake Detection

| Dataset | Scale | Types of Deepfakes | Key Strengths | Limitations |
|---|---|---|---|---|
| FaceForensics++ (FF++) | 1,000 videos × 4 methods | Face swap, expression reenactment | 3 compression levels; structured manipulation types | Poor generalization to in-the-wild forgeries |
| DFDC | >100,000 videos | Face swap, audio-visual forgeries | High demographic diversity; real-world conditions | No technique-level labels |
| Celeb-DF v2 | 5,639 videos; 59 celebrities | High-fidelity face swap | Advanced face-swapping; reduced temporal flickering | Predominantly Caucasian subjects (88%) |
| FFIW 10K | 10,000 videos (5K real / 5K fake) | Face swap in unconstrained settings | Perfectly balanced classes; eliminates majority-class bias | Limited manipulation diversity |

significant gap in reliable real-world forensic deployment. [39]

## 6. BENCHMARK DATASETS FOR DEEPFAKE DETECTION

Benchmark datasets are the fundamental pillar for developing and evaluating deepfake detection models, providing the homogenous amounts necessary to train neural networks to distinguish between authentic and manipulated content [42]. These datasets have evolved through multiple generations, meaningfully increasing in size, visual quality, and diversity to keep pace with advancing generative methods, such as Generative Adversarial Networks (GANs) and diffusion models [43, 44]. Early research focused mainly on visual-only datasets, but there is a growing trend towards multimodal datasets that include coordinated audio-visual data forgeries. [44] The most influential benchmark, FaceForensics++, provides 4,000+ videos across four manipulation techniques and three compression levels (Raw, HQ, LQ), simulating real-world social media "laundering" [45]. To address the need for scale and diversity, the Deepfake Detection Challenge (DFDC) introduced over 128,000 videos featuring diverse ethnicities, ages, and environmental perturbations, which is essential for testing the robustness [46] of the model [47]. Celeb-DF v2 significantly improved visual realism through the application of sophisticated face-swapping techniques combined with Kalman smoothing to minimize inter-frame inconsistencies; however, the dataset suffers from a notable demographic imbalance, with Caucasian subjects accounting for 88% of its content. [8][48]. Finally, FFIW 10 K was introduced to mitigate training bias by balancing the distribution of fake to real content (5,000 real and 5,000 manipulated), which ensures more reliable accuracy metrics [47].

**Table 3** provides a detailed comparative analysis of these benchmark datasets, highlighting their scale, manipulation techniques, and key advantages and limitations.

## 7. DEEP LEARNING MODELS FOR DEEPFAKE DETECTION

Before deep learning became the dominant paradigm, detection systems depended on handcrafted features such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), paired with classical classifiers including Support Vector Machines (SVM) and Random Forests. While these methods established early detection baselines, they proved insufficient in capturing the subtle, high-frequency artifacts introduced by modern GAN and diffusion-based synthesis, a limitation that drove the field's transition toward learned feature representations using deep learning methods. [10]

### 7.1. Classical CNN Architecture

Convolutional Neural Networks (CNNs):have served as the foundational backbone of deepfake forensics, extracting discriminative features and identifying subtle visual artifacts, including inconsistencies in texture, lighting, and blending boundaries [20], [45]. Early detection frameworks employed VGG architectures, leveraging their deeply stacked hierarchical layers and small-kernel convolutional filters to capture fine-grained manipulation artifacts [49]. However, their large parameter count constrains real-time detection deployment [50]. To mitigate the vanishing gradient problem in deeper networks, ResNet introduced the concept of residual learning., enabling the extraction of more complex features across many video compression levels. [51, 52] While ResNet effectively identifies "face warping artifacts " its high computational demands often make it impractical in resource-constrained environments and real-time [53]. Xception is now widely regarded as the state-of-the-art backbone

for deepfake detection, and uses depth-wise separable convolutions to handle spatial and cross-channel data separately. which makes it more efficient in detecting alterations even in high-compression videos, it consistently outperforms older models such as VGG and ResNet in standard benchmarks. [51] [54]

## 7.2. Lightweight Backbone Architecture

The deployment of deepfake detection models in resource-constrained environments requires a deliberate trade-off between architectural depth and computational efficiency. [55] To address this challenge, many lightweight backbone models have been discussed in the literature, one of the most widely discussed is **first The MobileNet family (V1, V2, V3)** This model involves replacing standard convolutions with Depthwise Separable Convolutions, significantly reducing operations while maintaining competitive feature extraction. [56] [48] [57] MobileNetV1 focused on cost-effective processing by spatial factorization, [10] whereas MobileNetV2 improved feature preservation through Inverted Residuals and Linear Bottlenecks (MBConv modules), which are critical for identifying subtle facial artifacts. [48], [58] Second, a "Compound Scaling" approach was presented by EfficientNet (B0 and B3), which consistently balanced network depth, breadth, and resolution [6]. Using Squeeze-and-Excitation (SE) blocks for channel-wise feature recalibration, EfficientNet-B0 strikes an impressive balance, achieving approximately 92.45% accuracy with just 0.45 GFLOPs. [57]EfficientNet-B3 is a deeper option that offers improved representational capacity for capturing complicated skin texture anomalies and face motion dependencies for situations where memory is allowed.[20]

## 7.3. Model Compression and Optimization Techniques

Bridging the gap between high-performance deep learning and real-time edge deployment requires a range of model-optimization strategies aimed at reducing computational demands while preserving detection accuracy. Among these, weight pruning improves model efficiency by eliminating redundant parameters and filters that contribute marginally to the classification output [60 Quantization further reduces memory footprint and accelerates inference by lowering numerical precision — for instance, transitioning from FP32 to INT8 — yielding reductions in model size of up to 12.4× and inference latency improvements of up to 10.8×. Post-training, Batch Normalization (BN) folding is commonly applied to absorb normalization parameters into the preceding convolutional layers, thereby simplifying the inference pipeline. To offset the reduced representational capacity of compressed models, Knowledge Distillation (KD) has emerged as a widely adopted technique, wherein a lightweight student model is trained to replicate the intermediate feature representations of a more complex teacher network. Recent innovations, such as Graph-based Relational Reasoning frameworks (e.g., spatial-spectral temporal graph neural networks (SSTGNN)), demonstrate that such optimized architectures can attain competitive, state-of-the-art performance with up to 42× fewer parameters, rendering them well-suited for forensic applications in resource-constrained settings. [59]

## 7.4. Transfer Learning

Transfer Learning (TL) has established itself as a cornerstone methodology in deepfake detection, enabling models to transfer knowledge acquired from large-scale source domains, such as ImageNet, to specialized forensic detection tasks. [1] Rather than training models from random initialization, fine-tuning facilitates the reuse of rich, generalized feature representations, considerably reducing both the computational requirements and the volume of labeled training data needed [11] [60]. In practice, pre-trained backbone architectures, including Xception, ResNet, and MobileNet, are widely adopted as feature extractors, given their exposure to millions of diverse images during pre-training. This prior knowledge enables them to recognize complex spatial patterns and subtle manipulation artifacts more effectively than models trained exclusively on deepfake-specific datasets. [55], [48]. However, adapting these large pretrained weights to meet the strict efficiency and real-time performance requirements of mobile forensics remains a persistent challenge [54] [51] [55].

## 7.5. Domain adaptation

Domain adaptation addresses the critical performance degradation observed when detection models are applied to datasets outside their training distribution, with a consistent drop of 10–20% in AUC reported across cross-dataset evaluations [6] [10]. Two principal strategies have emerged: (1) adversarial domain alignment, which trains a domain discriminator to align feature distributions across source and target domains, thereby reducing dataset-specific bias, and (2) Maximum Mean Discrepancy (MMD), which minimizes the statistical distance between source and target feature representations. While these approaches can recover 5–12% of the generalization gap, they require access to target-domain data during training, which is a fundamental constraint in real-world forensic scenarios where the forgery method is unknown a priori [39]. Furthermore, self-supervised pre-training methods, such as contrastive learning and masked autoencoders, have emerged as promising alternatives, enabling models to learn rich representations

without requiring large labeled deepfake corpora; however, their application in this domain remains underexplored. [42] [1] [61]

## 8. RECENT ADVANCEMENTS IN DEEPFAKE DETECTION

Recent studies have focused toward developing architectures that balance high detection accuracy with computational efficiency for real-time applications. These studies can be categorized into three primary research directions.

### 8.1. Hybrid Spatio-Temporal Frameworks:

A significant body of work focuses on capturing both the frame-level artifacts and temporal inconsistencies. Chikkanna et al. (2024) [62]

Maguluri et al. (2025) [61] utilized Multi-task Cascaded Convolutional Networks (MTCNN) for face localization, followed by EfficientNet/CNN

and LSTM networks, achieving accuracies of up to 97.9%. Similarly, Maheswari et al. (2024) [63] and AlMuhaideb et al. (2025) [56] optimized this hybrid approach using MobileNet and Gated Recurrent Units (GRU) to reduce latency for social media and mobile deployment. Meanwhile, Cyril et al. (2025) [64] and Zafar et al. (2025) [47]demonstrated high precision using Xception/inception-based hybrids, but noted that the high computational complexity of these backbones remains a barrier to edge-device integration.

### 8.2. Lightweight & Edge-Optimized Architectures:

Researchers have introduced specialized lightweight models to address resource constraints. Tauseef et al. (2025) [65] developed "Deep FaceGuard," leveraging TensorFlow Lite to achieve 45 FPS on edge devices, while Siddiqui and Kim (2025) [66] utilized HOG+LBP features with XGBoost to prove that non-deep learning classifiers can match complex models in efficiency. Furthermore, Lv et al. [67] proposed SFMFNet, which reduces computational costs by 87.9% through spatial-frequency fusion, and [58] (2025) introduced "HybridNet" achieving 97% accuracy with only 7.2M parameters by capturing spectral-spatial artifacts.

### 8.3. Specialized Detection & Robustness Strategies:

Innovative approaches have emerged to improve generalization and interpretability of these models. Farooq et al. (2025) [68] fused facial landmarks with biological heart rate features to avoid the "black-box" nature of CNNs, and Petmezas et al. (2025) [69] utilized 3D Morphable Models (3DMM) to detect identity mismatches. To enhance reliability in noisy IoT environments, Zhou et al. (2024) [70] introduced a Bayesian Inference Weighting Bayesian Inference Weighting (BIW) technique. Sharma et al. (2024) [71] focused on spectral anomalies in video conferencing via DCGAN discriminators. Lastly, Murty et al. (2025) [72] expanded the hybrid scope to audio deepfakes, achieving 99.84% accuracy using Mel-Frequency Cepstral Coefficients (MFCC) and GRU units.

**Table 4** summarizes the technical specifications, datasets, and performance metrics of these seminal studies, providing a comparative overview of the current SOTA methodologies.

## 9. DISCUSSION

A comprehensive analysis of the recent literature reveals a fundamental paradigm shift in the field of deepfake forensics. While earlier detection frameworks prioritized maximizing classification accuracy using computationally intensive architectures, research has increasingly moved toward lightweight hybrid architectures optimized for edge deployment and real-time detection.

**Accuracy vs. Efficiency:** Lightweight backbone architectures, such as MobileNet, when integrated with temporal sequence modules like (LSTM/GRU), have demonstrated high accuracies exceeding 96% [56], [63]. Sustaining genuine real-time inference remains a formidable challenge, which is largely attributable to the sequential processing latency inherent in the recurrent components. Consequently, hardware-aware optimization strategies, including operator fusion and quantization-aware training, are increasingly indispensable for ensuring efficient model execution on mobile and edge processors [59] [11].

**Generalization Gap:** While contemporary models perform well on standard benchmarks, their accuracy tends to degrade by 10–20% when subjected to compression artifacts when exposed to real-world social media compression [67]. Including spatial frequency fusion and biological signals (such as heart rate or eye blinking) is a promising strategy, as such physiological signals are inherently difficult for generative models to reliably synthesize [68].

**Explainability in Deepfake Detection:** future trends in deepfake detection A progressive transition from what is called 'black-box' systems towards Explainable AI (XAI) frameworks capable of producing transparent and interpretable detection decisions [73]. Current methodologies take advantage of visualization tools such as Grad-CAM and LIME to generate spatially localized heatmaps capable of identifying manipulated facial regions alongside interpretable architectures. However, significant research challenges remain in balancing the trade-off between model interpretability and cross-dataset generalization. [8] Eventually, integrating XAI with lightweight models is

**Table 4.** Summary of Recent Deepfake Detection Methods

| Reference | Year | Architecture (Backbone + Hybrid) | Dataset | Accuracy / AUC | Speed / Efficiency | Target Platform |
|---|---|---|---|---|---|---|
| [74] | 2024 | MesoNet4 + ResNet101 | FF++ | 98.7% | 30–50 FPS | GPU-based |
| [71] | 2024 | CED-DCGAN (Ensemble) | GAN-Gen | 98.2% | High Resource | Workstation |
| [62] | 2024 | EfficientNet-B0 + LSTM | Custom | 97.9% | Lightweight | Real-time Edge |
| [68] | 2025 | XGBoost + Heart Rate Features | WLDR | 0.95 AUC | Interpretable | Forensic Lab |
| [67] | 2024 | SFMFNet (Spatial-Frequency) | FF++, etc. | 0.86 AUC | 87.9% Cost ↓ | Real-time |
| [61] | 2025 | CNN + LSTM (MTCNN) | FF++, DFDC | 96.4% | Not Reported | Server-side |
| [69] | 2025 | 3DMM + CNN-LSTM-Trans. | VoxCeleb2 | 97% AUC | High Complexity | Identity Verif. |
| [65] | 2025 | DeepFaceGuard (EffNet/Xception) | FF++, DFDC | 96.5% | 45 FPS (TFLite) | Mobile Edge |
| [66] | 2025 | HOG+LBP+KAZE + XGBoost | FF++, C-DF | 92%–96% | Highly Efficient | Mobile Devices |
| [70] | 2024 | Bayesian Inference Weighting | IoT-Data | Robustness-focu AUC gain repor vs. baseline | Robustness Focus | IoT / Edge |
| [47] | 2025 | InceptionResNetV2 + LSTM | Custom | 97% | High Complexity | High-end PC |
| [64] | 2025 | Xception + LSTM | FF++, C-DF | 82% (Cross) | Web-App | Web Service |
| [72] | 2025 | CNN + LSTM + GRU (Audio) | In-the-Wild | 99.8% | High Speed | Audio Forensics |
| [63] | 2024 | MobileNet + LSTM + Grad-CAM | Custom | 98% | Low Latency | Law Enforcement |
| [56] | 2025 | LightFakeDetect (MBNet+GRU) | C-DF, DFDC | 98.2% | Efficient | Mobile Edge |
| [48] | 2025 | HybridNet (Freq-Spatial) | C-DF, FF++ | 97% | 7.2M Params | Edge Devices |

vital to ensure the reliability of real-time detectors needed to satisfy both computational efficiency and the severe evidentiary standards required in digital forensics.

## 10. CONCLUSION

This survey examined 75 peer-reviewed studies published between 2020 and 2026 and identified a clear directional shift in deepfake detection. The field has progressively moved away from computationally intensive single-modal CNN classifiers toward hybrid spatiotemporal architectures that combine CNN backbones with LSTM, GRU, or Transformer modules to capture both spatial and temporal manipulation cues. Concurrently, the adoption of efficient backbones, such as MobileNet and EfficientNet, paired with compression strategies, including pruning and quantization, has made high-accuracy detection increasingly viable in resource-constrained and real-time deployment environments. However, several critical challenges remain. Cross-dataset generalization remains a significant bottleneck, with detection accuracy consistently degrading by 10–20% when models are applied beyond their training distribution, as models often fail when exposed to unseen forgery types or diffusion-based synthetic media. Furthermore, the "black-box" nature of current detectors limits their reliability in legal and forensic contexts.

Several limitations of this review point directly to productive avenues for future research in this area. The temporal scope of 2020–2026, while appropriate for capturing the current generation of lightweight architectures, leaves foundational pre-2020 contributions unexamined. The heterogeneity of evaluation protocols across the included studies, including inconsistent datasets, metrics, and experimental conditions, also prevented direct quantitative synthesis, reinforcing the need for community-wide standardization of benchmarking practices. Most critically, this survey identified the absence of unified frameworks that simultaneously achieve computational efficiency, cross-domain robustness, and uncertainty quantification as the most consequential open gap in the field. Future work should treat these not as three separate challenges but as a single integrated design objective, pursued through multimodal detection strategies spanning visual, audio, and textual modalities alongside deeper integration of Explainable AI. As generative

technologies continue to advance in sophistication and accessibility, developing deployable, interpretable, and generalizable detection systems remains an urgent imperative for digital security, media integrity, and public trust.

## REFERENCES


[1] A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *WIREs Data Min. Knowl. Discov.*, 2024. DOI: 10.1002/widm.1520.

[2] R. Chauhan, I. Kansal, R. Popli, R. Kumar, and A. Sharma, "Current state of deepfake detection and generation: A review," *Recent Adv. Electr. Electron. Eng.*, vol. 18, no. 5, pp. 515–539, Jan. 2025. DOI: 10.2174/0123520965269334231117112747.

[3] V. K. Sharma, R. Garg, and Q. Caudron, "A systematic literature review on deepfake detection techniques," *Multimed. Tools Appl.*, vol. 84, no. 20, pp. 22 187–22 229, Aug. 2024. DOI: 10.1007/S11042-024-19906-1.

[4] R. Sunil, P. Mer, A. Diwan, R. Mahadeva, and A. Sharma, "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation," *Heliyon*, Feb. 2025. DOI: 10.1016/j.heliyon.2025.e42273.

[5] L. A. Passos et al., "A review of deep learning-based approaches for deepfake content detection," *Expert Syst.*, vol. 41, no. 8, Feb. 2024. DOI: 10.1111/EXSY.13570.

[6] S. A. Khan and D. T. Dang-Nguyen, "Deepfake detection: Analyzing model generalization across architectures, datasets, and pre-training paradigms," *IEEE Access*, vol. 12, pp. 1880–1908, 2024. DOI: 10.1109/access.2023.3348450.

[7] G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad, "A comprehensive review of deepfake detection using advanced machine learning and fusion methods," *Electronics*, Jan. 2024. DOI: 10.3390/electronics13010095.

[8] P. Edwards, J. C. Nebel, D. Greenhill, and X. Liang, "A review of deepfake techniques: Architecture, detection, and datasets," *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3477257.

[9] M. Alrashoud, "Deepfake video detection methods, approaches, and challenges," *Alex. Eng. J.*, Jun. 2025. DOI: 10.1016/j.aej.2025.04.007.

[10] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25 494–25 513, 2022. DOI: 10.1109/ACCESS.2022.3154404.

[11] A. Karathanasis, J. Violos, and I. Kompatsiaris, "A comparative analysis of compression and transfer learning techniques in deepfake detection models," *Mathematics*, vol. 13, no. 5, p. 887, Mar. 2025. DOI: 10.3390/math13050887.

[12] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, "A survey on the detection and impacts of deepfakes in visual, audio, and textual formats," *IEEE Access*, vol. 11, pp. 144 497–144 529, 2023. DOI: 10.1109/ACCESS.2023.3344653.

[13] J. Kang, S. K. Ji, S. Lee, D. Jang, and J. U. Hou, "Detection enhancement for various deepfake types based on residual noise and manipulation traces," *IEEE Access*, vol. 10, pp. 69 031–69 040, 2022. DOI: 10.1109/ACCESS.2022.3185121.

[14] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, unknown.

[15] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on discrepancies between faces and their context," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6111–6121, 2022. DOI: 10.1109/TPAMI.2021.3093446.

[16] G. S. Hsu, C. H. Tsai, and H. Y. Wu, "Dual-generator face reenactment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 632–640. DOI: 10.1109/CVPR52688.2022.00072.

[17] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," *Commun. ACM*, vol. 62, no. 1, pp. 96–104, 2020. DOI: 10.1145/3292039.

[18] N. Li and B. A. Plummer, *Supervised attribute information removal and reconstruction for image manipulation*, arXiv preprint arXiv:2207.06555, 2022. [Online]. Available: http://arxiv.org/abs/2207.06555.

[19] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020. DOI: 10.1016/j.inffus.2020.06.014.

[20] T. Zhang, L. Deng, L. Zhang, and X. Dang, "Deep learning in face synthesis: A survey on deepfakes," in *2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET)*, IEEE, Aug. 2020, pp. 67–70. DOI: 10.1109/CCET50901.2020.9213159.

[21] Z. Akhtar, "Deepfakes generation and detection: A short survey," *J. Imaging*, vol. 9, no. 1, p. 18, Jan. 2023. DOI: 10.3390/jimaging9010018.

[22] F. Sharevski, A. Zeidieh, J. V. Loop, and P. Jachim, "Blind and low vision individuals' detection of audio deepfakes," in *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, ACM, Dec. 2024, pp. 4867–4881. DOI: 10.1145/3658644.3690305.

[23] P. Dhabe, N. Choudhary, A. Vidhale, Y. Munde, M. Sayyed, and N. Zanwar, "Advanced sequential modeling for deepfake audio identification," *Int. J. for Res. Appl. Sci. Eng. Technol. (IJRASET)*, vol. 12, 2024. DOI: 10.22214/ijraset.2024.65027.

[24] R. Walambe, P. Chaudhary, A. Bajaj, A. S. Rathore, V. Jain, and K. Kotecha, "Generative adversarial networks for mitigating bias in disinformation," in *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, IEEE, 2023. DOI: 10.1109/INC457730.2023.10262880.

[25] M. Abdel-Basset, N. Moustafa, and H. Hawash, "Generative adversarial networks (gans)," in *Deep Learning Approaches for Security Threats in IoT Environments*, John Wiley & Sons, Nov. 2022, pp. 271–285. DOI: 10.1002/9781119884170.ch12.

[26] M. R. Rahman, S. Ummie, M. E. Rahman, M. N. A. Siddiky, and M. R. Rahman, "An overview of generative adversarial networks and their variants," *Preprints*, Apr. 2025. DOI: 10.20944/preprints202504.0071.v1.

[27] W. Sun, "Generative adversarial networks (gans): Principles, development, challenges, and diverse applications," *Appl. Comput. Eng.*, vol. 165, no. 1, pp. 152–158, Jul. 2025. DOI: 10.54254/2755-2721/2025.LD25482.

[28] Preeti, M. Kumar, and H. K. Sharma, "A gan-based model of deepfake detection in social media," *Procedia Comput. Sci.*, vol. 218, pp. 2153–2162, Jan. 2023. DOI: 10.1016/j.procs.2023.01.191.

[29] D. P. Kingma and M. Welling, *An Introduction to Variational Autoencoders* (Foundations and Trends in Machine Learning). Now Publishers, 2019. DOI: 10.1561/2200000056.

[30] C. Lei, "Unsupervised learning: Deep generative model," in *Machine Learning and Artificial Intelligence*, Springer, 2021, pp. 183–215. DOI: 10.1007/978-981-16-2233-5_9.

[31] F. Ye and A. G. Bors, "Deep mixture generative autoencoders," *IEEE Trans. on Neural Networks Learn. Syst.*, vol. 33, no. 10, pp. 5789–5803, Oct. 2022. DOI: 10.1109/TNNLS.2021.3071401.

[32] C. Bhattacharyya, H. Wang, F. Zhang, S. Kim, and X. Zhu, "Diffusion deepfake," *arXiv preprint arXiv:2404.01579*, Apr. 2024. arXiv: 2404.01579 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2404.01579.

[33] J. Ricker, S. Damm, T. Holz, and A. Fischer, "Towards the detection of diffusion model deepfakes," in *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, vol. 4, 2024, pp. 446–457. DOI: 10.5220/0012422000003660.

[34] Y. Chen, N. A. H. Haldar, N. Akhtar, and A. Mian, "Text-image guided diffusion model for generating deepfake celebrity interactions," in *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2023, pp. 348–355. DOI: 10.1109/DICTA60407.2023.00055.

[35] G. Pei et al., "Deepfake generation and detection: A benchmark and survey," *arXiv preprint arXiv:2403.17881*, Mar. 2024. arXiv: 2403.17881 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2403.17881.

[36] B. C. Soundarya and H. L. Gururaj, "Deepfake detection: Critical review of state-of-the-art approaches and future perspectives," *Discov. Appl. Sci.*, Jan. 2026. DOI: 10.1007/s42452-025-08174-9.

[37] M. Alrashoud, "Deepfake video detection methods, approaches, and challenges," *Alex. Eng. J.*, vol. 125, pp. 265–277, Jun. 2025. DOI: 10.1016/j.aej.2025.04.007.

[38] D. A. Dung and H. T. T. Binh, "Gdegan: Graphical discriminative embedding gan for tabular data," in *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2022. DOI: 10.1109/DSAA54385.2022.10032445.

[39] F. Abbas and A. Taeihagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," *Expert Syst. with Appl.*, vol. 252, p. 124 260, Oct. 2024. DOI: 10.1016/j.eswa.2024.124260.

[40] R. Sunil, P. Mer, A. Diwan, R. Mahadeva, and A. Sharma, "Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation," *Heliyon*, vol. 11, no. 3, e42273, Feb. 2025. DOI: 10.1016/j.heliyon.2025.e42273.

[41] S. K. Sharma, A. AlEnizi, M. Kumar, O. Alfarraj, and M. Alowaidi, "Detection of real-time deep fakes and face forgery in video conferencing employing generative adversarial networks," *Heliyon*, vol. 10, no. 17, e37163, Sep. 2024. DOI: 10.1016/j.heliyon.2024.e37163.

[42] O. A. H. H. Al-Dulaimi and S. Kurnaz, "A hybrid cnn-lstm approach for precision deepfake image detection based on transfer learning," *Electronics*, vol. 13, no. 9, May 2024. DOI: 10.3390/electronics13091662.

[43] G. Pei et al., "Deepfake generation and detection: A benchmark and survey," *arXiv preprint arXiv:2403.17881*, May 2024. arXiv: 2403.17881. [Online]. Available: https://arxiv.org/abs/2403.17881.

[44] Y. Patel et al., "Deepfake generation and detection: Case study and challenges," *IEEE Access*, vol. 11, pp. 143 296–143 323, 2023. DOI: 10.1109/ACCESS.2023.3342107.

[45] R. Khan et al., "Comparative study of deep learning techniques for deepfake video detection," *ICT Express*, Dec. 2024. DOI: 10.1016/j.icte.2024.09.018.

[46] D. Amir, L. M. Ibrahim, D. A. Sultan, and L. M. Ibrahim, "A comprehensive survey on deepfake detection techniques," *Int. J. Intell. Syst. Appl. Eng.*, 2023, Available online: https://www.researchgate.net/publication/376523314.

[47] F. Zafar, T. A. Khan, S. Akbar, M. T. Ubaid, S. Javaid, and K. A. Kadir, "A hybrid deep learning framework for deepfake detection using temporal and spatial features," *IEEE Access*, vol. 13, pp. 79 560–79 570, 2025. DOI: 10.1109/ACCESS.2025.3566008.

[48] A. Kumar, "Hybridnet: Advancing deepfake detection through residual, se, and depthwise convolutions," *IEEE Access*, vol. 13, pp. 207 541–207 552, 2025. DOI: 10.1109/ACCESS.2025.3640106.

[49] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, IEEE, 2015, pp. 730–734. DOI: 10.1109/ACPR.2015.7486599.

[50] T. T. Nguyen et al., "Deep learning for deepfakes creation and detection: A survey," *Comput. Vis. Image Underst.*, vol. 223, p. 103 525, Oct. 2022. DOI: 10.1016/j.cviu.2022.103525.

[51] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 1–11. DOI: 10.1109/ICCV.2019.00009.

[52] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2020, pp. 3204–3213. DOI: 10.1109/CVPR42600.2020.00327.

[53] R. Lanzino, F. Fontana, A. Diko, M. R. Marini, and L. Cinque, "Faster than lies: Real-time deepfake detection using binary neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2024, pp. 3771–3780. DOI: 10.1109/CVPRW63382.2024.00381.

[54] S. Degadwala and V. M. Patel, "Advancements in deepfake detection: A review of emerging techniques and technologies," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 5, pp. 127–139, Nov. 2024. DOI: 10.32628/CSEIT24105811.

[55] T. Shahriar, "Comparative analysis of lightweight deep learning models for memory-constrained devices," *arXiv preprint arXiv:2505.03303*, May 2025. arXiv: 2505.03303. [Online]. Available: https://arxiv.org/abs/2505.03303.

[56] S. AlMuhaideb, H. Alshaya, L. Almutairi, D. Alomran, and S. T. Alhamed, "Lightfakedetect: A lightweight model for deepfake detection in videos that focuses on facial regions," *Mathematics*, vol. 13, no. 19, p. 3088, Sep. 2025. DOI: 10.3390/math13193088.

[57] F. Zafar, T. A. Khan, S. Akbar, M. T. Ubaid, S. Javaid, and K. A. Kadir, "A hybrid deep learning framework for deepfake detection using temporal and spatial features," *IEEE Access*, vol. 13, pp. 79 560–79 570, 2025. DOI: 10.1109/ACCESS.2025.3566008.

[58] S. Kaur, P. Kumar, and P. Kumaraguru, "Deepfakes: Temporal sequential analysis to detect face-swapped video clips using convolutional long short-term memory," *J. Electron. Imaging*, vol. 29, 2020. DOI: 10.1117/1.JEI.29.

[59] R. Tao, Z. Qin, Y. Ding, C. Tan, J. Wang, and W. Wang, "Unlocking the potential of lightweight quantized models for deepfake detection," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Sep. 2025, pp. 520–528. DOI: 10.24963/IJCAI.2025/59.

[60] S. Suratkar and F. Kazi, "Deep fake video detection using transfer learning approach," *Arab. J. for Sci. Eng.*, vol. 48, no. 8, pp. 9727–9737, Aug. 2023. DOI: 10.1007/s13369-022-07321-3.

[61] L. V. M. Maguluri, H. N. V. Kothamasu, and S. D. Johnson, "Hybrid deepfake detection using cnn for spatial analysis and lstm for temporal consistency," *Int. J. Innov. Sci. Res. Technol.*, pp. 381–387, May 2025. DOI: 10.38124/IJISRT/25MAY346.

[62] S. Chikkanna, C. Mrutyunjaya, R. Manjappa, C. M. Lakshmanagowda, R. T. Venkatesh, and S. G. Murthy, "A hybrid deep learning framework for multimodal deepfake detection," *Int. J. Saf. Secur. Eng.*, vol. 15, no. 10, pp. 1995–2004, Oct. 2025. DOI: 10.18280/ijsse.151002.

[63] S. Maheswari, V. D. Kumar, D. A. Kumar, R. Jahnavi, and C. Laharee, "Real-time deepfake detection using a hybrid mobilenet-lstm model for enhanced media integrity," in *Advances in Intelligent Systems and Computing*, Atlantis Press, 2025, pp. 980–991. DOI: 10.2991/978-94-6463-866-0_79.

[64] M. A. Cyril, J. P. Sunkavalli, and D. Prasanth, "Deepguard: A hybrid cnn-lstm framework for robust deepfake video detection with spatiotemporal analysis," in *Proceedings of the 3rd International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, IEEE, 2025, pp. 2071–2076. DOI: 10.1109/ICSCDS65426.2025.11166809.

[65] M. Tauseef, L. S. Viji, and J. Fenwick, "Deepfaceguard: A lightweight cnn framework for robust face manipulation detection," in *3rd IEEE International Conference on Networks, Multimedia and Information Technology (NMITCON)*, IEEE, 2025. DOI: 10.1109/NMITCON65824.2025.11188219.

[66] S. M. Yasir and H. Kim, "Lightweight deepfake detection based on multi-feature fusion," *Appl. Sci.*, vol. 15, no. 4, p. 1954, Feb. 2025. DOI: 10.3390/app15041954.

[67] L. Lv, T. Wang, M. Huang, R. Liu, and Y. Wang, "A spatial-frequency aware multi-scale fusion network for real-time deepfake detection," in Springer, 2026, pp. 536–550. DOI: 10.1007/978-981-95-5676-2_36.

[68] M. U. Farooq, A. Javed, K. M. Malik, and M. A. Raza, "A lightweight and interpretable deepfakes detection framework," *arXiv preprint arXiv:2501.11927*, Jan. 2025. arXiv: 2501.11927. [Online]. Available: https://arxiv.org/abs/2501.11927.

[69] G. Petmezas, V. Vanian, K. Konstantoudakis, E. E. I. Almaloglou, and D. Zarpalas, "Video deepfake detection using a hybrid cnn-lstm-transformer model for identity verification," *Multimed. Tools Appl.*, vol. 84, no. 33, pp. 40 617–40 636, Mar. 2025. DOI: 10.1007/s11042-024-20548-6.

[70] L. Zhou, C. Ma, Z. Wang, Y. Zhang, X. Shi, and L. Wu, "Robust frame-level detection for deepfake videos with lightweight bayesian inference weighting," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 13 018–13 028, Apr. 2024. DOI: 10.1109/JIOT.2023.3337128.

[71] S. K. Sharma, A. AlEnizi, M. Kumar, O. Alfarraj, and M. Alowaidi, "Detection of real-time deep fakes and face forgery in video conferencing employing generative adversarial networks," *Heliyon*, vol. 10, no. 17, e37163, Sep. 2024. DOI: 10.1016/j.heliyon.2024.e37163.

[72] M. Murty, S. Tomar, and S. G. Koolagudi, "A hybrid deep learning framework for real and deepfake voice detection," *Circuits, Syst. Signal Process.*, pp. 1–27, Jan. 2026. DOI: 10.1007/s00034-025-03464-4.

[73] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Appl. Intell.*, vol. 53, no. 4, pp. 3974–4026, 2023. DOI: 10.1007/s10489-022-03766-z.

[74] M. Javed, Z. Zhang, F. H. Dahri, and A. A. Laghari, "Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach," *Electronics*, vol. 13, no. 15, p. 2947, Aug. 2024. DOI: 10 . 3390 / electronics13152947.