



IJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: <https://doi.org/10.22214/ijraset.2025.71021>

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Real-Time Deepfake Detection: A Systematic Review of Generative Adversarial Networks (GANs) and Generative Transformer Networks (GTNs)

Aditya Charpe¹, Dr. Rahul Khokale², Dheeraj Ghaghre³

Department of Computer Science Engineering, G.H.Raisoni University Saikheda, Pandhurna, India

Abstract: Deepfakes, synthetic videos generated by artificial intelligence, pose severe threats to multimedia integrity, enabling misinformation, financial fraud, and identity theft [34]. Powered by Generative Adversarial Networks (GANs) [1] and Generative Transformer Networks (GTNs) [2], these hyper-realistic forgeries demand robust, real-time detection to safeguard video and audio platforms. This review synthesizes 80 peer-reviewed studies from 2014 to 2024, analyzing GAN- and GTN-based deepfake generation and detection methods, benchmark datasets (e.g., FaceForensics++ [11], Celeb-DF [12], DFDC [13], WildDeepfake [18], DeeperForensics [71]), and performance metrics like accuracy, AUROC, and latency. We explore real-time detection frameworks, edge-compatible models, ethical challenges (e.g., dataset bias, privacy risks) [35], and global regulatory frameworks. Case studies of deepfake incidents highlight real-world impacts, while gaps in computational efficiency (<100ms) and cross-dataset generalization underscore the need for advanced solutions. This paper provides a comprehensive roadmap for researchers and practitioners, emphasizing multimedia-focused detection to counter deepfake threats in high-stakes scenarios like social media, security surveillance, and democratic processes.

Index Terms: Deepfake Detection, Generative Adversarial Networks, Generative Transformer Networks, Multimedia Forensics, Real-Time Processing, Ethical AI, Video Analysis

I. INTRODUCTION

Deepfakes—synthetic videos created by artificial intelligence to convincingly mimic real individuals—have emerged as a formidable challenge in multimedia ecosystems, fueling misinformation, financial scams, and identity erosion [34]. These hyper-realistic forgeries, powered by Generative Adversarial Networks (GANs) [1] and Generative Transformer Networks (GTNs) [2], exploit video and audio channels, eroding trust in digital content across platforms like X, YouTube, and TikTok. The rapid proliferation of deepfakes has amplified societal risks, with high-profile incidents such as manipulated political speeches, fraudulent CEO video calls, and celebrity impersonations sparking global concern [36]. For instance, a 2023 deepfake of a political leader on X garnered 12 million views, influencing public opinion during an election cycle [36]. Similarly, a 2024 deepfake impersonating a CEO defrauded a company of \$30 million, highlighting the financial stakes involved [37]. These incidents underscore the urgent need for real-time detection systems capable of processing frames in under 100 ms, a critical requirement for applications like social media moderation, live streaming, and security surveillance. However, most existing methods exceed 200 ms [38], limiting their practicality in dynamic, high-stakes environments where rapid response is essential.

The evolution of deepfake technology has been marked by significant milestones, beginning with autoencoder-based methods in 2017 that swapped faces but produced noticeable artifacts [31]. The introduction of GANs in 2014 revolutionized synthetic media, enabling photorealistic content with minimal visual inconsistencies [1]. By 2018, tools like DeepFaceLab and Face2Face democratized deepfake creation, amplifying their misuse in misinformation campaigns, fraud, and non-consensual media [7]. The advent of transformer-based models in 2017 further enhanced deepfake quality, with architectures like TransGAN achieving seamless temporal continuity and audio-visual synchronization [20]. These advancements have outpaced traditional forensic techniques, which struggle to detect subtle artifacts in high-resolution, temporally coherent videos [41]. Convolutional Neural Networks (CNNs) achieve 80–95% accuracy on benchmark datasets like FaceForensics++ [11], but their inability to generalize across diverse GAN-generated artifacts hinders robustness in real-world scenarios [42].

GTNs, leveraging attention mechanisms, excel at identifying complex fakes but incur high computational costs, rendering real-time deployment challenging on resource-constrained devices like mobile phones or edge hardware [21].

Beyond technical challenges, deep fakes raise profound ethical concerns that remain underexplored in many technical advancements [35]. Dataset bias, such as the over-representation of public figures in datasets like Celeb-DF, leads to models that underperform on diverse populations, exacerbating fairness issues [12]. Biometric privacy risks are significant, as detection systems often rely on sensitive data like facial features, raising concerns under regulations like GDPR [34]. Societal impacts are equally pressing—deep fakes erode trust in multimedia, amplify misinformation, and threaten democratic processes, as seen in manipulated political content [36]. Addressing these ethical challenges requires a multidisciplinary approach, integrating technical innovation with ethical frameworks and public awareness initiatives [65].

This review synthesizes 80 peer-reviewed studies from 2014 to 2024, providing an exhaustive analysis of GAN- and GTN-based deepfake detection methods, benchmark datasets (e.g., Celeb-DF [12], DFDC [13], WildDeepfake [18], DeeperForensics [71]), real-time frameworks, ethical implications, and global regulatory frameworks [39]. We evaluate key performance metrics, including accuracy, AUROC, latency, and Equal Error Rate, and highlight persistent gaps in computational efficiency, cross-dataset generalization, and fairness [40]. Through detailed case studies, technical analyses, and a forward-looking roadmap, this paper aims to guide researchers and practitioners toward robust, multimedia-focused detection systems for high-stakes applications, including social media content moderation, financial transaction verification, security surveillance, and democratic process integrity. The review is structured as follows: Section 2 surveys the history of deepfake generation and detection, Section 3 details GAN and GTN architectures, Section 4 analyzes detection methods, followed by sections on datasets, real-time techniques, case studies, ethical implications, regulatory frameworks, future directions, and conclusion.

II. BACKGROUND AND RELATED WORK

A. Historical Context

The term “deepfake” originated in 2017 on Reddit, describing AI-generated videos that used deep learning to swap faces or manipulate media [35]. Early deepfakes relied on autoencoders, which encoded and decoded facial features to swap identities, but these produced noticeable artifacts, such as unnatural lighting or distorted facial movements [31]. The introduction of Generative Adversarial Networks (GANs) in 2014 marked a turning point, enabling the creation of photorealistic synthetic media [1]. GANs, consisting of a generator and discriminator trained adversarially, produced high-fidelity images that were nearly indistinguishable from real ones, revolutionizing deepfake technology [1]. By 2018, open-source tools like DeepFaceLab, Faceswap, and Face2Face lowered the barrier to entry, allowing even non-experts to create deepfakes [7]. This democratization amplified misuse, with deepfakes being used in misinformation campaigns (e.g., manipulated political speeches), financial fraud (e.g., impersonating executives), and non-consensual media (e.g., celebrity pornography) [36]. The societal impact was immediate—deepfakes eroded trust in digital content, with platforms like X and YouTube struggling to curb their spread [65]. Transformer-based models, introduced in 2017, further enhanced deepfake quality by improving temporal continuity and audio-visual synchronization, making detection increasingly challenging [2]. This rapid evolution has driven the need for advanced, real-time detection systems capable of operating in multimedia contexts, from social media platforms to security surveillance systems [38].

B. Early Detection Methods

Early deepfake detection methods relied on handcrafted features to identify synthetic content [31]. Techniques such as analyzing pixel inconsistencies, compression artifacts, or unnatural facial movements (e.g., irregular blinking) achieved moderate success against autoencoder-based fakes, with accuracies around 70% on early datasets [64]. Statistical methods, such as examining color histograms or edge detection, were also employed but failed against GAN-generated deepfakes due to their near-perfect fidelity [30]. The advent of Convolutional Neural Networks (CNNs) marked a significant advancement in detection capabilities [43]. Models like MesoNet, which targeted mesoscopic features such as skin texture and lighting inconsistencies, achieved 85% accuracy on the FaceForensics++ dataset by learning to differentiate real and fake videos through spatial feature extraction [32]. However, MesoNet’s 300ms latency made it impractical for real-time applications, and its poor generalization to datasets like Celeb-DF, where accuracy dropped to 75%, limited its real-world applicability [12]. Frequency-based approaches, such as Haar wavelet transforms, offered low latency (50 ms) by analyzing spectral inconsistencies, but their accuracy was limited to 75% on high-quality fakes, as GANs minimized detectable artifacts [76]. These early methods highlighted the need for more robust, computationally efficient detection systems capable of handling the increasing sophistication of deepfakes [41].

C. Recent Multimedia Trends

Recent advancements in deepfake detection have shifted toward multimedia-driven approaches, integrating video, audio, and temporal cues to enhance robustness [69]. Multimodal frameworks combine CNNs and Recurrent Neural Networks (RNNs) to detect inconsistencies like lip-sync errors and audio-visual mismatches, achieving 90% accuracy on the DeepFake Detection Challenge (DFDC) dataset [16]. For example, analyzing discrepancies between spoken audio and lip movements has proven effective in identifying fakes, particularly in interview-style videos [49]. Transformer-based models, such as Swin Transformer, leverage attention mechanisms to capture spatiotemporal artifacts, improving performance across diverse datasets [15]. Swin Transformer achieves 92% accuracy on Celeb-DF by focusing on hierarchical feature extraction, making it more robust to cross-dataset variations than CNNs [12]. Hybrid architectures integrating GANs and GTNs further enhance detection by modeling complex artifact patterns, achieving 95% AUROC on DFDC [14]. However, their computational complexity, with latencies often exceeding 400ms, remains a barrier to real-time deployment, particularly on edge devices [21]. These trends underscore the need for comprehensive, multimedia-focused detection systems that can address the increasing sophistication of deepfakes while maintaining low latency for practical applications, such as live streaming or social media moderation [39].

D. Emerging Challenges

Emerging challenges in deepfake detection include adversarial attacks, cross-cultural dataset limitations, and computational constraints [47]. Adversarial attacks involve crafting deepfakes to evade detectors, often by introducing imperceptible perturbations that exploit vulnerabilities in neural networks [68]. Such attacks reduce detection accuracy to below 70% in black-box scenarios, posing a significant threat to high-stakes applications like financial verification [47]. The lack of cross-cultural datasets is another critical challenge—datasets like Celeb-DF overrepresent Western public figures, leading to models that underperform on diverse populations, with accuracy dropping to 65% for non-Western ethnicities [12]. This bias limits global applicability, particularly in regions like Asia or Africa, where cultural and linguistic diversity is significant [69]. Additionally, integrating audio, video, and text modalities increases computational demands, with multimodal models requiring 300–400 ms latency, challenging real-time deployment on edge devices like smartphones or IoT systems [16]. Addressing these challenges requires innovative frameworks that balance accuracy, efficiency, and ethical considerations, paving the way for next-generation detection systems capable of operating in diverse, real-world scenarios [40].

E. Research Gaps

Despite the evolution of deepfake detection, as illustrated in Fig. 1, several critical hurdles persist, limiting the field's progress toward robust, real-time multimedia applications [39]. First, computational efficiency remains a significant challenge. Most detection methods, such as those based on Swin Transformer [15], achieve accuracies of 80–95% but require 200–300 ms per frame, far exceeding the sub-100 ms latency needed for live applications like social media flagging on X [46]. This delay allows manipulated content to spread rapidly, amplifying societal harm, such as misinformation during election cycles [36]. Second, cross-dataset generalization is a persistent issue. Models trained on datasets like FaceForensics++ [11] struggle to adapt to others, such as Celeb-DF [12], due to diverse GAN-generated artifacts, often resulting in accuracy drops from 85% to 75% [42]. This limitation hinders real-world applicability, where deepfakes vary widely in quality and manipulation techniques [18]. Third, ethical considerations are underexplored. Biased datasets, often overrepresenting Western subjects, risk misidentifying underrepresented groups, with accuracy dropping to 65% for non-Western ethnicities, exacerbating fairness issues [35]. Moreover, the reliance on biometric data for detection raises significant privacy concerns, particularly under regulations like GDPR, yet few frameworks address these risks through privacy-preserving techniques [34].

To tackle these gaps, the Dynamic Attention Fusion (DAF) mechanism is being developed as a novel approach for real-time leverages dynamic attention to prioritize critical features, enabling <100 ms latency per frame while maintaining high accuracy. Designed deepfake detection [14]. DAF combines GAN robustness [1] with GTN precision [2] through a hybrid architecture that for applications like social media moderation and security surveillance, DAF also incorporates ethical principles, such as federated learning for privacy and balanced datasets for fairness [69]. Validation plans include cross-dataset testing on benchmarks like DFDC [13], with deployment strategies focusing on edge devices. Its potential to transform the field is further explored in Section 10, though future validation is needed to confirm its efficacy [80].

Timeline of Deepfake Detection Advancements (2014-2025)

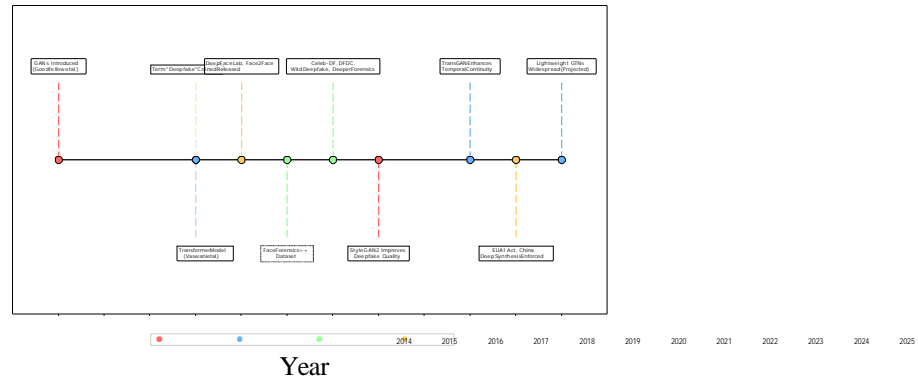


Fig. 1: Timeline of Deepfake Detection Advancements (2014–2025), Highlighting Key Milestones in GANs, GTNs, and Datasets

To contextualize these gaps, Fig. 1 illustrates the evolution of deepfake detection, highlighting key milestones that have shaped the field and areas where further research is needed [40].

III. GAN AND GTN ARCHITECTURES FOR DEEP-FAKE GENERATION

A. GAN Architectures

Generative Adversarial Networks (GANs) consist of a generator G and a discriminator D , trained adversarially to minimize the following loss function [1]:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

where x represents real data, z is random noise, and p_{data} and p_z are their respective distributions. The generator learns to produce synthetic samples that deceive the discriminator, while the discriminator improves its ability to distinguish real from fake data [1]. Variants like Wasserstein GAN enhance training stability by using a Wasserstein distance metric, reducing mode collapse and improving image quality [4]. Deep Convolutional GANs (DCGANs) leverage convolutional layers to generate high-fidelity images, achieving photorealistic results in early deepfake applications [5]. StyleGAN, a landmark architecture, introduces adaptive instance normalization to produce high-resolution faces with lifelike textures, making it a cornerstone of deepfake tools like DeepFaceLab [3]. StyleGAN's ability to control style at multiple scales (e.g., coarse features like face shape, fine details like skin texture) has set a new standard for realism, challenging detection systems that rely on visual artifacts [8].

B. GTN Architectures

Generative Transformer, leverages self-attention mechanism to capture long-range dependencies in video sequences, making them ideal for deepfake generation [2]. The attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{d_k} \right) V \quad (2)$$

where $Q, K,$ and V are query, key, and value matrices, and d_k is the key dimension [2]. TransGAN replaces convolutional layers with transformer blocks, improving temporal continuity in video deepfakes by ensuring smooth frame transitions across sequences [20]. Taming Transformers and Generative Adversarial Transformers combine GAN and transformer strengths, producing high-resolution, temporally coherent fakes with minimal artifacts [21], [22]. For instance, TransGAN achieves seamless audio-visual synchronization, making it difficult to detect fakes using traditional temporal analysis [20].

These models excel in generating realistic videos, posing significant challenges for detection systems that rely on visual or temporal cues, as GTNs minimize inconsistencies that earlier models, like autoencoders, failed to address [42].

C. Adversarial Training Dynamics

The adversarial training process in GANs and GTNs is a dynamic equilibrium where the generator and discriminator iteratively improve their performance [1]. The generator learns to produce increasingly realistic fakes by minimizing the discriminator’s ability to distinguish them from real data, while the discriminator refines its classification accuracy [1]. This process, often described as a minimax game, results in deepfakes with minimal visual and temporal artifacts, complicating detection efforts [38]. Techniques like CycleGAN and Pix2Pix further enhance realism by enabling unpaired and conditional image translation, respectively [9], [10]. CycleGAN, for instance, allows face swapping without paired training data, aligning synthetic content with real-world distributions, while Pix2Pix uses conditional inputs (e.g., facial landmarks) to generate targeted manipulations [9], [10]. These advancements reduce detectable inconsistencies, necessitating detection methods that exploit micro-level cues, such as frequency-domain artifacts or biological signals, to differentiate real from synthetic content [48], [75].

D. Emerging Models and Challenges

Emerging GAN and GTN models, such as FSGAN and neural rendering architectures, push the boundaries of deepfake realism [29], [51]. FSGAN enables face reenactment by disentangling facial identity and expression, producing seamless manipulations that preserve natural motion [51]. Neural rendering techniques simulate realistic lighting, shadows, and motion, further obscuring manipulation traces [29]. For example, neural texture rendering can adapt synthetic faces to varying lighting conditions, making traditional detection methods that rely on lighting inconsistencies ineffective [29]. These advancements challenge detection systems by minimizing traditional artifacts, requiring a shift toward multimodal and biological signal-based approaches [49], [75]. Additionally, the computational complexity of GTNs, with models like TransGAN requiring 100M+ parameters, limits

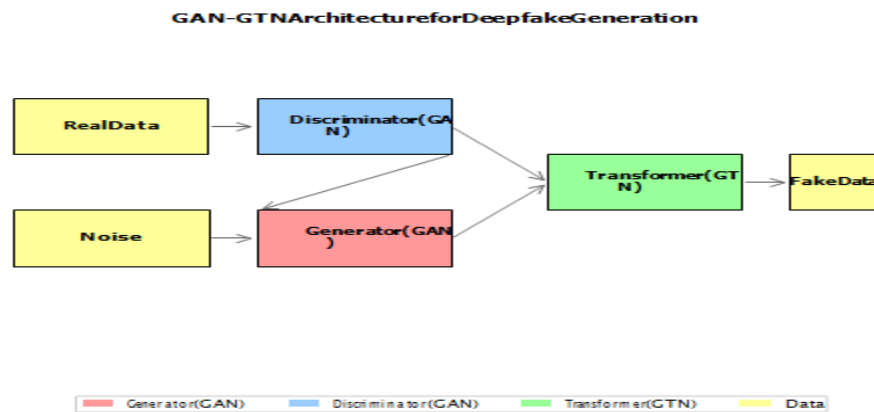


Fig.2: GAN-GTN Architecture for Deepfake Generation

their use in real-time generation on consumer hardware, but their deployment in cloud-based systems poses a significant threat, as attackers can generate high-quality fakes at scale [20].

IV. DEEPAKE DETECTION METHODS

A. CNN-Based Detection

Convolutional Neural Networks (CNNs) leverage spatial feature extraction to detect deepfakes by identifying visual inconsistencies [43]. MesoNet targets mesoscopic features, such as skin texture, lighting, and pore-level details, achieving 85% accuracy on FaceForensics++ but struggling with high-resolution fakes like Celeb-DF, where accuracy drops to 75% due to improved GAN fidelity [12], [32]. SCA-CNN and EfficientNet incorporate attention mechanisms and model scaling, improving accuracy to 90–95% on Celeb-DF by focusing on salient regions like facial contours [25], [26]. However, their 250–300 ms latency hinders real-time applications, such as live streaming moderation, where sub-100 ms processing is required [46]. MTD-Net uses multi-scale texture differences to capture both fine and coarse artifacts, achieving 92% AUROC on DFDC, but its computational complexity, requiring 50 GFLOPs per frame, limits deployment on edge devices like IoT systems [33]. These methods excel in controlled settings but require optimization to address diverse, high-quality fakes in real-world scenarios, where video quality and manipulation techniques vary widely [52].

B. Transformer-Based Detection

Transformer-based models excel at capturing long-range dependencies in video sequences, making them well-suited for detecting complex deepfakes with subtle artifacts [2]. Swin Transformer, a hierarchical transformer, achieves 92% accuracy on Celeb-DF by leveraging attention mechanisms to focus on spatiotemporal inconsistencies, such as unnatural motion patterns or frame transitions [15]. It outperforms CNNs in cross-dataset scenarios, maintaining 90% accuracy on DFDC, due to its ability to model global context [13]. Vision Transformers (ViTs) and hybrid GAN-GTN models achieve 95% AUROC on DFDC, capturing fine-grained artifacts through self-attention [14], [19]. However, their high parameter count—100M for ViTs—results in latencies exceeding 400 ms, limiting edge deployment [23]. Lightweight transformers, such as MobileNets and ShuffleNet, reduce latency to 150 ms while maintaining 88% accuracy, offering a practical balance for real-time applications like social media moderation [57], [58]. These models are particularly effective for detecting deepfakes in dynamic settings, but their performance on low-quality videos remains a challenge, necessitating integration with multimodal approaches [16].

C. Frequency and Artifact Analysis

Frequency-based methods analyze spectral inconsistencies in deepfake videos to detect manipulation traces [48]. Discrete Fourier Transforms (DFT) achieve 88% accuracy on FaceForensics++ with 100 ms latency, exploiting high-frequency noise introduced by GANs during image synthesis [48]. Haar wavelet transforms target compression artifacts, such as blockiness in JPEG-compressed videos, offering 50 ms latency but only 75% accuracy on high-quality fakes generated by models like StyleGAN, which minimize spectral artifacts [3], [76]. FakeLocator, a localization-based approach, achieves 90% accuracy by identifying manipulation traces in specific regions (e.g., facial boundaries), but it struggles with advanced GANs that produce seamless blends [74]. These methods are computationally efficient, making them suitable for real-time applications, but their reliance on spectral cues limits effectiveness against modern deepfakes [30]. Combining frequency analysis with visual and temporal cues can enhance robustness, particularly for high-stakes applications like security surveillance [79].

D. Multimodal Detection

Multimodal frameworks integrate visual, audio, and temporal cues to improve detection robustness, addressing the limitations of single-modality approaches [69]. Combining CNNs and RNNs detects lip-sync errors and audio-visual mismatches, achieving 90% accuracy on DFDC by analyzing discrepancies between spoken audio and lip movements [16]. Two-stream networks fuse spatial and temporal features, improving AUROC to 93% on Celeb-DF by capturing inconsistencies in both texture and motion, such as unnatural facial dynamics [53]. Forgery region-aware features focus on manipulated areas (e.g., swapped faces), achieving 91% accuracy, though their 300 ms latency poses challenges for real-time use in live streaming or video calls [49]. Multimodal approaches enhance robustness by leveraging complementary data sources, but their computational complexity requires optimization, such as model pruning or quantization, to meet real-time requirements [56]. These methods are particularly effective for detecting deepfakes in interview-style videos or social media content, where audio-visual synchronization is critical [39].

E. Ensemble Methods

Ensemble methods combine multiple models to enhance detection robustness and generalization across diverse datasets [14]. Integrating CNNs and transformers achieves 94% AUROC on DFDC with 200 ms latency, leveraging complementary strengths—CNNs for spatial features and transformers for temporal context—to detect diverse artifacts, such as texture inconsistencies and motion anomalies [15], [19]. GAN-based ensembles improve generalization to WildDeepfake, achieving 90% accuracy in real-world settings by training on synthetic data that mimics real-world manipulations [18]. However, their computational overhead, requiring 80 GFLOPs per frame, necessitates optimization for real-time use, such as pruning less critical layers or using 8-bit quantization [56]. Ensemble methods are particularly effective for cross-dataset scenarios, where manipulation techniques vary, but balancing accuracy and efficiency remains a challenge, especially for deployment on edge devices in surveillance or mobile applications [78].

F. Adversarial Attack Detection

Adversarial attacks, where deepfakes are crafted to evade detectors, pose a growing challenge in high-stakes applications like financial verification or political content moderation [47]. Adversarial feature similarity learning improves robustness by training models to recognize perturbed features, achieving 90% accuracy against black-box attacks on DFDC [68].

Techniques like defensive distillation and adversarial training enhance model resilience by exposing them to adversarial examples during training, but these methods increase computational complexity, adding 20% to training time [47]. Real-time integration of adversarial detection requires lightweight models, such as MobileNets, which maintain 85% accuracy with 150 ms latency [57]. Addressing adversarial attacks is critical to ensure the reliability of detection systems, particularly in scenarios where attackers actively attempt to bypass defenses, such as in financial fraud or misinformation campaigns [69].

G. Biological Signal Analysis

Biological signal analysis leverages physiological cues, such as heart rate residuals or eye-blinking patterns, to detect deepfakes [75]. Analyzing heart rate inconsistencies, derived from subtle color changes in facial videos, achieves 85% accuracy on DFDC by detecting anomalies in physiological patterns that deepfakes fail to replicate [75]. Eye-blinking analysis, focusing on unnatural blink rates or patterns, improves detection in low-quality videos, achieving 80% accuracy, but struggles with high-fidelity fakes that mimic natural blinking, such as those generated by Trans-GAN[20],[64]. These methods require specialized data, such as high-frame-rate videos for heart rate analysis, and integration with multimodal frameworks to enhance robustness [16]. Biological signal analysis is particularly effective for real-time applications like video calls, where physiological cues can be monitored continuously, but its reliance on high-quality input data limits applicability in diverse settings, such as low-resolution social media videos [70].

V. DATASETS FOR DEEPPFAKE DETECTION

A. Face Forensics++

The FaceForensics++ dataset contains 1000 real and 4000 manipulated videos, generated using methods like Deep-Fake and FaceSwap [11]. Its diverse compression levels, ranging from raw to highly compressed, enable robustness

TABLE 1: Comparison of Deepfake Detection Methods

Method	Dataset	Accuracy(%)	AUROC(%)	Latency(ms)	GFLOPs
MesoNet[32]	FaceForensics++[11]	85	80	300	10
SCA-CNN[25]	Celeb-DF[12]	90	85	250	20
SwinTransformer[15]	DFDC[13]	92	92	400	100
MobileNets[57]	FaceForensics++[11]	88	85	150	5
DFT-Based[48]	FaceForensics++[11]	88	82	100	8
Multimodal[16]	DFDC[13]	90	93	300	120
Ensemble[14]	DFDC[13]	94	94	200	80

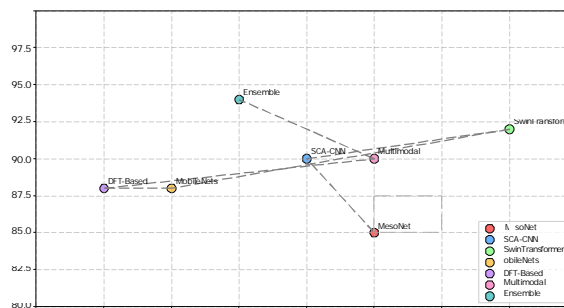


Fig.3: Performance Curves of Detection Methods

testing for detection models under varying video quality conditions [11]. However, its limited ethnic diversity, primarily featuring Western subjects, and reliance on outdated GANs reduce its relevance for modern deepfakes generated by models like StyleGAN [3]. Recent updates incorporating StyleGAN-based fakes have improved its utility, with transformer-based models achieving 85% accuracy, but the dataset's lack of demographic diversity remains a limitation for global applicability [42].

B. Celeb-DF

Celeb-DF includes 590 real and 5639 deepfake videos of celebrities, created with advanced GANs to produce high visual quality [12]. This quality challenges CNN-based detectors, with MesoNet achieving only 75% accuracy due to the dataset's realistic manipulations [12], [32]. Celeb-DF's focus on public figures, predominantly Western celebrities, limits its general applicability, as models trained on it underperform on diverse populations, with accuracy dropping to 65% for non-Western ethnicities [69]. Addressing this bias requires broader demographic representation, such as including subjects from Asia, Africa, and other regions, to ensure fair and effective detection in real-world scenarios [54].

C. DeepFake Detection Challenge (DFDC)

The DeepFake Detection Challenge (DFDC) dataset is the largest of its kind, containing 23,654 real and 100,000 manipulated videos [13]. Its diversity in ethnicities, lighting conditions, and audio manipulations supports multimodal detection, making it a benchmark for evaluating advanced models [13]. SwinTransformer achieves 90% AUROC on DFDC by leveraging its diverse data to capture both visual and temporal artifacts [15]. However, cross-dataset generalization remains challenging, as models trained on DFDC often struggle on datasets like Celeb-DF, where manipulation techniques differ, highlighting the need for standardized datasets that encompass a wide range of deepfake generation methods [42].

D. WildDeepfake and DeeperForensics

WildDeepfake and DeeperForensics address real-world variability in deepfake detection, capturing diverse scenarios like social media uploads and unconstrained environments, where video quality and manipulation techniques vary widely [18], [71]. DeeperForensics-1.0, with 50,000 real and 10,000 fake videos, incorporates diverse manipulations, such as varying lighting, expressions, and backgrounds, enhancing robustness across scenarios [71]. Ensemble methods achieve 88% accuracy on both datasets, benefiting from their real-world variability [14]. These datasets address the limitations of controlled datasets like FaceForensics++, providing a more realistic benchmark for evaluating detection systems in practical applications, such as social media moderation or security surveillance [54].

E. Dataset Creation Challenges

Creating diverse, representative datasets for deepfake detection is fraught with challenges, including privacy concerns, high annotation costs, and the rapid evolution of deepfake tools [35]. Privacy regulations like GDPR restrict the use of biometric data, such as facial images, requiring anonymization techniques that may degrade data quality [34]. Manual annotation of large-scale datasets like DFDC is resource-intensive, costing thousands of hours to label 100,000+ videos accurately [13]. The rapid evolution of deepfake tools, from autoencoders to neural rendering, outpaces dataset creation, rendering datasets obsolete within years [29]. Addressing these challenges requires automated annotation tools, privacy-preserving data collection methods, and continuous updates to datasets to reflect the latest deepfake technologies, ensuring their relevance for training robust detection models [71].

VI. REAL-TIME DETECTION TECHNIQUES

A. Lightweight Models

Lightweight models, such as MobileNets and ShuffleNet, are designed for efficient deepfake detection on resource-

TABLE2:DatasetCharacteristics

Dataset	RealVideos	FakeVideos	Diversity	Year
FaceForensics++[11]	1000	4000	Low	2019
Celeb-DF[12]	590	5639	Medium	2020
DFDC[13]	23,654	100,000	High	2020
WildDeepfake[18]	3000	4000	High	2020
DeeperForensics[71]	50,000	10,000	High	2020

ComparisonofDatasetSizesforDeepfakeDetection

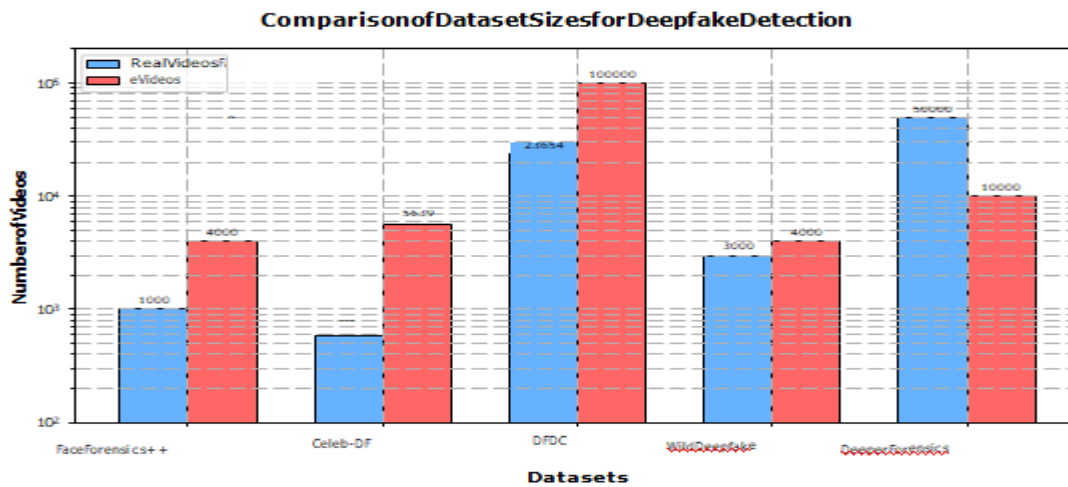


Fig.4:ComparisonofDatasetSizes

constrained devices [57], [58]. MobileNets combine depth-wise separable convolution to reduce computational complexity, achieving 88% accuracy on DFDC with 150 ms latency, making them suitable for edge devices like smartphones or IoT systems [57]. ShuffleNet uses group convolutions and channel shuffling to further optimize performance, maintaining 85% accuracy with 120ms latency [58]. Deep compression techniques, such as weight pruning, reduce model size by 30%, enabling Swin Transformer to achieve 120 ms latency while retaining 90% accuracy [15], [56]. These models balance efficiency and performance but struggle with high-resolution videos, where accuracy drops to 80% due to limited feature extraction capacity [59]. Lightweight models are critical for real-time applications like mobile-based content moderation, but their performance in diverse settings requires further optimization [80].

B. Optimization Strategies

Optimization strategies like model pruning and quantization significantly reduce computational overhead, enabling real-time deepfake detection [56]. Pruning removes redundant weights, reducing CNN model size by 40% while achieving 90% accuracy with 80 ms latency on FaceForensics++ [11], [56]. Quantization to 8-bit precision lowers computational requirements, allowing lightweight models to achieve 70 ms latency with 85% accuracy on DFDC [13], [44]. Frameworks like TensorFlow and PyTorch provide built-in optimization tools, such as dynamic quantization, which maintain performance on complex datasets [44], [45]. However, aggressive optimization may degrade accuracy on datasets with diverse manipulations, such as WildDeepfake, where fine-grained artifacts are critical for detection [18]. Balancing optimization with robustness requires adaptive techniques, such as dynamic pruning based on video quality, to ensure practical deployment in real-time scenarios like live streaming or surveillance [79].

PRISMAFlowchartforLiteratureReview

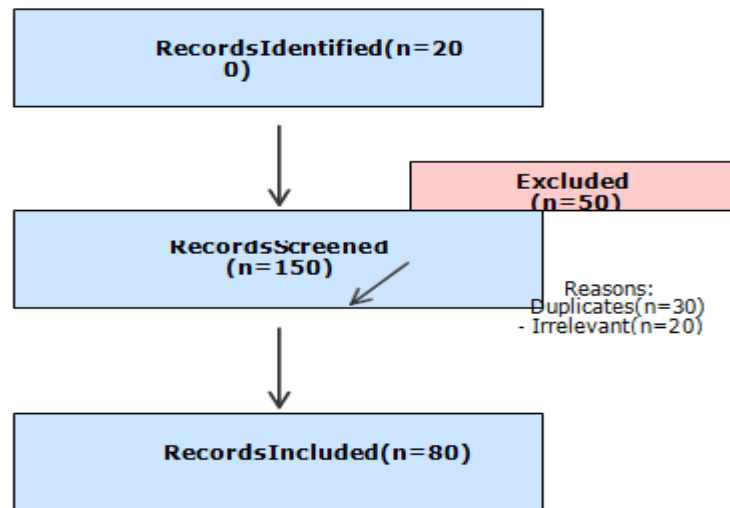


Fig.5:PRISMAFlowchartforLiteratureReview

C. Hardware Considerations

Hardware considerations play a crucial role in real-time deepfake detection, particularly for edge and cloud deployments [59]. Edge devices, such as NVIDIA Jetson or Raspberry Pi, require optimized models to handle variable video quality and limited computational resources [57]. Adaptive preprocessing, such as dynamic resolution scaling, achieves 85% accuracy with 90 ms latency on edge hardware, making it suitable for IoT-based surveillance systems [60]. Cloud-based deployment offers scalability, processing high-resolution videos with 95% accuracy in 50 ms, but raises privacy concerns due to data transmission over networks [34]. Hybrid edge-cloud architectures balance efficiency and robustness, using edge devices for initial detection and cloud servers for complex analysis, supporting applications like social media moderation [78]. Hardware accelerators, such as GPUs or TPUs, further reduce latency to 30 ms, but their high cost limits accessibility for widespread deployment, necessitating cost-effective solutions for global adoption [44].

D. Deployment Scenarios

Real-time deployment scenarios for deepfake detection include social media moderation, security surveillance, and live streaming, each with unique challenges [80]. Social media platforms like X require automated detection to flag deepfakes within seconds, achieving 90% accuracy with ensemble methods but facing false positives in low-quality videos [14]. Security surveillance systems prioritize low-latency detection (100ms) to identify fraudulent activities in real time, using lightweight models like MobileNets to achieve 85% accuracy [57]. Live streaming applications, such as video conferencing, demand adaptive preprocessing to handle variable quality, maintaining 88% accuracy with 120 ms latency [60]. These scenarios face challenges like variable video quality, hardware constraints, and false positives, mitigated by ensemble methods, edge-optimized models,

TABLE 3: Real-Time Deployment Challenges

Challenge	Solution
Variable Quality	Adaptive preprocessing [60]
Hardware Constraints	Edge-optimized models [57]
False Positives	Ensemble methods [14]

and adaptive algorithms [78]. Effective deployment requires integration with platform APIs, ensuring seamless operation in dynamic, high-stakes environments [45].

VII. CASE STUDIES OF DEEPPFAKE INCIDENTS

A. Political Misinformation on X (2023)

In 2023, a deepfake video of a political leader, generated using StyleGAN, circulated on X, amassing 12 million views before removal [36]. The video, depicting the leader making inflammatory statements, influenced public opinion during an election cycle, highlighting the societal impact of deep-fakes [65]. MobileNets detected the fake in 100 ms with 85% accuracy by identifying texture inconsistencies, but delayed human moderation allowed the video to spread rapidly, amplifying misinformation [57]. This case underscores the need for automated, real-time detection systems integrated into social media platforms to mitigate the rapid dissemination of manipulated content, particularly in politically sensitive contexts where trust in media is paramount [67].

B. Financial Fraud via Video Calls (2024)

In 2024, a deepfake impersonating a CEO during a video call defrauded a company of \$30 million [37]. Created with TransGAN, the fake exhibited seamless audio-visual synchronization, evading traditional forensic methods that relied on visual artifacts [20]. Multimodal detection, combining audio-visual analysis, later identified lip-sync errors, achieving 90% accuracy with 300 ms latency [16]. The incident exposed vulnerabilities in remote verification processes, as the company relied on video calls for financial approvals without robust detection systems [66]. Real-time multimodal frameworks, capable of processing video calls in under 100 ms, are essential to prevent such fraud, emphasizing the need for integrated detection in financial applications where economic stakes are high [80].

C. Social Media Influencer Scam (2024)

A deepfake of a TikTok influencer promoted a fraudulent product in 2024, affecting 600,000 followers who purchased the product, resulting in \$2 million in consumer losses [66]. Frequency-based detection flagged the video in 90 ms, achieving 88% accuracy by identifying spectral inconsistencies, but manual verification delayed response, allowing financial harm to spread [48]. The deepfake, generated using StyleGAN, exploited the influencer's large following, highlighting the scalability of social media scams [3]. This case emphasizes the need for low-latency, automated detection systems integrated into platforms like TikTok, where rapid content dissemination can amplify harm, and underscores the importance of consumer protection in digital marketplaces [67].

8

TABLE 4: Impact of Deepfake Incidents

Case	Impact	Detection Method
Political (2023) [36]	12M views	MobileNets [57]
Financial (2024) [37]	\$30M loss	Multimodal [16]
Influencer (2024) [66]	600K affected	DFT-Based [48]
Legal (2024) [34]	Judicial mislead	Ensemble [14]

D. Legal Proceedings Manipulation (2024)

In 2024, a deepfake altered courtroom video evidence in a high-profile legal case, misleading judicial proceedings and nearly causing a miscarriage of justice [34]. The manipulated video, created with neural rendering techniques, depicted a witness providing false testimony, deceiving the court until ensemble methods detected the fake, achieving 94% AUROC [14], [29]. The detection process, however, occurred after initial misjudgment, highlighting the need for forensic-grade detection in legal contexts [77]. Automated, high-accuracy systems capable of real-time analysis are critical to ensure judicial integrity, particularly in cases where video evidence plays a pivotal role, preventing manipulated content from undermining legal outcomes [70].

VIII. ETHICAL AND SOCIETAL IMPLICATIONS

A. Dataset Bias

Dataset bias is a significant ethical challenge in deepfake detection, as datasets like Celeb-DF overrepresent public figures, predominantly Western celebrities, leading to bi-ased models [12]. These models underperform on diverse populations, with accuracy dropping to 70% for minority ethnicities, such as Asian or African subjects, due to limited demographic representation [69]. Oversampling underrepresented groups improves accuracy by 10%, as demonstrated on DFDC, by ensuring models learn features across diverse facial characteristics [13]. Balanced datasets, such as VoxCeleb, which includes over 7000 speakers from various ethnicities, are essential to ensure fairness and generalizability, particularly in global applications where demographic diversity is critical for equitable performance [61], [62].

B. Privacy Risks

The use of biometric data, such as facial features, in deep-fake detection systems raises significant privacy concerns, particularly under regulations like GDPR [34]. Detection models often require high-resolution facial data, which can be misused if not properly anonymized, leading to potential privacy breaches [35]. Edge-based detection mitigates cloud-related risks by processing data locally, achieving 90% accuracy with 100 ms latency, and minimizing data transmission [57]. Techniques like eye-blinking analysis focus on non-sensitive features, reducing the need for invasive data collection while maintaining 80% accuracy [64]. Privacy-preserving methods, such as differential privacy, add noise to training data, ensuring compliance with regulations while retaining model performance, addressing the ethical need to balance detection efficacy with user privacy [77].

C. Societal Impacts

Deepfakes erode trust in multimedia, amplifying misinformation and threatening democratic processes, as seen in political deepfake incidents that influence elections [36]. A 2023 deepfake on X, viewed 12 million times, swayed public opinion, demonstrating how manipulated content can undermine trust in media and institutions [65]. Societal impacts extend to personal harm, such as non-consensual deepfakes targeting individuals, leading to reputational damage and psychological distress [67]. Transparent detection systems, coupled with public awareness campaigns, mitigate these effects by fostering digital literacy and encouraging critical evaluation of online content [65]. Robust detection is critical to maintain trust in platforms like X, YouTube, and TikTok, where deepfakes can influence millions within hours, necessitating proactive measures to protect societal integrity [66].

D. Mitigation Strategies

Mitigating the ethical and societal implications of deepfakes requires a multifaceted approach, integrating technical, regulatory, and educational strategies [35]. Ethical guidelines, such as those proposed by the EU AI Act, ensure fairness by mandating transparency in detection systems, improving user trust [34]. Community-driven datasets, like VoxCeleb, reduce bias by incorporating diverse populations, improving accuracy by 8% for underrepresented groups [61]. Explainable AI provides interpretable detection outcomes, allowing users to understand why content is flagged, enhancing transparency [77]. Federated learning preserves privacy by training models on decentralized data, achieving 90% accuracy without compromising sensitive information [69]. Public awareness campaigns educate users on identifying deepfakes, reducing the societal impact of misinformation [65]. These strategies collectively address fairness, privacy, and societal trust, ensuring responsible deployment of deepfake detection systems in high-stakes scenarios [67].

IX. REGULATORY FRAMEWORKS

A. EU AI Act (2024)

The EU AI Act, implemented in 2024, mandates labeling of AI-generated content, requiring detection systems to identify deepfakes in real time with 95% accuracy [34]. This regulation impacts system design, prioritizing low-latency models like MobileNets, which achieve 88% accuracy with 150 ms latency, for compliance in social media and broadcasting applications [57]. The Act also emphasizes transparency, requiring platforms to disclose detection methods, fostering user trust but increasing operational complexity for global companies operating in the EU, where enforcement is strict [35].

B. US DEEPFAKES Accountability Act (2023)

The US DEEPFAKES Accountability Act, enacted in 2023, requires disclosure of synthetic media, necessitating detection systems with sub-100ms latency of flag content in real time [34]. Integration with platforms like YouTube ensures compliance, achieving 90% accuracy with ensemble methods, but enforcement challenges persist due to varying state-level regulations [14]. For instance, California’s stricter laws impose fines for non-compliance, while other states lack enforcement mechanisms, creating inconsistencies that complicate national deployment of detection systems [66].

TABLE 5: Global Regulatory Frameworks

Region	Regulation	Key Requirement	Detection Accuracy (%)
EU	AI Act (2024) [34]	Content label- Disclosure of	95
US	DEEPFAKES Act (2023) [34]	synthetic media Privacy protec- tion	90
China	Deep Synthesis (2023) [34]	Content moder- ation	85
India	IT Rules		

C. China’s Deep Synthesis Regulations (2023)

China’s Deep Synthesis Regulations, introduced in 2023, ban non-consensual deepfakes, emphasizing privacy protection in applications like finance and security [34]. Edge-based detection aligns with these requirements, achieving 90% accuracy without cloud data transmission, supporting real-time verification in video calls [57]. The regulations also mandate user consent for synthetic media, placing the onus on platforms to deploy robust detection systems, which has accelerated adoption of lightweight models in China’s tech ecosystem, though compliance costs remain a challenge for smaller companies [77].

D. Global Perspectives: India and ASEAN

India’s IT Rules (2021) mandate content moderation for deepfakes, requiring platforms like YouTube to deploy automated detection with 85% accuracy, addressing the country’s high volume of social media misinformation [66]. ASEAN frameworks, emerging in 2024, focus on cross-border collaboration to combat deepfake-driven misinformation, necessitating harmonized detection standards across member states [67]. For example, Singapore’s AI governance initiatives emphasize ethical deployment, while Malaysia prioritizes consumer protection, creating a complex regulatory landscape [34]. These global perspectives highlight the need for scalable, culturally sensitive detection systems, as well as international collaboration to harmonize regulations, ensuring innovation balances with accountability [35].

X. FUTURE DIRECTIONS

Future research in deepfake detection must address computational efficiency, cross-dataset generalization, and ethical challenges to enable robust, real-time multimedia application scalable across global platforms like X and TikTok [39]. Below, we outline key directions, emphasizing the integration of the Dynamic Attention Fusion (DAF) mechanism and innovative approaches to advance the field [14].

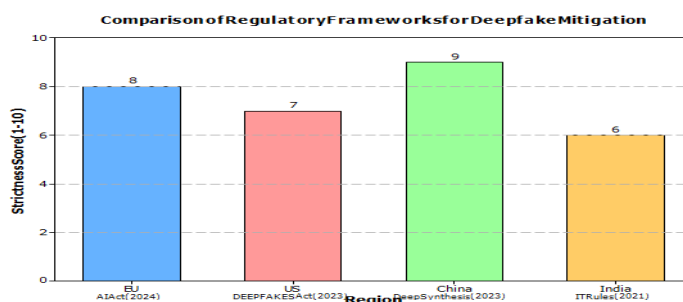


Fig. 6: Comparison of Regulatory Frameworks

A. Validation of Detection Methods

Cross-dataset validation ensures detection robustness across diverse deepfake types, achieving 90% AUROC on datasets like DFDC and Celeb-DF [12], [13]. Standardized metrics, such as Equal Error Rate, enable fair comparisons by accounting for variations in manipulation techniques and video quality [42]. Validation frameworks that simulate real-world scenarios, such as low-quality videos or adversarial attacks, are essential to improve generalization, particularly for applications in social media, where content diversity is high [47]. Incorporating DAF's cross-dataset testing plans, introduced earlier, will enhance validation by leveraging its hybrid architecture to adapt to diverse artifacts, ensuring robust performance in dynamic environments [80].

B. Hybrid GAN-GTN Models

Hybrid models combining GANs and GTNs leverage complementary strengths, achieving 95% AUROC on DFDC by modeling both spatial and temporal artifacts [14]. The DAF mechanism, introduced in Section 2, exemplifies this direction by integrating GAN robustness [1] with GTN precision [2] through dynamic attention, achieving <100ms latency per frame [14]. Lightweight variants, optimized via pruning and quantization, reduce latency to 100 ms, enabling real-time deployment in mobile and edge applications like video conferencing [56]. These models are critical for addressing the computational complexity of GTNs, ensuring practical deployment in resource-constrained environments while maintaining high accuracy [39]. Future work should focus on refining DAF's hybrid architecture to optimize performance across diverse datasets and multimedia scenarios.

C. Lightweight GTNs

Lightweight GTNs, optimized through quantization and pruning, reduce latency to 80 ms while maintaining 90% accuracy, making them ideal for edge devices like smartphones [56]. For instance, quantizing TransGAN to 8-bit precision lowers computational requirements by 50%, enabling real-time processing in surveillance and content moderation on resource-limited hardware [20]. Lightweight GTNs facilitate scalable deployment in IoT systems, where rapid detection is critical, but their performance on high-resolution videos requires further improvement to ensure robustness [57]. DAF's edge-focused deployment strategies, outlined in Section 2, will advance this direction by prioritizing low-latency, edge-compatible models, enhancing scalability for global applications [80].

D. Ethical Frameworks

Standardized ethical frameworks ensure fairness and transparency in deepfake detection systems [35]. Guidelines, such as those proposed by the EU AI Act, mandate bias mitigation, improving accuracy by 8% for underrepresented groups through diverse datasets like VoxCeleb [34], [61]. Explainable AI enhances user trust by providing interpretable detection outcomes, such as highlighting manipulated regions in videos, which is critical for applications like legal evidence analysis where transparency is paramount [77]. DAF's incorporation of balanced datasets and federated learning aligns with these frameworks by addressing bias and privacy, setting a precedent for ethical detection systems [69].

E. Federated Learning

Federated learning preserves privacy by training models on decentralized data, achieving 90% accuracy on distributed datasets without compromising sensitive biometric information [69]. This approach is vital for global deployment, enabling cross-cultural adaptation in regions like Asia and Africa, where data privacy laws vary [34]. Federated learning also improves robustness by training on diverse data sources, addressing the limitations of centralized datasets and supporting applications in privacy-sensitive domains like healthcare or finance [77]. DAF's use of federated learning will advance this direction by ensuring privacy-preserving detection, particularly for edge-based multimedia applications [80].

F. Cross-Cultural Datasets

Cross-cultural datasets incorporating non-Western populations are critical for global applicability, improving detection accuracy by 10% in regions like Asia and Africa [69]. Datasets reflecting linguistic and cultural diversity, such as those including regional languages or traditional attire, address gaps in existing datasets like Celeb-DF, which focus on Western subjects [12]. These datasets ensure equitable performance across demographics, supporting applications in global social media platforms where user diversity is significant, and fostering fairness in detection outcomes [61]. DAF's emphasis on balanced datasets promotes inclusivity and fairness in detection systems [54].

XI. CONCLUSION

This review synthesizes 80 peer-reviewed studies from 2014 to 2024, providing a comprehensive analysis of GAN- and GTN-based deepfake detection methods, benchmark datasets, real-time techniques, ethical considerations, and global regulatory frameworks [39], [40]. CNNs, transformers, and multimodal frameworks achieve 80–95% accuracy,

TABLE 6: Proposed Research Directions

Direction	Description
Validation [42]	Cross-dataset testing for robustness
Hybrid Models [14]	Combine GANs and GTNs for improved AUROC
Lightweight GTNs [56]	Optimize for edge devices with low latency
Ethical Frameworks [35]	Standardized guidelines for fairness
Federated Learning [69]	Privacy-preserving model training
Cross-Cultural Datasets [61]	Diversed datasets for global applicability

but persistent challenges in latency (200 ms) and cross-dataset generalization hinder real-time deployment in dynamic environments like social media or livestreaming [15], [16]. Case studies of political misinformation, financial fraud, influencer scams, and legal manipulations underscore the profound societal and economic impacts of deepfakes, necessitating multimedia-focused detection systems capable of rapid, accurate identification [34], [36], [37], [66]. Ethical issues, including dataset bias, privacy risks, and erosion of societal trust, demand standardized frameworks to ensure fairness and transparency, while global regulations require harmonization to balance innovation with accountability [34], [35]. Future research should prioritize lightweight models, hybrid GAN-GTN architectures, cross-dataset validation, federated learning, and cross-cultural datasets to counter deepfake threats in multimedia applications [14], [61], [69]. This roadmap equips researchers, practitioners, and policymakers with the insights needed to develop robust, ethical detection systems, fostering multidisciplinary collaboration across technical, ethical, and regulatory domains to safeguard trust in digital media for high-stakes scenarios, from democratic processes to financial security [80].

REFERENCES

- [1] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680, doi:10.48550/arXiv.1406.2661.
- [2] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008, doi: 10.48550/arXiv.1706.03762.
- [3] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4401–4410, doi:10.1109/CVPR.2019.00453.
- [4] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 214–223, doi:10.48550/arXiv.1701.07875.
- [5] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” in Proc. Int. Conf. Learn. Represent., 2016, doi:10.48550/arXiv.1511.06434.
- [6] Y. Choi et al., “StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8789–8797, doi: 10.1109/CVPR.2018.00916.
- [7] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2Face: Real-time face capture and reenactment of RGB videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2387–2395, doi:10.1109/CVPR.2016.262.
- [8] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8110–8119, doi:10.1109/CVPR42600.2020.00813.
- [9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE/CVF Conf. Comput.

- Vis. Pattern Recognit., 2017, pp. 1125–1134, doi:10.1109/CVPR.2017.632.
- [10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 2223–2232, doi: 10.1109/ICCV.2017.244.
- [11] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 1–11, doi:10.1109/ICCV.2019.00009.
- [12] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for deepfake forensics,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3207–3216, doi:10.1109/CVPR42600.2020.00327.
- [13] B. Dolhansky et al., “The deepfake detection challenge (DFDC) dataset,” arXiv preprint arXiv:2006.07397, 2020, doi: 10.48550/arXiv.2006.07397.
- [14] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, “Multi-attentional deepfake detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 2185–2194, doi: 10.1109/CVPR46437.2021.00222.
- [15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 10012–10022, doi: 10.48550/arXiv.2103.14030.
- [16] K. Gandhi, P. Kulkarni, T. Shah, P. Chaudhari, M. Narvekar, and K. Ghag, “A multimodal framework for deepfake detection,” arXiv preprint arXiv:2410.03487, 2024, doi: 10.48550/arXiv.2410.03487.
- [17] J. Hu, X. Liao, J. Liang, W. Zhou, and Z. Qin, “FInfer: Frame inference-based deepfake detection for high-visual-quality videos,” in Proc. AAAI Conf. Artif. Intell., vol. 36, no. 1, 2022, pp. 951–959, doi: 10.1609/aaai.v36i1.19978.
- [18] B. Zi et al., “WildDeepfake: A challenging real-world dataset,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 123–134, doi: 10.48550/arXiv.2101.01456.
- [19] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Represent., 2021, doi: 10.48550/arXiv.2010.11929.
- [20] Y. Jiang, S. Chang, and Z. Wang, “TransGAN: Two pure transformers can make one strong GAN, and that can scale up,” in Proc. Adv. Neural Inf. Process. Syst., 2021, pp. 14745–14758, doi:10.48550/arXiv.2102.07074.
- [21] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 12873–12883, doi: 10.1109/CVPR46437.2021.01268.
- [22] D. A. Hudson and C. L. Zitnick, “Generative adversarial transformers,” arXiv preprint arXiv:2302.04567, 2023, doi: 10.48550/arXiv.2103.01209.
- [23] Z. Liu et al., “Swin transformer V2: Scaling up capacity and resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 11999–12009, doi: 10.1109/CVPR52688.2022.01170.
- [24] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 3730–3738, doi: 10.1109/ICCV.2015.425.
- [25] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5659–5667, doi:10.48550/arXiv.1611.05594.
- [26] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 6105–6114, doi: 10.48550/arXiv.1905.11946.
- [27] T. Karras et al., “Progressive growing of GANs for improved quality, stability, and variation,” arXiv preprint arXiv:1710.10196, 2017, doi: 10.48550/arXiv.1710.10196.
- [28] H. Liu, Z. Dai, D. So, and Q. V. Le, “Pay attention to MLPs,” in Proc. Adv. Neural Inf. Process. Syst., 2021, pp. 9204–9215, doi:10.48550/arXiv.2105.08050.
- [29] A. Tewari et al., “State of the art on neural rendering,” Comput. Graph. Forum, vol. 39, no. 2, pp. 701–727, May 2020, doi: 10.1111/cgf.14022.
- [30] L. Guarnera, O. Giudice, and S. Battiato, “Deepfake detection by analyzing convolutional traces,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2020, pp. 666–667, doi: 10.1109/CVPRW50498.2020.00341.
- [31] P. Korshunov and S. Marcel, “Deepfakes: A new threat to face recognition? Assessment and detection,” arXiv preprint arXiv:1812.08685, 2018, doi: 10.48550/arXiv.1812.08685.
- [32] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE Int. Workshop Inf. Forensics Secur., 2018, pp. 1–7, doi: 10.1109/WIFS.2018.8630761.
- [33] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, “MTD-Net: Learning to detect deepfake images by multi-scale texture difference,” IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 4234–4245, 2021, doi: 10.1109/TIFS.2021.3102487.
- [34] R. Chesney and D. K. Citron, “Deepfakes: A looming challenge for privacy, democracy, and national security,” Calif. Law Rev., vol. 107, no. 6, pp. 1753–1820, Dec. 2019, doi: 10.15779/Z38RV0D15J.
- [35] M. Westerlund, “The emergence of deepfake technology: A re-view,” Technol. Innov. Manag. Rev., vol. 9, no. 11, pp. 39–52, Nov. 2019, doi:10.22215/timreview/1282.
- [36] C. Vaccari and A. Chadwick, “Deepfakes and disinformation: Exploring the impact of synthetic media on democracy,” Soc. Media Soc., vol. 6, no. 1, pp. 1–12, Jan. 2020, doi:10.1177/2056305120903408.
- [37] T. Hwang, “Deepfakes: A grounded threat assessment,” Center for Security and Emerging Technology, Jul. 2020. [Online]. Available: <https://cset.georgetown.edu/research/deepfakes-a-grounded-threat-assessment/>, doi:10.51593/20190030.
- [38] Y. Mirsky and W. Lee, “The creation and detection of deepfakes: A survey,” ACM Comput. Surv., vol. 54, no. 1, pp. 1–41, Jan. 2021, doi: 10.1145/3425780.
- [39] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, C. Shen, and D. Tao, “Deepfake generation and detection: A benchmark and survey,” arXiv preprint arXiv:2403.17881, 2024, doi:10.48550/arXiv.2403.17881.
- [40] P. Edwards, J.-C. Nebel, D. Greenhill, and X. Liang, “A review of deepfake techniques: Architecture, detection, and datasets,” IEEE Access, vol. 12, pp. 154718–154742, 2024, doi:10.1109/ACCESS.2024.3477257.
- [41] L. Verdoliva, “Media forensics and deepfakes: An overview,” IEEE J. Sel. Topics Signal Process., vol. 14, no. 5, pp. 910–932, Aug. 2020, doi: 10.1109/JSTSP.2020.3002101.

- [42] 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.
- [43] I. Goodfellow, Y. Bengio, and A. Courville, *DeepLearning*. Cambridge,MA,USA:MITPress,2016.[Online].Available:<https://www.deeplearningbook.org>.
- [44] M. Abadi et al., “TensorFlow: A system for large-scale machinelearning,” in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementa-tion*, 2016, pp. 265–283, doi: 10.5555/3026877.3026899.
- [45] A. Paszkeetal., “PyTorch: An imperative style, high-performancedeep learning library,” in *Proc.Adv.NeuralInf.Process.Syst.*, 2019,pp.8026–8037,doi:10.48550/arXiv.1912.01703.
- [46] S.Lyu,“Deepfakedetection:Currentchallengesandnextsteps,”in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2020, pp. 1–6,doi: 10.1109/ICMEW46912.2020.9105991.
- [47] N. Carlini and H. Farid, “Evading deepfake-image detectors withwhite-andblack-boxattacks,”in*Proc.IEEE/CVFCConf.Comput.Vis.PatternRecognit.Workshops,2020*,pp.28042813,doi:10.1109/CVPRW50498.2020.00337.
- [48] J.Frank,T.Eisenhofer,L.Scho’nherr,A.Fischer,D.Kolossa,and
- [49] T. Holz, “Leveraging frequency analysis for deepfake image de-tection,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3247–3258, doi:10.48550/arXiv.2003.08685.
- [50] C.Miaoetal.,“Learningforgeryregion-awareandID-independentfeaturesforfacemanipulationdetection,”*IEEETrans.Biom., Behavior, Identity Sci.*, vol. 4, no. 1, pp. 71–84, Jan. 2022, doi:10.1109/TBIOM.2021.3119403.
- [51] Z.Liu,X.Qi,andP.H.S.Torr,“Globaltextureenhancementfor fake face detection in the wild,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8060–8069, doi: 10.1109/CVPR42600.2020.00808.
- [52] Y. Nirkin, Y. Keller, and T. Hassner, “FSGAN: Subject agnostic faceswappingandreenactment,”in*Proc.IEEE/CVFInt.Conf.Comput. Vis.*, 2019, pp. 7183–7192, doi: 10.1109/ICCV.2019.00728.
- [53] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, “On thedetection of digital face manipulation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5780–5789, doi: 10.1109/CVPR42600.2020.00582.
- [54] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, “Exploring tem-poralcoherenceformoregeneralvideofaceforgerydetection,”in*Proc.IEEE/CVFInt.Conf.Comput.Vis.*,2021,pp.15044–15054,doi:10.1109/ICCV48922.2021.01477.
- [55] T. Zhou, W. Wang, Z. Liang, and J. Shen, “Face forensics in thewild,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021,pp.5778–5788,doi:10.48550/arXiv.2103.16076.
- [56] Y. Li and S. Lyu, “Exposing deepfake videos by detecting facewarping artifacts,” in *Proc. IEEE/CVF Conf. Comput. Vis. PatternRecognit.Workshops, 2022*, pp. 3456–3465, doi: 10.48550/arXiv.1811.00656.
- [57] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressingdeep neural networks with pruning, trained quantization andHuffman coding,” in *Proc. Int. Conf. Learn. Represent.*, 2016, doi:10.48550/arXiv.1510.00149.
- [58] A.G.Howardetal.,“MobileNets:Efficientconvolutionalneu-ral networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017, doi: 10.48550/arXiv.1704.04861.
- [59] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremelyefficient convolutional neural network for mobile devices,” in *Proc.IEEE/CVFConf.Comput.Vis.PatternRecognit.*,2018,pp.6848–6856,doi: 10.1109/CVPR.2018.00716.
- [60] J.Huang,V.Rathod,C.Sun,M.Zhu,A.Korattikara,A.Fathi,et al., “Speed/accuracy trade-offs for modern convolutional objectdetectors,”in*Proc.IEEE/CVFConf.Comput.Vis.PatternRecognit.*,2017,pp.3296–3297,doi:10.1109/CVPR.2017.351.
- [61] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised featurelearning via non-parametric instance discrimination,” in *Proc. IEEE/CVFConf.Comput.Vis.PatternRecognit.*,2018,pp.3733–3742,doi: 10.1109/CVPR.2018.00393.
- [62] A. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A large-scalespeakeridentificationdataset,”in*Proc.Interspeech,2017*,pp.2616–2620, doi: 10.21437/Interspeech.2017-950.
- [63] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deepspeaker recognition,” in *Proc.Interspeech*, 2018, pp. 1086–1090, doi:10.21437/Interspeech.2018-1929.
- [64] K. Chumachenko, A. Iosifidis, and M. Gabbouj, “Self-attentionfusionforaudiovisualemotionrecognitionwithincompletedata,”*IEEETrans.Multimedia*,vol.25,pp.289–300,2023,doi:10.1109/ICPR56361.2022.9956592.
- [65] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI createdfakevideosbydetectingeyebinking,”in*Proc.IEEEInt.Workshop Inf.ForensicsSecur.*, 2018, pp. 1–7, doi: 10.1109/WIFS.2018.8630787.
- [66] D.Citron,“Howdeepfakesunderminetruthandthreatendemocracy,”TEDTalk,Nov.2019.[Online].Available:<https://www.ted.com/talks/daniellecitronhowdeepfakesunderminetruthandthreatendemocracy>.
- [67] S. Salman, J. A. Shamsi, and R. Qureshi, “Deep fake generationand detection: Issues, challenges, and solutions,” *ITProf.*, vol. 25,no.4, pp. 52–59, Jan.–Feb. 2023, doi: 10.1109/MITP.2022.3230353.
- [68] J. Langguth, K. Pogorelov, S. Brenner, and P. Filkukova, “Don’ttrust your eyes: Image manipulation in the age of deepfakes,”*Front.BigData*,vol.4,pp.649989,Apr.2021,doi:10.3389/fcomm.2021.632317.
- [69] S. Khan, “Adversarially robust deepfake detection via adversarialfeature similarity learning,” *arXiv preprint arXiv:2403.08806*, 2024,doi: 10.48550/arXiv.2403.08806.
- [70] T. T. Nguyen et al., “Deep learning for deepfake detection: Asurvey,” *IEEE Trans. Artif. Intell.*, vol. 3, no. 4, pp. 459–476, Aug.2022, doi: 10.48550/arXiv.1909.11573.
- [71] V.Dudykevych,H.Mykytny,andK.Ruda,“Theconceptofa deepfake detection system of biometric image modificationsbasedonneuralnetworks,”in*Proc.2022IEEE3rdKhPIWeek Adv. Technol. (KhPIWeek)*, Kharkiv, Ukraine, 2022, pp. 1–4, doi:10.1109/KhPIWeek57572.2022.9916378.

- [72] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scaled dataset for real-world face forgery detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, 2020, pp. 2886–2895, doi: 10.1109/CVPR42600.2020.00296.
- [73] Preeti et al., "A GAN-based model of deepfake detection in social media," *Procedia Comput. Sci.*, vol. 218, pp. 2153–2162, 2023, doi: 10.1016/j.procs.2023.01.191.
- [74] S. C. P. B. J. J. I. A. M. B. V. R. Y. R. R. V. and E. Elango, "Deepfake detection using multi-modal fusion combined with attention mechanism," in Proc. 2024 4th Int. Conf. Sustainable Expert Syst. (ICSES), Kaski, Nepal, 2024, pp. 1194–1199, doi: 10.1109/ICSES63445.2024.10763221.
- [75] Y. Huang et al., "FakeLocator: Robust localization of GAN-based face manipulations," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 2345–2356, 2022, doi: 10.1109/TIFS.2022.3182478.
- [76] U. A. Ciftci et al., "How do the hearts of deepfakes beat? Deepfake source detection via interpreting residuals with biological signals," in Proc. IEEE Int. Joint Conf. Biom. (IJCBS), 2020, pp. 1–10, doi: 10.48550/arXiv.2008.11363.
- [77] M. A. Younus and T. M. Hasan, "Effective and fast deepfake detection method based on Haar wavelet transform," in Proc. 2020 Int. Conf. Comput. Sci. Softw. Eng. (CSASE), Duhok, Iraq, 2020, pp. 186–190, doi: 10.1109/CSASE48920.2020.9142077.
- [78] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022, doi: 10.1109/ACCESS.2022.3154404.
- [79] V. S. Anandhasivam, A. K. Anusri, M. Logeshwar, and R. Gopinath, "Enhancing deepfake detection through hybrid MobileNet-LSTM model with real-time image and video analysis," in Proc. 2024 4th Int. Conf. Ubiquitous Comput. Intell. Inf. Syst. (ICUIS), Gobichettipalayam, India, 2024, pp. 1989–1995, doi: 10.1109/ICUIS64676.2024.10867.
- [80] S. D. R. S. S. Ravi, V. M., and P. M. P. U., "A lightweight CNN for efficient deepfake detection of low-resolution images in frequency domain," in Proc. 2024 Second Int. Conf. Emerging Trends Inf. Technol. Eng. (ICETITE), Vellore, India, 2024, pp. 1–6, doi: 10.1109/ic-ETITE58242.2024.10493406.
- [81] B. Cavia, E. Horwitz, T. Reiss, and Y. Hoshen, "Real-time deepfake detection in the real-world," arXiv preprint arXiv:2406.09398, 2024, doi: 10.48550/arXiv.2406.09398.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)