Sami Seppälä

Review of deepfake detection methods

Final report

School of Technology and Innovations
Bachelor's thesis in Energy and Information Technology
Automation and Information Technology

Vaasa 2025

UNIVERSITY OF VAASA
School of Technology and Innovation
Author: Sami Seppälä
Title of the Thesis: Review of deepfake detection methods : Final report
Degree: Bachelor of Engineering Sciences
Program: Automation and Information Technology
Supervisor: Tomi Pasanen
Year: 2025 **Page Count:** 31

ABSTRACT:
This bachelor's thesis reviews methods for detecting deepfake media, examining peer-reviewed literature published between 2020 and 2025. This review also briefly covers deepfake generation methods such as GANs, autoencoders, and diffusion models. Detection methods are grouped into twelve categories, including CNN-based classification and physiological signal analysis. While many methods score above 95% accuracy on standard benchmarks, they struggle when tested on data they were not trained on, and video compression makes detection even harder. Methods based on physiological signal measurement and identity features performed better in robustness testing. The review concludes with recommendations for future research priorities.

KEYWORDS: Deepfake, Artificial Intelligence, Deep learning, Information integrity

Contents

Abbreviations

General/Technical:

AI – Artificial Intelligence

AUC – Area Under the Curve

CNN – Convolutional Neural Network

DCGAN – Deep Convolutional Generative Adversarial Network

GAN – Generative Adversarial Network

RL – Reinforcement Learning

rPPG – Remote Photoplethysmography

VAE – Variational Autoencoder

Datasets:

DFDC – DeepFake Detection Challenge

DFDCP – DeepFake Detection Challenge Preview

Method-Specific:

ADAL – Artifacts-Disentangled Adversarial Learning

ADT – Anti-Deepfake Transformer

AST – Audio Spectrogram Transformer

I2G – Inconsistency Image Generator

PCL – Pair-wise self-Consistency Learning

SBI – Self-Blended Image

VFD – Voice-Face Matching Detection

1 Introduction

Deepfakes are fake pieces of media, either images, videos, or audio, that are altered in a way that the person or people depicted in the media take on some other individual's likeness or voice believably, but without participating (Zhang, 2022). Deepfakes can also be used just to alter a person's facial expression, without changing their likeness to resemble someone else. They are created using methods based on artificial intelligence and machine learning (Kietzmann, 2020), and they are becoming more believable and easier to create, as the technology is becoming more common and more accessible.

Currently, deepfakes present many opportunities for malicious activities, and this threat raises many security concerns (Masood et al., 2023). This has led to many methods being developed for detecting and combating deepfakes, often employing different approaches to the problem. These methods are split between machine learning based methods and feature-based detection methods (Zhang, 2022). Different methods are used for detecting fake images, video, and audio (Nguyen et al., 2022).

This review aims to investigate literature regarding methods for detecting deepfakes currently in use or development and briefly summarize deepfakes as a phenomenon and as a piece of technology. It is structured as follows: Section 2 provides the literature review, beginning with the technical foundations of deepfake creation before examining current detection methods organized by approach. Section 3 synthesizes the findings, comparing performance across methods and identifying key factors that influence detection accuracy. Section 4 concludes with a summary and recommendations for future research.

1.1 Background and significance

Manipulated pieces of media have existed since the early days of photography, film, and audio recording. However, recent advancements in artificial intelligence (AI), machine learning, and deep learning have introduced highly sophisticated new tools and techniques for creating synthetic media (Rana et al., 2022). These tools and techniques have become increasingly accessible and produce more realistic results (Zhang, 2022), making

it possible for almost anyone to create altered images, video, or audio in an instant that is difficult to distinguish from real content.

The first appearance of deepfakes happened in the year 2017 when a Reddit user, named “deepfakes”, began posting pornographic videos where the performers’ faces were substituted with celebrity faces, without their consent, using deep learning (Mirsky & Lee, 2022). Later, in 2018, a video of former United States president Barack Obama giving a speech on the matter was published by BuzzFeed; it was created using the same software used to create the videos posted to Reddit. Since these events, large numbers of deepfake videos have begun to emerge, and continue to do so to this day.

As a piece of technology, deepfakes have multiple positive applications, and for example, are being used for enhancements in filmmaking and virtual reality (Yu et al., 2021). Deepfake technology has legitimate uses, but it is far more often discussed in the context of misuse, as in privacy violations, political manipulation, and misinformation. Deepfakes have already been used to damage personal privacy and spread political misinformation, and as the technology improves, the potential for abuse grows with it (Zhang, 2022).

As deepfake quality improves, both humans and algorithms are finding it harder to tell real from fake (Nguyen et al., 2022). This makes reliable detection methods more important than ever. Addressing this has led to the development of deepfake detection methods that often employ the same deep learning techniques that are used to create them in the first place, such as generative adversarial networks (GANs) and convolutional neural networks (CNNs) (Rana et al., 2022).

1.2 Literature search and selection strategy

This literature review integrates findings from peer-reviewed articles published between 2020 and 2025. These articles focus on different detection methods for deepfake media. Relevant sources were retrieved from well-established academic databases and publisher platforms, including IEEE Xplore, arXiv, ACM Digital Library, PubMed, PubMed Central, ScienceDirect, and SpringerLink, as well as Taylor & Francis, Wiley Online Library, CVF Open Access, MDPI, and AAAI Digital Library.

To identify relevant studies, Semantic Scholar and Google Scholar were used to explore databases. Keyword searches included search terms such as “Deepfake detection”, “AI-generated content”, “synthetic media detection”, “deepfake video detection”, “audio deepfake recognition”, “machine learning in deepfake detection”, “GAN-based deepfake detection”, and “deep learning based deepfake detection”. Results were filtered based on publication year, peer-review status, and their relevance to deepfake detection techniques.

2 Literature review

This section reviews recent research on deepfake detection, focusing on detection methods, benchmarking datasets, and key challenges; it also briefly discusses methods for generating deepfakes to provide context for detection approaches.

2.1 Technical Foundations of Deepfake Creation

To better understand and develop methods for detecting deepfakes, it is essential to know how synthetic media is created. This section provides a technical overview of the primary machine learning approaches used for deepfake generation. Understanding the creation methods helps identify the characteristic artifacts and patterns that detection algorithms exploit to distinguish fake content.

Deepfakes can be categorized by media type and the nature of manipulation. Visual deepfakes can be split into four categories: reenactment, replacement, editing, and synthesis (Mirsky & Lee, 2022). An example of reenactment would be the use of a source face or body to drive the expression or pose of a target. Face swapping would be classified as replacement. Editing focuses on methods for facial attribute editing, for example, while synthesis involves generating photorealistic images or videos of individuals who do not exist in real life.

Audio deepfakes commonly include text-to-speech synthesis, in which the target's voice reads written text aloud, and voice conversion, in which the speech of a source speaker is altered to resemble the target's voice.

The following subsections examine the core architectural components that underpin these techniques.

2.1.1 Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) are a fundamental deep learning architecture that has become the backbone of computer vision applications (Zhang, 2022), as their ability to capture hierarchical patterns within data makes them highly effective for processing

visual information (Mirsky & Lee, 2022). In deepfake generation, CNNs serve as building blocks within larger generative frameworks, providing the computational mechanisms for feature extraction and image reconstruction. A CNN produces a feature map by sliding learned filters systematically across the input image, detecting local patterns at each position. The feature map represents the presence and location of specific visual patterns across the image (Mirsky & Lee, 2022).

Unlike generative adversarial networks (GANs), which constitute a complete adversarial training framework, CNNs are most often used as components within deepfake generation pipelines. They are commonly integrated into encoder-decoder architectures, autoencoders, and as the generator or discriminator networks within GANs themselves (Masood et al., 2023; Mirsky & Lee, 2022). This CNN-based approach with GANs was first introduced by Radford et al. as Deep Convolutional GAN (DCGAN) (Radford et al., 2016), shortly after GANs themselves were first introduced (Goodfellow et al., 2014).

CNNs can adapt to various deepfake creation tasks, offering clear benefits but also several drawbacks. CNNs excel in image processing efficiency, especially when compared to fully connected networks, and they produce high-fidelity images with realistic features within the deepfake generation context (Masood et al., 2023; Mirsky & Lee, 2022). By stacking multiple convolutional layers, CNNs are able to capture more complex variations in human faces, with each layer learning increasingly complex feature representations from the previous layer (Nguyen et al., 2022). Furthermore, CNNs are able to handle different levels of detail simultaneously – from fine-grained textures to overall facial structure (Nguyen et al., 2022). However, CNN-based methods require large amounts of training data, and are often subject-specific (Masood et al., 2023). They are also prone to visual artifacts in generated content, especially when significant modification is needed (Masood et al., 2023).

2.1.2 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) were first introduced by Goodfellow et al. (Goodfellow et al., 2014) as a framework for generating realistic synthetic data. The framework uses an adversarial training process in which a generator and a discriminator

compete. The generator learns to produce synthetic data that mimics real data as closely as possible to fool the discriminator, while the discriminator is trained to distinguish between fake and real samples. This process improves the generator's ability to produce convincing synthetic outputs that the discriminator struggles to differentiate from real data. Once training is complete, the discriminator is no longer required, and the generator can synthesize new data independently (Mirsky & Lee, 2022).

In the context of deepfake creation, GANs have become the dominant tool for face manipulation. The generator network learns to map facial features from a source to a target person, while the discriminator ensures the output appears realistic. Popular deepfake applications use variants of GANs, such as CycleGAN (Zhu et al., 2020) and StyleGAN (Karras et al., 2019), which have been specifically adapted for manipulation tasks and facial synthesis.

The adversarial training process makes GANs particularly effective for deepfake generation because the discriminator acts as a quality control mechanism, pushing the generator to synthesize better samples of fake faces that eventually fool the automated training system and potentially a human observer presented with images created by the generator.

2.1.3 Encoder-Decoders, Autoencoders, and VAEs

Encoder-decoders are a fundamental architecture used in deepfake generation. Specifically, they are a type of neural network architecture designed to learn efficient representations of data through compression and reconstruction; the minimum structural requirement is that they consist of at least two networks: the encoder and the decoder. The encoder compresses the input data into a lower-dimension latent representation – a compact numerical form capturing essential features, while the decoder produces the output from this compressed representation (Mirsky & Lee, 2022).

An autoencoder is a specific type of encoder-decoder network that aims to reconstruct the input as output. To facilitate reconstruction, the encoder and decoder need to be dimensionally compatible (Minaee et al., 2020; Mirsky & Lee, 2022). Autoencoders were used in the original 2017 Reddit deepfake generation network, where a shared

encoder was paired with person-specific decoders. These components were trained in parallel as two separate autoencoders. This enabled the encoder to map the features of its inputs, such as pose and expression, into a shared latent space. Face swapping was then achieved by encoding the source person's face with the shared encoder, and reconstructing it with the target person's decoder (Mirsky & Lee, 2022). The shared encoder learns identity-independent facial features, while each decoder learns person-specific reconstruction, thus enabling identity transfer while preserving expression and pose.

Variational autoencoders (VAEs) are an advanced form of autoencoders that include probabilistic elements by learning distributions in latent space rather than fixed points (Mirsky & Lee, 2022). First introduced in 2013 (Kingma & Welling, 2013), VAEs did away with mapping features to specific coordinates, instead encoding to a distribution – typically Gaussian – characterized by mean and variance. This probabilistic approach enables smooth interpolation between expressions and generation of natural variations not present in training data, thus producing more realistic deepfakes as a result (Pei et al., 2024).

2.1.4 Diffusion models

Diffusion models represent a more recent approach to deepfake generation. They are probabilistic generative models that create realistic high-quality content by reversing a noise-infusion process, gradually transforming random noise into realistic images or videos (Ayodele R. Akinyele et al., 2024; Chen et al., 2024). These models work through forward and reverse processes; first adding noise to data over multiple steps, then learning to denoise the data and generate new realistic content (Ayodele R. Akinyele et al., 2024). This approach enables high-quality synthesis with greater stability and control compared to earlier methods, such as GANs and VAEs, and has been widely adopted in computer vision and audio generation applications (Chen et al., 2024). However, diffusion models are computationally expensive due to their multi-step generation process, which is a considerable limit regarding their scalability (Ayodele R. Akinyele et al., 2024).

2.2 Deepfake Detection Methods

The following subsections present detection methods organized by technical approach, with consistent reporting of accuracy metrics to enable comparison in Section 3. Detection methods are typically evaluated in two settings: within-dataset, where training and testing use different portions of the same dataset, and cross-dataset, where the model is tested on entirely different data than it was trained on.

2.2.1 CNN-Based Detection Methods

Convolutional Neural Networks have become the foundation of most detection approaches. Researchers commonly use CNN architectures such as XceptionNet and EfficientNet (Dolhansky et al., 2020), ResNet (T. Zhao et al., 2021), and Inception-ResNet (Singh et al., 2021) as backbone networks, the base model that extracts features before classification, for binary classification tasks, distinguishing real from fake content.

To improve detection capabilities, several studies have combined CNNs with specialized modules. The multi-attentional network by Zhao et al. used multiple spatial attention heads, components that identify the most relevant regions in an image, to focus on local discriminative features while enhancing textural information from shallow features (H. Zhao et al., 2021). This approach achieved 97.60% accuracy and 99.29% AUC (area under the curve, a classification metric where 1.0 is perfect and 0.5 is random chance) on high-quality FaceForensics++ data, demonstrating state-of-the-art performance (H. Zhao et al., 2021). Raza et al. proposed a hybrid combining a VGG16 with another CNN architecture, achieving 95% precision and 94% accuracy (Raza et al., 2022).

For spatiotemporal analysis, some researchers turned to 3D CNNs. Almestekawy et al. employed an enhanced 3D CNN with spatiotemporal attention in a Siamese architecture, which processes two inputs through identical networks to compare them, combining texture features with deep learning. This approach improved accuracy by 7.9% and achieved AUC scores of up to 97.51% in the same-dataset scenarios and 95.44% in cross-dataset evaluation (Almestekawy et al., 2024).

2.2.2 Transformer-Based Approaches

Transformer architectures have emerged as alternatives to CNNs. Khan et al. developed a hybrid transformer network with early feature fusion, combining XceptionNet and EfficientNet-B4 as feature extractors trained jointly as a single system with a BERT-styled transformer. Their model achieved 95.00% accuracy on face swapping and 90.00% on NeuralTextures in FaceForensics++ (Khan & Dang-Nguyen, 2022).

Wang et al. took a different approach with the Anti-Deepfake Transformer (ADT), which modeled both global and local information via attention-based modules, multi-forensic modules, and variant residual connections. By capturing both fine details and overall structure, this approach addressed the limitations of CNN methods that rely too heavily on local texture information, achieving state-of-the-art performance in cross-dataset evaluation (P. Wang et al., 2022).

Transformers have also shown promise for audio deepfakes. Channing et al. evaluated transformer-based models, including AST and Wav2Vec-based architectures. AST achieved 85% accuracy on FakeAVCeleb with 0.985 AUC, while Wav2Vec reached 81% accuracy with 0.990 AUC; both of these transformer models outperforming traditional methods such as Gradient Boosting Decision Trees (Channing et al., 2024).

2.2.3 Physiological Measurement Methods

Physiological signals present in videos provide distinctive cues for detection. Hernandez-Ortega et al. developed DeepFakesON-Phys using remote photoplethysmography (rPPG) to analyze subtle color changes in skin that reveal blood flow patterns (Hernandez-Ortega et al., 2022). Their method employed a Convolutional Attention Network to extract spatial and temporal information from video frames. Single-frame detection achieved over 98% AUC on both Celeb-DF v2 and DFDC databases, and when combining scores from consecutive frames, the system reached 100% accuracy on Celeb-DF v2 (Hernandez-Ortega et al., 2022).

The strength of these methods lies in exploiting physiological signals that are difficult for current deepfake generation techniques to replicate (Juefei-Xu et al., 2022). However,

they require careful extraction of physiological features and can be sensitive to video quality and lighting conditions.

2.2.4 Multimodal and Audio-Visual Methods

Several approaches have leveraged consistency between audio and visual modalities. Cheng et al. proposed voice-face matching based on the principle that individuals exhibit high homogeneity between voice and face, while deepfakes often involve mismatched identities (Cheng et al., 2023). Their VFD method first trained on a large general dataset, then refined the model on deepfake-specific data, achieving 86.11% AUC on FakeAVCeleb and outperforming baselines by nearly 2% (Cheng et al., 2023).

Taking a different approach, Wang et al. developed ART-AVDF using articulatory representation learning, employing an audio encoder to extract articulatory features and a lip encoder trained through self-supervised learning, where the model learns patterns from unlabeled data (Y. Wang & Huang, 2024). Their system integrated a multimodal joint-fusion module to exploit inherent audio-visual consistency, achieving significant performance improvements over comparable models.

Muppalla et al. combined audio-visual features with fine-grained deepfake classification, categorizing samples into four types based on modality-specific labels (Muppalla et al., 2023). Using Capsule networks and Swin Transformers, their approach achieved 99.20% accuracy with Capsule Forensics on FakeAVCeleb. They tested both feature fusion, combining raw data representations, and score fusion, combining final classification outputs.

2.2.5 Temporal Consistency and Recurrent Approaches

Temporal analysis exploits inconsistencies across frames in deepfake videos. Chintha et al. combined convolutional latent representations with bidirectional recurrent structures and entropy-based cost functions. Their XcepTemporal model achieved 100% accuracy on FaceForensics++ for both frame-level and video-level detection, enabling identification of both spatial and temporal signatures of deepfakes (Chintha et al., 2020).

To address the specific challenge of compressed deepfakes, Hu et al. proposed a two-stream method that analyzes frame-level and temporal-level features. The frame-level stream gradually pruned the network to prevent overfitting, as in learning noise patterns specific to training data rather than general features, to compression artifacts, while the temporality-level stream extracted temporal correlation features. This approach outperformed state-of-the-art methods on compressed videos (Hu, Liao, Wang, et al., 2022).

Liu et al. developed a detector that leverages temporal consistency to distinguish between clean and perturbed videos, achieving 100% detection accuracy for weakly adversarial deepfakes. Their work also introduced a framework for generating high-quality adversarial deepfake videos using optical flow to restrict the temporal coherence of adversarial perturbations (Liu et al., 2023).

Caldelli et al. took a more direct approach, utilizing optical flow fields, a technique which tracks pixel movement between consecutive frames, to detect motion dissimilarities in video sequences. Their method achieved 97% accuracy for uncompressed videos (C0), 91% for lightly compressed (C23), and 76% for heavily compressed video (C40) in the same-forgery scenarios. Notably, this approach demonstrated superior robustness in cross-forgery scenarios compared to frame-based methods (Caldelli et al., 2021).

2.2.6 Identity-Based and Metric Learning Methods

Identity-aware approaches characterize individuals through biometric traits. Cozzolino et al. introduced ID-Reveal, which learned temporal facial features specific to how individuals move while talking through metric learning, training the model to measure similarity between examples, and adversarial training. What makes this approach notable is that it required only real videos for training, not fake data. ID-Reveal achieved more than 15% average improvement in accuracy for facial reenactment on highly compressed videos compared to supervised approaches (Cozzolino et al., 2021).

Dong et al. identified a different problem: “Implicit Identity Leakage”, where binary classifiers unexpectedly learned identity representations rather than forgery artifacts, hindering generalization. Their ID-unaware Deepfake Detection Model with Artifact

Detection Module addressed this issue, achieving 99.70% AUC on FaceForensics++ with ResNet-34 and 93.88% on Celeb-DF with EfficientNet-b4 (Dong et al., 2023).

Rather than analyzing faces in isolation, Nirkin et al. detected face swapping by identifying discrepancies between faces and their context. Their method employed a face identification network analyzing the tightly segmented face region, alongside a context recognition network examining surrounding features such as hair, ears, and the neck. This approach achieved state-of-the-art results on FaceForensics++ and Celeb-DF-v2 benchmarks (Nirkin et al., 2022).

2.2.7 Artifact-Based and Self-Consistency Methods

Several approaches have focused on detecting manipulation artifacts. Zhao et al. proposed pair-wise self-consistency learning (PCL), which detects deepfakes by measuring the inconsistency of source features within forged images. Their method introduced an inconsistency image generator (I2G) for creating training data, improving the average AUC from 96.45% to 98.05% in within-dataset evaluation and from 86.03% to 92.18% in cross-dataset evaluation (T. Zhao et al., 2021).

Li et al. tackled the challenge of separating meaningful artifacts from noise with the Artifacts-Disentangled Adversarial Learning (ADAL) framework. Their Multi-scale Feature Separator precisely transmitted artifact features, while Artifacts Cycle Consistency Loss enabled pixel-level supervised training. ADAL achieved 97.71% accuracy and 99.51% AUC on FaceForensics++, though performance dropped to 79.87% accuracy and 84.62% AUC on the more challenging Celeb-DFv2 (X. Li et al., 2023).

Shiohara et al. took a novel approach with self-blended images (SBIs), synthetic training data generated by blending pseudo-source and target images from a single pristine image. This technique reproduced common forgery artifacts, such as blending boundaries, without requiring actual forged photos for training. The results were promising, outperforming baselines by 4.90% on DFDC and 11.78% on DFDCP in cross-dataset evaluation (Shiohara & Yamasaki, 2022).

A different strategy came from Li et al., who proposed forensic symmetry using a multi-stream learning architecture with two feature extractors capturing symmetry and

similarity features from face patches. Their approach achieved 95.99% accuracy at the image level and 99.43% at the video level for DF-TIMIT, with a maximum AUC of 99.44% at the image level (G. Li et al., 2023).

2.2.8 Frame Inference and Prediction Methods

Hu et al. developed FInfer, a frame-inference-based detection framework designed for high-quality deepfakes. The approach learned to reference representations of current and future frames, using an autoregressive model to predict upcoming facial representations from current ones. The key insight being that real videos show higher correlation between predicted and actual frames than fake ones. FInfer achieved 90.47% accuracy and 93.30% AUC on Celeb-DF (Hu, Liao, Liang, et al., 2022).

2.2.9 Adversarial Training and Robustness Methods

Adversarial training has shown promise for improving model generalization to unseen manipulations. Wang et al. used additive, spatial-transformed, and blurring-based adversarial examples to strengthen detection methods. Their approach with two generators (Two-Gen-BAT) improved EfficientNet accuracy from 81.35% to 84.10% and Xception accuracy from 96.77% to 97.45% on FaceForensics++. More importantly, cross-dataset performance saw significant gains, Xception jumped from 54.88% to 64.85% on Celeb-DF (Z. Wang et al., 2022).

2.2.10 Color Space and Frequency Domain Methods

Mo et al. explored a different angle, analyzing differences in color space components to improve discrimination rates. By combining color-space channel recombinations with a channel attention mechanism, their Xception-based model achieved up to 99.10% accuracy on the same face generation task. Notably, the model maintained 98.71% accuracy even with JPEG compression factor of 100, demonstrating strong robustness (Mo et al., 2022).

2.2.11 Continual and Reinforcement Learning Approaches

As new deepfake methods emerge, detectors need to adapt without forgetting previous ones. Li et al. (2023) addressed this through continual learning, evaluating XceptionNet, ResNet-50, and various incremental learning strategies, including NSCIL, LRCIL, iCaRL, and LUCIR. Vision Transformer-based methods like DyTox achieved the best results, around 86% average accuracy with a memory budget of 100, outperforming CNN-based methods (C. Li et al., 2023).

Nadimpalli et al. took an unconventional approach, formulating deepfake detection as a hybrid of supervised learning and reinforcement learning. Their RL agent selected optimal augmentations for each test sample individually, with classification scores averaged to determine the final result. This approach achieved 0.952 AUC on DeeperForensics-1.0 and 0.669 on Celeb-DF in cross-dataset evaluation (Nadimpalli & Rattani, 2022).

2.2.12 Dual-Level and Multi-Task Frameworks

Pu et al. proposed a dual-level collaborative framework that tackles frame-level and video-level forgeries simultaneously, using a joint loss function optimizing both the AUC score and error rate. The key advantage of this multitask structure is that frame-level and video-level detection reinforce each other. Their AUC-based loss function also handled imbalanced data better than focal loss, resulting in improved robustness to video quality variations and stronger cross-dataset generalization (Pu et al., 2022).

3 Synthesis of findings

3.1 Performance Across Different Approaches

Detection performance varied substantially across methods and evaluation scenarios. Within-dataset evaluation typically yielded excellent results, with many CNN-based approaches achieving over 95% accuracy on FaceForensics++ (H. Zhao et al., 2021). However, methods optimized for specific datasets often experienced dramatic performance degradation when tested on unseen data (Cozzolino et al., 2021).

Factors that explain these differences in performance can be identified as:

- Dataset quality and manipulation diversity
- Compression and video quality effects
- Temporal vs. spatial analysis trade-offs
- Artifact-specific vs. generic feature learning

Studies using high-quality datasets like Celeb-DF, which contains more realistic deepfakes, showed lower detection rates compared to studies that used FaceForensics++ (Nadimpalli & Rattani, 2022). Wang et al. found that when trained on FaceForensics++ and tested on Celeb-DF, Xception baseline performance dropped to 54.88%, though adversarial training improved this to 64.85% (Z. Wang et al., 2022). This pattern is more indicative of the quality gap between training and testing data rather than inherent method limitations.

Detection accuracy suffered significantly with increasing compression. Caldelli et al. demonstrated that optical flow methods maintained 97% accuracy on uncompressed videos (C0) but dropped to 91% for C23 and 76% for C40 compression levels (Caldelli et al., 2021). Similarly, Zhao et al. noted their multi-attentional network achieved 97.60% accuracy on high-quality FaceForensics++ but only 88.69% on low-quality versions (H. Zhao et al., 2021). Physiological measurement approaches proved particularly robust to compression, with DeepFakesON-Phys maintaining over 98% AUC even on compressed data (Hernandez-Ortega et al., 2022).

Methods emphasizing temporal consistency generally showed better cross-manipulation generalization but required more computational resources. Chintha et al.'s

recurrent approach achieved perfect 100% accuracy on FaceForensics++ (Chintha et al., 2020), while optical flow-based methods demonstrated superior cross-forgery robustness (Caldelli et al., 2021). However, single-frame methods like the multi-attentional network traded some temporal robustness for computational efficiency while still achieving competitive performance (H. Zhao et al., 2021).

Methods learning manipulation-specific artifacts often excelled within their training domain but struggled with novel forgery techniques. In contrast, approaches like self-blended images that learned generic forgery patterns showed improved cross-dataset generalization, outperforming baselines by 4.90% on DFDC and 11.78% on DFDCP (Shiohara & Yamasaki, 2022). Similarly, identity-based methods like ID-Reveal achieved more than 15% improvement on compressed videos by focusing on identity-level features rather than manipulation-specific artifacts (Cozzolino et al., 2021).

3.2 Method Specific Strengths and Limitations

3.2.1 CNN-Based Methods

CNN-based methods form the backbone of many detection systems; however, they do display some fundamental limitations. Wang et al. noted that CNNs' overreliance on local texture information hindered generalization to unseen data (P. Wang et al., 2022). This limitation manifested as the "Implicit Identity Leakage" phenomenon identified by Dong et al., in which binary classifiers unexpectedly learned identity representations rather than forgery artifacts. The ID-unaware approach addressing this issue improved cross-dataset AUC from 91.15% to 93.88% on Celeb-DF (Dong et al., 2023).

3.2.2 Transformer Approaches

Transformer-based methods showed promise for improved generalization by modeling both global and local information (P. Wang et al., 2022). Khan et al.'s hybrid transformer achieved 95% accuracy across multiple FaceForensics++ subsets (Khan & Dang-Nguyen, 2022). For audio deepfakes, transformers like AST and Wav2Vec substantially

outperformed traditional methods, though they lacked the interpretability of hand-crafted feature approaches (Channing et al., 2024).

3.2.3 Multimodal Methods

Audio-visual approaches leveraged cross-modal consistency but did not always surpass single-modality detection (Muppalla et al., 2023). Cheng et al.'s voice-face matching achieved 86.11% AUC on FakeAVCeleb, representing a notable improvement over vision-only baselines. However, these methods faced challenges when both modalities were manipulated simultaneously (Cheng et al., 2023). The effectiveness of multimodal approaches depended critically on the availability of synchronized, high-quality audio-visual data.

3.2.4 Physiological Signals

Biological signal-based detection exploits features that are difficult for current deepfake techniques to replicate (Juefei-Xu et al., 2022). Hernandez-Ortega et al.'s rPPG-based method achieved over 98% AUC, demonstrating the power of physiological cues (Hernandez-Ortega et al., 2022). However, these approaches required careful feature extraction and could be sensitive to video quality and lighting conditions.

4 Conclusion

4.1 Summary of findings

This review covered detection approaches ranging from CNN-based classifiers to physiological signal analysis. Most of these methods work well in controlled settings but applying them to real-world content is a different matter.

Within-dataset evaluation consistently achieves high detection accuracy, with many methods often surpassing 95% accuracy on standard benchmarks, such as FaceForensics++. However, cross-dataset generalization remains an issue, as methods trained on one dataset frequently experience dramatic performance degradation when tested on unseen data. In other words, many detectors learn to recognize artifacts from specific datasets or manipulation methods. They fail to learn general signs of tampering.

Another persistent challenge is presented by video compression. Detection accuracy suffers greatly as compression increases, with some methods losing over 20 percentage points between uncompressed and heavily compressed video. This poses problems for real-world deployments of these approaches, as many deepfakes are distributed via social media platforms that apply compression to shared content to reduce the content's data footprint. Malicious actors can also use this to their advantage and reduce the video quality on purpose to evade deepfake detection systems.

Physiological approaches, especially remote photoplethysmography (rPPG), handled compression well. These methods detect subtle signals like heartbeat patterns in skin color. Current deepfake generation technologies cannot replicate these signals reliably. Identity-based methods that focus on behavioral consistency instead of manipulation artifacts showed improved performance on compressed video. Approaches using synthetic training data designed to capture generic manipulation patterns, such as self-blended images, showed better cross-dataset generalization than methods trained on specific manipulation types.

Methods that extend beyond single-frame spatial analysis each address part of the problem but introduce their own tradeoffs. Temporal analysis methods traded high computational cost for better cross-manipulation performance when compared to frame-

level approaches. Transformers can capture both global and local detail, which may help them avoid the texture-dependence problem that limits CNNs. Multimodal approaches leverage anomalies in audio-visual consistency and are effective in situations where only one modality is manipulated but are less effective when both are faked simultaneously. A single approach is rarely enough to cover all scenarios where detection takes place.

4.2 Recommendations for future research

Current literature leaves several gaps that require further investigation. Poor cross-dataset generalization limits practical applications. This needs more attention in future work. Evaluation should be done using only previously unseen manipulation types rather than held-out samples from the same distribution. Diffusion-based generation is a newer problem. Most current detectors were built to catch GAN or autoencoder outputs, and it is unclear how well they handle diffusion-generated content.

Real-world deployment considerations receive only minor attention in current research. Most studies evaluate on curated datasets under controlled conditions, which leaves questions regarding computational efficiency, latency requirements and integration into current content moderation pipelines unexplored. There is also a practical need for methods that can run locally on a phone or laptop, without needing to send every video to a cloud server for analysis.

On a broader level, generation and detection methods are locked in an arms race. As detection methods become more sophisticated, the generation methods evolve to evade them. A long-term solution would likely require content authenticity verification at the point of creation, in addition to the detection methods discussed in this paper.

Finally, the interpretability of detection decisions remains underdeveloped. Most current methods function as black boxes, providing binary classifications without explanations on how they reached that conclusion. This would enhance trust in these methods in real-world use, such as content moderation or legal proceedings.

As deepfakes continue to evolve, detection methods must develop at a similar pace. If detection does not keep up, the credibility of all digital media is at risk, and with it, the public’s trust in what they see or hear online.

References

Almestekawy, A., Zayed, H. H., & Taha, A. (2024). Deepfake detection: Enhancing performance with spatiotemporal texture and deep learning feature fusion. *Egyptian Informatics Journal*, *27*, 100535. https://doi.org/10.1016/j.eij.2024.100535

Ayodele R. Akinyele, Frederick Ogunseye, Adewale Asimolowo, Geoffrey Munyaneza, Oluwatosin Mudele, & Oluwole Olakunle Ajayi. (2024). Advancements in diffusion models for high-resolution image and short form video generation. *GSC Advanced Research and Reviews*, *21*(2), 508–520. https://doi.org/10.30574/gscarr.2024.21.2.0441

Caldelli, R., Galteri, L., Amerini, I., & Del Bimbo, A. (2021). Optical Flow based CNN for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters*, *146*, 31–37. https://doi.org/10.1016/j.patrec.2021.03.005

Channing, G., Sock, J., Clark, R., Torr, P., & Witt, C. S. de. (2024). *Toward Robust Real-World Audio Deepfake Detection: Closing the Explainability Gap* (No. arXiv:2410.07436). arXiv. https://doi.org/10.48550/arXiv.2410.07436

Chen, M., Mei, S., Fan, J., & Wang, M. (2024). Opportunities and challenges of diffusion models for generative AI. *National Science Review*, *11*(12), nwae348. https://doi.org/10.1093/nsr/nwae348

Cheng, H., Guo, Y., Wang, T., Li, Q., Chang, X., & Nie, L. (2023). Voice-Face Homogeneity Tells Deepfake. *ACM Trans. Multimedia Comput. Commun. Appl.*, *20*(3), 76:1-76:22. https://doi.org/10.1145/3625231

Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K., Hickerson, A., Wright, M., & Ptucha, R. (2020). Recurrent Convolutional Structures for Audio Spoof and Video Deepfake

Detection. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 1024–1037. IEEE Journal of Selected Topics in Signal Processing. https://doi.org/10.1109/JSTSP.2020.2999185

Cozzolino, D., Rössler, A., Thies, J., Nießner, M., & Verdoliva, L. (2021). ID-Reveal: Identity-aware DeepFake Video Detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15088–15097. https://doi.org/10.1109/ICCV48922.2021.01483

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). *The DeepFake Detection Challenge (DFDC) Dataset* (No. arXiv:2006.07397). arXiv. https://doi.org/10.48550/arXiv.2006.07397

Dong, S., Wang, J., Ji, R., Liang, J., Fan, H., & Ge, Z. (2023). Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3994–4004. https://doi.org/10.1109/CVPR52729.2023.00389

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* (No. arXiv:1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Hernandez-Ortega, J., Tolosana, R., Fierrez, J., & Morales, A. (2022). DeepFakes Detection Based on Heart Rate Estimation: Single- and Multi-frame. In C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, & C. Busch (Eds.), *Handbook of Digital Face Manipulation and Detection* (pp. 255–273). Springer International Publishing. https://doi.org/10.1007/978-3-030-87664-7_12

Hu, J., Liao, X., Liang, J., Zhou, W., & Qin, Z. (2022). FInfer: Frame Inference-Based Deepfake Detection for High-Visual-Quality Videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(1), Article 1. https://doi.org/10.1609/aaai.v36i1.19978

Hu, J., Liao, X., Wang, W., & Qin, Z. (2022). Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1089–1102. IEEE Transactions on Circuits and Systems for Video Technology. https://doi.org/10.1109/TCSVT.2021.3074259

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2022). Countering Malicious DeepFakes: Survey, Battleground, and Horizon. *International Journal of Computer Vision*, *130*(7), 1678–1734. https://doi.org/10.1007/s11263-022-01606-8

Karras, T., Laine, S., & Aila, T. (2019). *A Style-Based Generator Architecture for Generative Adversarial Networks* (No. arXiv:1812.04948). arXiv. https://doi.org/10.48550/arXiv.1812.04948

Khan, S. A., & Dang-Nguyen, D.-T. (2022). Hybrid Transformer Network for Deepfake Detection. *International Conference on Content-Based Multimedia Indexing*, 8–14. CBMI 2022: International Conference on Content-based Multimedia Indexing. https://doi.org/10.1145/3549555.3549588

Kietzmann, J. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135. https://doi.org/10.1016/j.bushor.2019.11.006

Kingma, D. P., & Welling, M. (2013). *Auto-Encoding Variational Bayes* (No. arXiv:1312.6114). arXiv. https://doi.org/10.48550/arXiv.1312.6114

Li, C., Huang, Z., Paudel, D. P., Wang, Y., Shahbazi, M., Hong, X., & Van Gool, L. (2023). A Continual Deepfake Detection Benchmark: Dataset, Methods, and Essentials. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1339–1349. https://doi.org/10.1109/WACV56688.2023.00139

Li, G., Zhao, X., & Cao, Y. (2023). Forensic Symmetry for DeepFakes. *IEEE Transactions on Information Forensics and Security*, *18*, 1095–1110. IEEE Transactions on Information Forensics and Security. https://doi.org/10.1109/TIFS.2023.3235579

Li, X., Ni, R., Yang, P., Fu, Z., & Zhao, Y. (2023). Artifacts-Disentangled Adversarial Learning for Deepfake Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(4), 1658–1670. IEEE Transactions on Circuits and Systems for Video Technology. https://doi.org/10.1109/TCSVT.2022.3217950

Liu, H., Zhou, W., Chen, D., Fang, H., Bian, H., Liu, K., Zhang, W., & Yu, N. (2023). Coherent adversarial deepfake video generation. *Signal Processing*, *203*, 108790. https://doi.org/10.1016/j.sigpro.2022.108790

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, *53*(4), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2020). *Image Segmentation Using Deep Learning: A Survey* (No. arXiv:2001.05566). arXiv. https://doi.org/10.48550/arXiv.2001.05566

Mirsky, Y., & Lee, W. (2022). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, *54*(1), 1–41. https://doi.org/10.1145/3425780

Mo, S., Lu, P., & Liu, X. (2022). AI-Generated Face Image Identification with Different Color Space Channel Combinations. *Sensors (Basel, Switzerland)*, *22*(21), 8228. https://doi.org/10.3390/s22218228

Muppalla, S., Jia, S., & Lyu, S. (2023). *Integrating Audio-Visual Features for Multimodal Deepfake Detection* (No. arXiv:2310.03827). arXiv. https://doi.org/10.48550/arXiv.2310.03827

Nadimpalli, A. V., & Rattani, A. (2022). *On Improving Cross-dataset Generalization of Deepfake Detectors* (No. arXiv:2204.04285). arXiv. https://doi.org/10.48550/arXiv.2204.04285

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, *223*, 103525. https://doi.org/10.1016/j.cviu.2022.103525

Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2022). DeepFake Detection Based on Discrepancies Between Faces and Their Context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 6111–6121. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2021.3093446

Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C., & Tao, D. (2024). *Deepfake Generation and Detection: A Benchmark and Survey* (No. arXiv:2403.17881; Version 4). arXiv. https://doi.org/10.48550/arXiv.2403.17881

Pu, W., Hu, J., Wang, X., Li, Y., Hu, S., Zhu, B., Song, R., Song, Q., Wu, X., & Lyu, S. (2022). Learning a deep dual-level network for robust DeepFake detection. *Pattern Recognition*, *130*, 108832. https://doi.org/10.1016/j.patcog.2022.108832

Radford, A., Metz, L., & Chintala, S. (2016). *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks* (No. arXiv:1511.06434). arXiv. https://doi.org/10.48550/arXiv.1511.06434

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake Detection: A Systematic Literature Review. *IEEE Access*, *10*, 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404

Raza, A., Munir, K., & Almutairi, M. (2022). A Novel Deep Learning Approach for Deepfake Image Detection. *Applied Sciences*, *12*(19), Article 19. https://doi.org/10.3390/app12199820

Shiohara, K., & Yamasaki, T. (2022). Detecting Deepfakes with Self-Blended Images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18699–18708. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR52688.2022.01816

Singh, R. K., Sarda, P. V., Aggarwal, S., & Vishwakarma, D. K. (2021). Demystifying deepfakes using deep learning. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 1290–1298. https://doi.org/10.1109/ICCMC51019.2021.9418477

Wang, P., Liu, K., Zhou, W., Zhou, H., Liu, H., Zhang, W., & Yu, N. (2022). ADT: Anti-Deepfake Transformer. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2899–2903. https://doi.org/10.1109/ICASSP43922.2022.9746888

Wang, Y., & Huang, H. (2024). Audio–visual deepfake detection using articulatory representation learning. *Computer Vision and Image Understanding*, *248*, 104133. https://doi.org/10.1016/j.cviu.2024.104133

Wang, Z., Guo, Y., & Zuo, W. (2022). Deepfake Forensics via an Adversarial Game. *IEEE Transactions on Image Processing*, *31*, 3541–3552. IEEE Transactions on Image Processing. https://doi.org/10.1109/TIP.2022.3172845

Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A Survey on Deepfake Video Detection. *IET Biometrics*, *10*(6), 607–624. https://doi.org/10.1049/bme2.12031

Zhang, T. (2022). Deepfake generation and detection, a survey. *Multimedia Tools and Applications*, *81*(5), 6259–6276. https://doi.org/10.1007/s11042-021-11733-y

Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D., & Yu, N. (2021). Multi-attentional Deepfake Detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2185–2194. https://doi.org/10.1109/CVPR46437.2021.00222

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., & Xia, W. (2021). Learning Self-Consistency for Deepfake Detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15003–15013. https://doi.org/10.1109/ICCV48922.2021.01475

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks* (No. arXiv:1703.10593). arXiv. https://doi.org/10.48550/arXiv.1703.10593