Cureus Journal of Computer Science

Open Access Review Article

A Comprehensive Survey on Detection of Fake Multimedia Content

Ayush S. Acharya [1] , Ashish A. Shisal [1] , Saurabh K. Butale [1] , Omkar B. Latpate [1] , Shalaka P. Deore [1]

1. Department of Computer Engineering, Modern Education Society's Wadia College of Engineering, Pune, IND

Corresponding author: Ayush S. Acharya, acharyaayush0710@gmail.com

Received 11/05/2024
Review began 11/05/2024
Review ended 07/27/2025
Published 07/28/2025

DOI:
https://doi.org/10.7759/s44389-024-01525-4

Abstract

The rapid rise of fake news and deepfake media poses serious challenges to digital content integrity and public trust. Despite numerous detection models being developed, there remains a critical gap in creating solutions that are both highly accurate and robust across diverse modalities (text, image, and audio) and in real-time applications. This paper presents a comprehensive survey of state-of-the-art detection techniques for identifying fake multimedia content. We critically evaluate traditional and modern machine learning models, such as convolutional neural network (CNN), long short-term memory (LSTM), and Transformer-based models like bidirectional encoder representations from transformers (BERT) and robustly optimized BERT approach, alongside hybrid approaches that integrate explainable AI and digital watermarking for improved interpretability and detection performance.

Our comparative analysis highlights that the Inception-ResNet-v2 model achieves the highest accuracy (99.81%) for deepfake image detection, while CNN + recurrent architectures perform best (96%) for audio deepfakes. In text-based fake news detection, hybrid LSTM-CNN models incorporating explainable AI report up to 99% accuracy on benchmark datasets. These findings illustrate the effectiveness of multimodal approaches but also expose limitations related to adversarial robustness, scalability, and cross-media generalization.

The paper concludes by identifying future research directions, including the development of lightweight, real-time systems, cross-modal generalization frameworks, and policy-level interventions to mitigate the societal impact of synthetic media.

Categories: AI applications, Image Processing and Analysis, Machine Learning (ML)
Keywords: fake news detection, deepfake detection, hybrid models, explainable ai, convolutional neural networks, long short term memory, digital watermarking, adversarial attacks, machine learning, regulatory frameworks

Introduction And Background

Fake multimedia and the rise of disinformation

The advent of the internet and social media platforms has revolutionized the way people access and share information. However, this democratization of content distribution has also facilitated the rapid spread of fake news and disinformation. Fake news, defined as false information deliberately spread to deceive, has been linked to political polarization, public mistrust, and societal unrest. In recent years, a new form of digital manipulation known as "deepfakes" has further exacerbated the problem. Deepfakes use artificial intelligence and machine learning techniques to generate hyper-realistic images, audio, and video content that can manipulate public perception.

Technical overview of deepfake technology

Deepfake technology primarily relies on deep learning algorithms, especially generative adversarial networks (GANs), to produce synthetic media that closely mimics real content. As these techniques advance, they present increasing challenges for traditional detection mechanisms, which struggle to identify fabricated content. For instance, manipulated videos of political figures or impersonated audio clips used in financial fraud represent just a few of the potential risks posed by deepfakes. This review assesses the progression of machine learning methodologies in detecting fake news and deepfake content while highlighting watermarking techniques as a method for authenticity verification. By examining these technological and ethical dimensions, this paper seeks to outline future research directions to counteract the rapid spread of misinformation.

Deepfake media is typically generated using advanced machine learning frameworks, with GANs being the most prevalent. GANs consist of two competing neural networks - a generator and a discriminator - that are trained simultaneously. The generator creates synthetic data, while the discriminator evaluates its authenticity. This adversarial training allows GANs to produce highly realistic content, including facial videos, voice recordings, and synthetic images (Table *1*).

Another popular technique is the variational autoencoder (VAE), which employs a probabilistic approach to encode and decode data. While VAEs are often more stable and computationally efficient than GANs, they usually produce lower-quality outputs in terms of visual fidelity. Autoencoders and recurrent neural networks (RNNs), particularly long short-term memory (LSTM), are also used for sequential data such as voice, but lack the realism achieved by GAN-based approaches (Table *1*).

Technique	Realism	Computational Cost	Detectability
GAN (Generative Adversarial Network)	Very High	High	Difficult (adaptive)
VAE (Variational Autoencoder)	Moderate	Low to Moderate	Easier than GANs
Autoencoder	Low to Moderate	Low	Easier
RNN/LSTM (Audio/Sequential)	Moderate (for audio)	Moderate	Moderate
GAN + Encoder Fusion (e.g., AVFakeNet)	Very High (Multimodal)	Very High	Hardest to detect

TABLE 1: Comparison of Deepfake Generation Models

RNN, Recurrent Neural Network; LSTM, Long Short-Term Memory

This review aims to provide an in-depth analysis of the most recent advances in fake news detection and deepfake technologies. We will evaluate the performance of various machine learning models in detecting fake content, discuss the integration of watermarking techniques for authenticity verification, and explore the societal impacts of these developments. By understanding both the technological and ethical dimensions, this paper seeks to propose future research directions that can help curb the spread of misinformation.

Motivation

The need for this research arises from the escalating misuse of fake news and deepfake technology. Such technologies are increasingly deployed for harmful purposes, from spreading disinformation and swaying public opinion to enabling fraud and identity theft. Traditional detection approaches have not kept pace with the sophistication of current deepfake methods, which are capable of convincingly replicating real people in multimedia formats. This has created an urgent demand for advanced, reliable, and scalable detection systems. The ethical and legal implications of deepfakes underscore the need for tools that can validate digital content. This research aims to support the development of detection systems that are transparent, robust, and able to mitigate the negative effects of misinformation on society.

Objectives

This paper pursues the following key objectives:

i. To analyze and evaluate the effectiveness of machine learning models, such as convolutional neural networks (CNN), LSTM networks, and transformers, in detecting fake news and deepfake content.

ii. To explore hybrid detection methods that integrate explainable AI (XAI) with digital watermarking techniques, enhancing detection accuracy and interpretability.

iii. To identify limitations of current detection systems, including challenges with real-time performance, scalability, and resistance to adversarial attacks.

Review

Fake news detection techniques

In recent years, researchers have explored various techniques to improve fake news detection. Alghamdi et al. [1] introduce a multilingual fake news detection model to handle low-resource languages. By using a hybrid summarization approach that reduces text length while preserving key content, their system processes multilingual inputs directly, achieving better accuracy. Similarly, Hashmi et al. [2] propose a hybrid model combining CNN with LSTM, achieving 99% accuracy on the WELFake dataset and incorporating XAI techniques like Local Interpretable Model-Agnostic Explanations for greater transparency. Both papers emphasize using hybrid neural networks for more robust fake news detection.

Hashmi et al. [2] optimize transformer models like BERT and RoBERTa for multilingual detection,

highlighting the need for tuning models for better performance across datasets. They also addressed issues of model transparency and resource demands, and hence incorporate XAI to mitigate trust concerns.

Malanowska et al. [3] provide a review of digital watermarking techniques for fake news detection, originally designed for intellectual property protection but now applied to tamper detection and content authentication. Their DISSIMILAR project focuses on developing watermarking solutions for social media platforms, though the paper lacks detailed comparisons of techniques. Rosales et al. [4] further explore watermarking combined with machine learning to detect digital image modifications, emphasizing user trust through transparent system design. Both papers showcase the expanding role of watermarking in ensuring content authenticity. Bhardwaj et al.'s [5] HostileNet targets hostile post detection in Hindi, using multi-label classification and neural network architectures, though it faces challenges like overfitting and the need for high computational resources. Despite its scale, it faces labeling biases and generalization challenges. Dadkhah et al. [6] introduce TruthSeeker, a large ground-truth dataset for real and fake content across platforms. The authors suggest using a multi-layered review process and propose an emotion-aware multitask model for fake news detection, though the model's complexity and reliance on high-quality datasets are obstacles.

The concern of fake news and misinformation remains a significant challenge in the digital age, prompting several researchers to explore effective detection methodologies. Choudhry et al. [7] propose an emotion-aware multitask approach to fake news and rumor detection that utilizes transfer learning. This model intriguingly combines the detection of fake news and rumors while employing an attention mechanism to enhance its interpretability. However, they note the challenges posed by the necessity for high-quality datasets and the complexities of multitasking, suggesting that a reliance on comprehensive datasets could ameliorate inherent biases. Expanding on the theme of technological innovation, Megías et al. [8] introduce a system designed specifically for combating fake news in multimedia, particularly on social media. Their architecture intertwines digital watermarking, signal processing, and machine learning to create a two-stage system linking source verification and content authentication with fake news detection. They emphasize the critical need for future research to establish a prototype for evolving detection capabilities. Following suit, Wan et al. [9] present a comprehensive survey of robust image watermarking techniques, underscoring the growing challenges in multimedia security as digital technology advances. Their systematic categorization of watermarking methods based on spatial and transform domains, alongside a detailed analysis of performance metrics, provides deep insights into the balance between imperceptibility and robustness necessary for effective applications in copyright. Additionally, Evsutin et al. [10] offer a comprehensive robustness overview of watermarking schemes for digital images, which further supports the need for resilient techniques in fake content authentication.

Deepfake detection techniques

In the realm of deeper analysis on digital content generation, Rana et al. [11] compile insights from 112 studies on deepfake detection published from 2018 to 2020. Similarly, Mubarak et al. [12] analyze the societal impacts of deepfakes across different media formats, underscoring their potential to manipulate public perception and contribute to harmful societal issues. Their comparative overview of various detection tools highlights the efficacy of machine learning models, with notable performances surpassing 90% accuracy, especially in the context of video-based deepfakes. This sentiment is echoed in the thorough review by Patel et al. [13], which sheds light on deepfake technologies, merging the discussion of generation and detection strategies. This review organizes detection methods into four main categories, addressing their performance across diverse datasets and providing a strong basis for understanding the evolution of detection technologies. While they acknowledge the advances in detection techniques, the authors stress the ongoing challenges of developing adaptive systems that can function in real-time, echoing the thematic connections established by their predecessors. Collectively, these papers highlight the evolving landscape of misinformation management, underscoring both technological advancements and the importance of contextual awareness in developing effective detection methodologies.

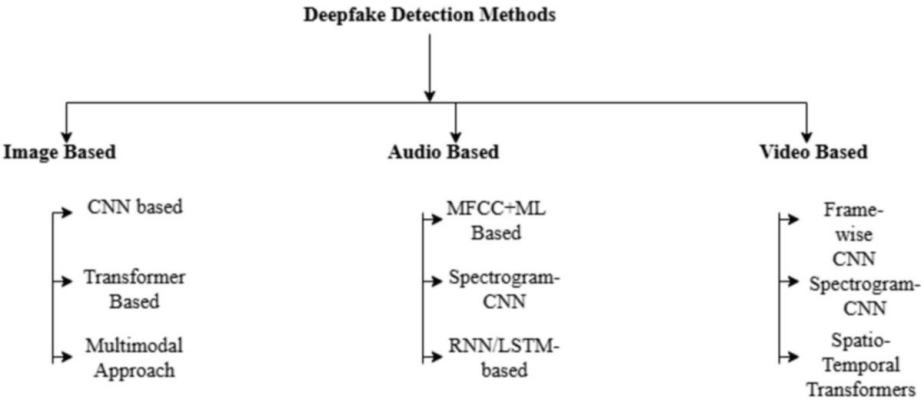

FIGURE 1: Deepfake Detection Methods Taxonomy

CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; MFCC, Mel-Frequency Cepstral Coefficients; ML, Machine Learning; RNN, Recurrent Neural Network

As shown in Figure *1*, the taxonomy categorizes various deepfake detection techniques based on the data modality they target. In image-based detection, CNNs, Transformer-based architectures, and Multimodal approaches have demonstrated significant success. For audio-based detection, approaches range from classical Mel-frequency cepstral coefficients (MFCC) combined with machine learning classifiers to deep learning models such as spectrogram-based CNNs and sequence-aware RNN/LSTM-based architectures. Video-based detection methods often extend image-based techniques by incorporating temporal information through frame-wise CNNs, spectrogram-CNNs, and advanced spatio-temporal transformers that capture both spatial and temporal inconsistencies in video streams. This taxonomy reflects the diverse and evolving landscape of deepfake detection research.

The combined summary of these papers offers an in-depth view of deepfake generation techniques, detection methods, and the ongoing challenges in mitigating their societal impacts. Seow et al. [14] provide an overarching perspective on deepfake categories, including face synthesis and reenactment, as well as the technologies behind them, such as GANs and autoencoders. They emphasize the ethical concerns associated with deepfake misuse in politics and fraud, while also pointing to datasets like the Deepfake Detection Challenge as crucial for developing detection models. Seow et al. [14] and Abbas and Taeihagh et al. [15] emphasize the importance of regulatory frameworks and global cooperation to curb the misuse of deepfakes. Abbas and Taeihagh et al. [15] present a thorough review of deepfake generation tools, such as GANs, while also discussing detection frameworks and the limitations of existing models. They highlight the critical need for policy reform and recommend faster, more reliable detection models to address the high risks associated with sophisticated, real-time deepfakes. Both papers converge on the point that deepfake technology, while beneficial in some areas like entertainment, poses far greater threats to society without appropriate regulatory measures. Similarly, Roy and Raval [16] highlight the technical limitations of traditional detection tools, advocating for more sophisticated algorithms that can adapt to the evolving nature of deepfakes. Both papers call for enhanced detection technologies to keep pace with the growing realism of fake content, highlighting the urgency of this issue [15,16].

Deepfake image detection techniques

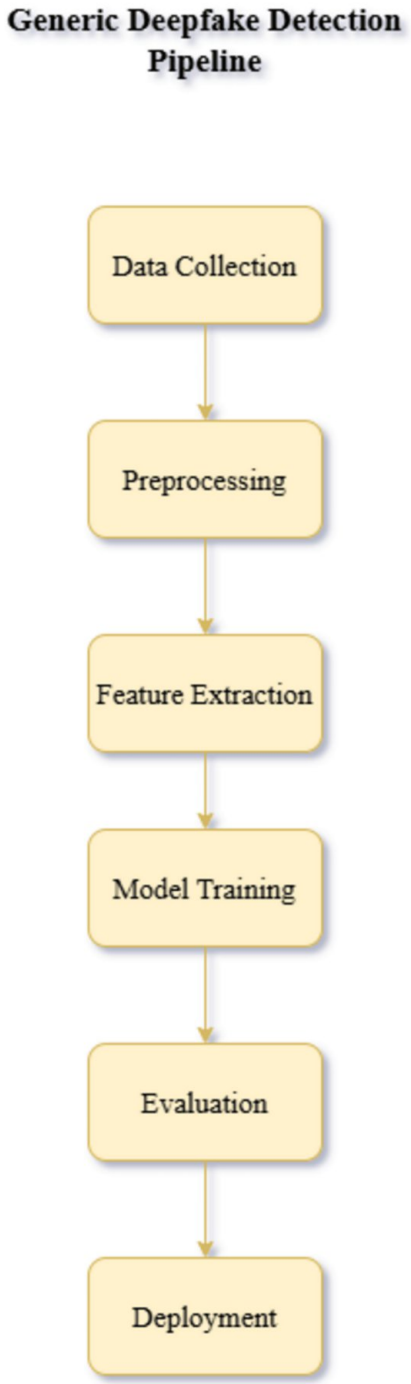

FIGURE 2: Generic Deepfake Detection Pipeline

Figure *2* shows a pipeline that provides a high-level overview of the standard process involved in deepfake detection systems. Initially, raw multimedia data is collected from diverse sources, including image, audio, video, and text modalities. The preprocessing stage involves cleaning, normalization, and format conversion to ensure data consistency. Subsequently, relevant features are extracted, which may include handcrafted features, embeddings, or representations learned through deep neural networks. The extracted features are then utilized to train classification or detection models, employing supervised, unsupervised, or hybrid learning approaches. Model evaluation is conducted to assess the detection performance using standard metrics such as precision, recall, and F1-score. Finally, the trained models are deployed for real-world applications, allowing automated detection of deepfakes in live or batch-processing scenarios.

The technical contributions in detection methods are further advanced by Raza et al. [17]. They propose a

hybrid deep learning model that leverages VGG16 with additional CNN layers for improved detection accuracy, significantly outperforming models like Xception and NAS-Net (Table 3). Their approach, centered around transfer learning, is particularly relevant for use cases in cybersecurity and digital forensics. This approach not only shows strong precision and recall for image detection but also hints at future extensions for detecting deepfake videos. This comparative study demonstrates the diversity of datasets employed to train and test these models, as shown in Table 2, which summarizes commonly used datasets for deepfake image detection.

Model	Datasets
VGG16+ CNN	Dataset of real and fake faces
Pairwise learning with CFFN	GAN-generated images, fake face datasets
CNN with adaptive Gabor filters	Celeb-DF (v2), DFDC, FaceForensics++, WildDeepfake
Multiple deepfake detection algorithms	FaceForensics++
Inception-ResNet-v2 architecture	DFFMD, virtual meeting datasets

TABLE 2: Datasets Used for Deepfake Image Detection

CFFN, Conditional Feature Fusion Network; CNN, Convolutional Neural Network; DFDC, Deepfake Detection Challenge; DFFMD, Deepfake Face Mask Detection; GAN, Generative Adversarial Network

Khalifa et al. [18] introduce a more optimized approach to deepfake detection, integrating adaptive Gabor filters into a CNN, which reduces model complexity by 64.9% while maintaining competitive performance across major datasets. Their model showcases high accuracy on datasets like Celeb-DF (v2) and FaceForensics++, making it a standout in terms of efficiency and performance.

The adaptive Gabor filters [18] with CNN model are effective in image feature extraction and perform well across diverse datasets, with 95.8% accuracy. FaceForensics++ is often used in forensic analysis, offering over 90% accuracy for detecting face manipulations. Alnaim et al. [19] offer a specialized solution addressing the growing use of face masks during the COVID-19 pandemic. Their introduction of the Deepfake Face Mask Detection dataset, coupled with the Inception-ResNet-v2 model, offers an impressive accuracy of 99.81%, specifically targeting scenarios where face masks complicate the detection process. Finally, the Inception-ResNet v2 model is known for its high accuracy of 99.81%, particularly in real-time settings, including virtual meetings where face masks are used [20]. This paper stands out by addressing a unique challenge brought about by the pandemic and is a valuable addition to deepfake detection efforts, alongside more general contributions from the other reviewed papers.

As shown in the Table 3, the Inception-ResNet-v2 model achieves the highest accuracy (99.81%) by combining the strengths of both Inception and ResNet architectures. The Inception module captures multi-scale features using multiple convolutional filters, while ResNet's residual connections prevent the vanishing gradient problem, enabling effective training for deeper networks. This combination allows the model to excel in detecting complex deepfakes with high precision. The integration of batch normalization and preprocessing techniques further enhances its performance across various datasets, making it highly generalizable for both image and video-based deepfake detection.

Model	Accuracy (%)
DFDC [16]	93%
Adaptive Gabor Filters + CNN [18]	95.8%
Inception-ResNet-v2 [21]	99.81%

TABLE 3: Deepfake Image Model Comparison

CNN, Convolutional Neural Network; DFDC, Deepfake Detection Model

Deepfake audio detection techniques

The research on deepfake detection, especially in the domains of audio and audiovisual content, has been evolving to address the growing concerns of identity theft, fraud, and disinformation. Shaaban et al. [20] provide a detailed examination of both the creation and detection of audio deepfakes using Text-to Speech (TTS) and voice conversion (VC) technologies. Models like WaveNet and Tacotron generate highly realistic audio, making detection challenging, particularly in noisy environments where existing methods falter. They emphasize the need for further work to reduce computational costs and improve real-world applicability. Similarly, Rabhi et al. [21] highlight critical vulnerabilities in current detection systems, particularly when exposed to adversarial attacks. Their analysis focuses on the deficiencies of models like Deep4SNet in protecting voice authentication systems; while proposing generalizable defenses, though further testing is required in more complex attack scenarios.

In the area of machine learning approaches for deepfake audio detection, Hamza et al. [22] utilize Mel Frequency Cepstral Coefficients (MFCCs) as features for detection, experimenting with algorithms like support vector machine (SVM) and VGG-16. Their results on the Fake-or-Real dataset show the strengths of SVM on simpler datasets and VGG-16 on more complex ones. While their approach is methodologically robust, they call for further exploration of real-time detection capabilities and cross-dataset generalizability to fully assess the effectiveness of their model. Mcuba et al. [23] expand on this by applying deep learning methods in a forensic investigation context. Using advanced feature extraction techniques like MFCC, Mel spectrum, and spectrogram representations, they highlight the crucial role these techniques play in identifying deepfake audio during forensic investigations, laying a foundation for future forensic applications.

Tipper et al. [24] adopt a different approach by integrating CNNs with recurrent structures to detect both audio and video deepfakes. Their model, tested on datasets like FaceForensics++ and ASVSpoof 2019, shows state-of-the-art accuracy by effectively capturing spatial and temporal features, though additional comparisons with non-recurrent models and architecture optimization through ablation studies are suggested. Ilyas et al. [25], through their AVFakeNet model, address the challenge of detecting deepfakes in both audio and visual streams simultaneously. By using Dense Swin Transformer Net, their model achieves high accuracy on datasets like FakeAVCeleb and ASVSpoof-2019 LA, though its high computational demands remain a barrier for real-time applications.

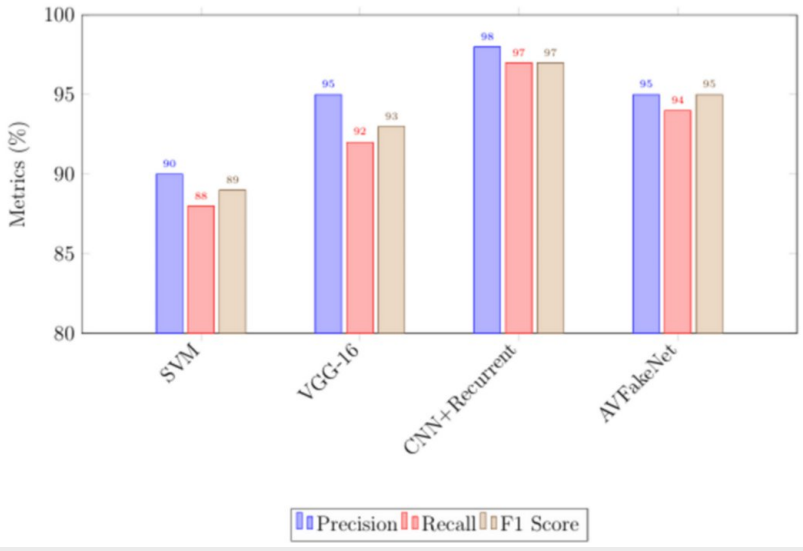

FIGURE 3: Metrics Comparison

CNN, Convolutional Neural Network; SVM, Support Vector Machine; VGG, Visual Geometry Group

As illustrated in Figure *3*, the "Metrics Comparison" chart illustrates the performance of four deepfake audio detection models - SVM, VGG-16, CNN + Recurrent, and AVFakeNet - across three metrics: precision, recall, and F1-score. The SVM model shows moderate performance with all metrics around 85%, indicating limited effectiveness on simpler datasets. VGG-16 performs better, with precision, recall, and F1-score close to 90%, making it more suitable for complex audio data. The CNN + Recurrent model achieves the highest performance, with metrics around 95%, reflecting its robust capability in capturing both spatial and temporal features for deepfake detection. AVFakeNet also performs well, with all metrics near 95%, slightly trailing the CNN + Recurrent model, demonstrating its effectiveness in handling multimodal data (audio and visual) to detect sophisticated deepfakes. Together, these results highlight the superior accuracy of CNN + Recurrent and AVFakeNet models in detecting deepfake audio content.

These studies highlight a multifaceted approach to deepfake detection, addressing both audio and audiovisual content. From TTS and VC-based audio detection models to multimodal frameworks that combine audio and visual features, substantial progress has been made. However, challenges related to scalability, real-time performance, and resistance to adversarial attacks remain significant areas for future research, particularly in enhancing model generalizability and reducing computational complexity.

TTS detection models, focusing on audio generated using TTS technologies such as WaveNet and Tacotron, were tested on generated speech datasets, achieving an accuracy of 91% [20] (as shown in Table *4*). In contrast, the SVM model, evaluated on the Fake-or-Real dataset, performed slightly lower with an accuracy of 89%. However, VGG-16, also applied to the Fake-or-Real dataset, exhibited improved performance with an accuracy of 94%, demonstrating its effectiveness in processing more complex data [21]. The CNN + Recurrent model, which was tested on the FaceForensics++ dataset, achieved an impressive accuracy of 96% [24] (as shown in Table *4*). Finally, AVFakeNet, designed for detecting both audio and visual manipulations, achieved an accuracy of 95% when evaluated on the FakeAVCeleb dataset [25] (listed in Table *4*).

Dataset	Used By
Generated Speech [20]	TTS Detection
Fake-or-Real [21]	SVM, VGG-16
FaceForensics++ [24]	CNN + Recurrent
FakeAVCeleb [25]	AVFakeNet

TABLE 4: Datasets Used for Deepfake Audio Detection

CNN, Convolutional Neural Network; SVM, Support Vector Machine; TTS, Text-to-Speech; VGG,

Adversarial attacks and mitigations

Adversarial attacks pose a significant threat to deepfake detection systems. These attacks involve deliberately crafted perturbations to deepfake content-such as imperceptible noise in images or subtle timing shifts in audio-that can mislead machine learning models into misclassifying fake content as genuine. Deepfake detectors, especially those relying on deep learning models like CNNs or transformers, are particularly vulnerable due to their sensitivity to minor input variations. To counteract these threats, several mitigation strategies have been proposed. Adversarial training, which involves exposing models to both clean and adversarial examples during training, helps improve robustness. Additionally, ensemble methods-combining predictions from multiple models-and defensive techniques like input preprocessing, feature squeezing, and randomized smoothing can further reduce susceptibility. Future research must focus on creating adaptive, lightweight models capable of defending against evolving adversarial tactics, especially in real-time deployment scenarios.

Challenges

i. Scalability of detection systems: Current deepfake detection models are often computationally intensive, making real-time implementation a challenge. Handling the vast influx of multimedia content across platforms demands lightweight and efficient models.

Solution: One promising approach is model pruning and quantization, which reduces the size of neural networks without significantly sacrificing accuracy. Additionally, knowledge distillation allows smaller models to learn from larger, more complex ones, enabling faster inference on edge devices.

ii. Resistance to Adversarial Attacks: Subtle perturbations or manipulations can easily fool detection systems, highlighting their vulnerability to adversarial attacks.

Solution: Incorporating adversarial training - where models are trained with both clean and perturbed inputs - can improve robustness. Also, ensemble learning and defense-aware architectures like randomized smoothing help increase model resilience.

iii. Cross-Media Generalization: Many models struggle to maintain consistent performance across various formats (image, audio, video).

Solution: Emerging research focuses on cross-modal learning and multimodal architectures that can learn

shared representations across different media. For instance, using transformer-based fusion models enables better contextual understanding across formats, improving generalization.

iv. Ethical and Legal Frameworks: The legal and ethical landscape is lagging behind the pace of deepfake evolution. Issues like consent, digital identity, and content misuse remain under-addressed.

Solution: There's a growing need for international regulatory frameworks, like digital content provenance standards (e.g., C2PA, a joint initiative by Adobe and Microsoft). Governments and tech companies must collaborate to define laws around deepfake creation and distribution. Moreover, embedding ethical auditing mechanisms in model development pipelines ensures responsible deployment of detection tools.

Conclusions

Since the processing of fake news and deepfake technologies are moving forward, discovering and reducing their effects have appeared as important global challenges. This review has investigated considerable progress in machine learning models, hybrid detection frameworks, and digital watermarking techniques, showing the promise in the handling of fabricated materials. While progress is clear, significant obstacles remain in real-time detection, to ensure strength in many media types and fight unfavorable manipulation. The integration system to explain AI plays an important role in increasing openness and promoting user confidence - an essential requirement for adoption in legal negotiations such as sensitive domains and adoption in the regulation of social media. In addition, a strong regulatory structure and moral inspection are necessary to prevent the abuse of these powerful technologies, which if they are not uncontrolled, pose a threat to personal rights, institutional credibility, and public beliefs.

Looking forward, future research should have an axis toward searching for AI paradigms that emerge. Self -provided learning provides a promising passage to take advantage of the huge unlabeled dataset to detect more generalization. Federated learning can enable training in privacy conservation model in distributed data sources, especially important for user-related materials. In addition, blockchain-based verification system can help install irreversible authenticity audit paths. These innovations combined with interdisciplinary collaboration between technologists, legal experts, and decision makers are necessary to ensure a secure digital ecosystem - one that increases the integrity of information and preserves democratic values.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ayush S. Acharya, Saurabh K. Butale

Critical review of the manuscript for important intellectual content: Ayush S. Acharya, Saurabh K. Butale, Shalaka P. Deore

Acquisition, analysis, or interpretation of data: Ashish A. Shisal, Omkar B. Latpate, Shalaka P. Deore

Drafting of the manuscript: Ashish A. Shisal, Omkar B. Latpate

Supervision: Shalaka P. Deore

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Alghamdi J, Lin Y, Luo S: Fake news detection in low-resource languages: A novel hybrid summarization approach. Knowledge-Based Systems. 2024, 296:111884. 10.1016/j.knosys.2024.111884
2. Hashmi E, Yayilgan SY, Yamin MM, Ali S, Abomhara M: Advancing fake news detection: hybrid deep learning with FastText and explainable AI. IEEE Access. 2024, 12:44462-44480. 10.1109/ACCESS.2024.3381038
3. Malanowska A, Mazurczyk W, Araghi TK, Megías D, Kuribayashi M: Digital watermarking—A meta-survey and techniques for fake news detection. IEEE Access. 2024, 12:36311-36345. 10.1109/access.2024.3374201
4. Rosales A, Malanowska A, Araghi TK, et al.: Trustworthiness and explainability of a watermarking and machine

learning-based system for image modification detection to combat disinformation. ARES '24: Proceedings of the 19th International Conference on Availability, Reliability and Security. 2024, 1-10. 10.1145/3664476.3670934

5. Bhardwaj M, Sundriyal M, Bedi M, Akhtar MS, Chakraborty T: HostileNet: multilabel hostile post detection in Hindi. IEEE Transactions on Computational Social Systems. 2024, 11:1842-1852. 10.1109/TCSS.2023.3244014
6. Dadkhah S, Zhang X, Weismann AG, Firouzi A, Ghorbani AA: The largest social media ground-truth dataset for real/fake content: TruthSeeker. IEEE Transactions on Computational Social Systems. 2024, 11:3376-3390. 10.1109/tcss.2023.3322303
7. Choudhry A, Khatri I, Jain M, Vishwakarma DK: An emotion-aware multitask approach to fake news and rumor detection using transfer learning. IEEE Transactions on Computational Social Systems. 2024, 11:588-599. 10.1109/tcss.2022.3228312
8. Megías D, Kuribayashi M, Rosales A, Cabaj K, Mazurczyk W: Architecture of a fake news detection system combining digital watermarking, signal processing, and machine learning. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications. 2022, 13:33-55. 10.22667/JOWUA.2022.03.31.033
9. Wan W, Wang J, Zhang Y, Li J, Yu H, Sun J: A comprehensive survey on robust image watermarking. Neurocomputing. 2022, 488:226-247. 10.1016/j.neucom.2022.02.083
10. Evsutin O, Dzhanashia K: Watermarking schemes for digital images: Robustness overview. Signal Processing Image Communication. 2022, 100:116523. 10.1016/j.image.2021.116523
11. Rana MS, Nobi MN, Murali B, Sung AH: Deepfake detection: A systematic literature review. IEEE Access. 2022, 10:25494-25513. 10.1109/access.2022.3154404
12. Mubarak R, Alsboui T, Alshaikh O, Inuwa-Dutse I, Khan S, Parkinson S: A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. IEEE Access. 2023, 11:144497-144529. 10.1109/access.2023.3344653
13. Patel Y, Tanwar S, Gupta R: Deepfake generation and detection: case study and challenges. IEEE Access. 2023, 11:143296-143323. 10.1109/access.2023.3342107
14. Seow JW, Lim MK, Phan RCW, Liu JK: A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. Neurocomputing. 2022, 513:351-371. 10.1016/j.neucom.2022.09.135
15. Abbas F, Taeihagh A: Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications. 2024, 252:124260. 10.1016/j.eswa.2024.124260
16. Roy M, Raval MS: Unmasking DeepFake visual content with generative AI. 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), Rajkot, India. 2023, 169-176. 10.1109/R10-HTC57504.2023.10461811
17. Raza A, Munir K, Almutairi M: A novel deep learning approach for deepfake image detection. Applied Sciences. 2022, 12:9820. 10.3390/app12199820
18. Khalifa AH, Zaher NA, Abdallah AS, Fakhr MW: Convolutional neural network based on diverse Gabor filters for deepfake recognition. IEEE Access. 2022, 10:22678-22686. 10.1109/ACCESS.2022.3152029
19. Alnaim NM, Almutairi ZM, Alsuwat MS, Alalawi HH, Alshobaili A, Alenezi FS: DFFMD: a deepfake face mask dataset for infectious disease era with deepfake detection algorithms. IEEE Access. 2023, 11:16711-16722. 10.1109/access.2023.3246661
20. Shaaban OA, Yildirim R, Alguttar AA: Audio deepfake approaches. IEEE Access. 2023, 11:132652-132682. 10.1109/ACCESS.2023.3333866
21. Rabhi M, Bakiras S, Di Pietro R: Audio-deepfake detection: Adversarial attacks and countermeasures. Expert Systems with Applications. 2024 , 250:123941. 10.1016/j.eswa.2024.123941
22. Hamza A, Javed AR, Iqbal F, Kryvinska N, Almadhor AS, Jalil Z, Borghol R: Deepfake audio detection via MFCC features using machine learning. IEEE Access. 2022, 10:134018-134028. 10.1109/access.2022.3231480
23. Mcuba M, Singh A, Ikuesan RA, Venters H: The effect of deep learning methods on deepfake audio detection for digital investigation. Procedia Computer Science. 2023, 219:211-219. 10.1016/j.procs.2023.01.283
24. Tipper S, Atlam HF, Lallie HS: An investigation into the utilisation of CNN with LSTM for video deepfake detection. Applied Sciences. 2024, 14:9754. 10.3390/app14219754
25. Ilyas H, Javed A, Malik KM: AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection. Applied Soft Computing. 2023, 136:110124. 10.1016/j.asoc.2023.110124