International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Artificial Intelligence–Based Detection of Deepfake Media on Social Platforms: Evaluating Machine Learning Models for Identifying Manipulated Video Content

Ratan Singh

Abstract

Deepfakes are synthetic or manipulated audio-visual media produced through deep learning techniques that can depict people saying or doing things that did not occur. Their increasing realism has intensified concerns about misinformation, fraud, privacy, political communication, and the evidentiary reliability of digital media. This paper presents a critical narrative review of deepfake detection research, comparing traditional forensic methods, classical machine learning approaches, and contemporary deep learning systems. The review examines spatial detectors based on image artefacts, temporal models that analyse frame-to-frame inconsistency, physiological approaches that use biological signals, and multimodal systems that combine visual and audio evidence. The literature indicates that deep learning methods generally outperform handcrafted forensic techniques on benchmark datasets because they learn complex features directly from data. However, benchmark superiority does not translate automatically into dependable real-world performance. Cross-dataset testing reveals substantial losses in accuracy when detectors encounter unfamiliar manipulation methods, compression, low resolution, demographic variation, or adversarial perturbations. The central conclusion is therefore conditional: deep learning provides the strongest current detection capability, but no single architecture is sufficiently general, explainable, efficient, and robust for universal deployment. Future progress requires diverse and continually updated datasets, cross-dataset evaluation, multimodal fusion, uncertainty-aware decisions, adversarial testing, provenance systems, and governance frameworks that protect privacy and due process. Deepfake detection should be treated as one component of a broader media-authenticity ecosystem rather than as a complete solution to synthetic-media harms.

Keywords: deepfake detection, synthetic media, digital forensics, convolutional neural networks, multimodal detection, generalisation, media authenticity

1. Introduction

1.1 Background of Deepfakes

Deepfakes are a form of synthetic media in which a person's likeness is digitally altered or replaced using advanced techniques from Artificial Intelligence. These manipulations can involve swapping faces in videos, generating entirely new human appearances, or altering speech to make it seem as though someone said something they never actually did. For example, a video might depict a well-known public figure delivering a statement or engaging in actions that never occurred in reality. Such content can appear highly

realistic, making it increasingly difficult for viewers to distinguish between authentic and fabricated media.

At the core of deepfake technology are Generative Adversarial Networks (GANs), a class of machine learning models designed to generate new data that closely resembles real-world inputs. GANs consist of two neural networks—a generator and a discriminator—that are trained simultaneously in a competitive process. The generator attempts to create convincing fake data, while the discriminator evaluates its authenticity. Over time, this adversarial process enables the system to produce highly realistic images, videos, and audio.

The foundation for deepfake development was laid long before the term itself was coined. Researchers in fields such as computer vision and image processing had been exploring ways to analyze, reconstruct, and manipulate visual data for decades. A major breakthrough came in 2014 when Ian Goodfellow introduced GANs, providing a powerful framework for automated content generation. This innovation significantly accelerated the development of synthetic media technologies.

The term “deepfake” first emerged in 2017 on the social media platform Reddit, when a user posted AI-generated face-swapped videos. These early examples quickly gained attention, highlighting both the impressive capabilities and the potential risks of the technology. In the following years, particularly throughout the 2020s, deepfake tools became more advanced, accessible, and user-friendly. Today, they can be created using widely available software, raising important questions about ethics, misinformation, privacy, and the future of digital trust.

1.2 Deepfakes as a Social and Economic Issue

Deepfakes have emerged as a significant socio-economic issue due to their potential to disrupt trust, manipulate markets, and exacerbate inequality in the digital economy . By manipulating Artificial Intelligence systems highly realistic audio and video content can be generated making it difficult to distinguish between truth and deception . This erosion of trust has serious implications for financial systems, where fake announcements or impersonations can influence stock prices and corporate reputations. Additionally, deepfakes disproportionately impact individuals with less access to legal or technological resources, such as victims of non-consensual synthetic media, thereby widening social inequalities.

A 2025 analysis of Artificial Intelligence (AI) scams in India reports that 47 percent of Indian adults have either been victims of, or know someone who has been a victim of, an AI voice-cloning or deepfake scam, nearly double the global average of 25 percent. The same report notes that 83 percent of Indian victims of AI voice scams suffered monetary loss, with almost half losing over INR50,000, highlighting the rapid growth of deepfake-enabled fraud in the Indian financial ecosystem. Within this rapidly evolving threat landscape, one of the most concerning developments for financial institutions is the emergence and proliferation of deepfakes as tools for cyber-enabled financial crime. (Observe Research Foundation) . Deepfakes have been used in many real life examples such as the ARUP ENGINEERING SCAM where an employee was duped into transferring 25 million dollars after attending a call with deepfakes versions of senior executives . Deepfake videos of ELON MUSK promoting fraudulent crypto schemes circulated widely on social media and many people were tricked into investing and they lost millions of dollars . Victims described the impersonations as “ indistinguishable “ from the real musk.

Not only do these scenarios taint the reputation of celebrities they also aid to the erosion of trust in online content which has become a critical socio-economic concern in the digital age, driven by the rapid spread of misinformation, manipulated media, and technologies such as deepfakes. As individuals increasingly

rely on digital platforms for news, communication, and financial decisions, the inability to verify the authenticity of content undermines confidence in institutions, media organizations, and even interpersonal interactions. This growing skepticism can have economic consequences, including market instability caused by false information and increased costs for businesses that must invest in verification and cybersecurity measures. Socially, it contributes to polarization, as people gravitate toward information that aligns with their beliefs rather than verified facts. Ultimately, the decline in trust weakens the foundation of informed decision-making and poses long-term risks to both democratic processes and economic stability.

1.3 Problem Statement

The rapid evolution of deepfake technology has introduced serious challenges for researchers, policymakers, and digital platforms. Advances in deep learning, particularly Generative Adversarial Networks (GANs) and other generative models, have enabled the creation of highly realistic synthetic media that is increasingly difficult to distinguish from authentic content. These technologies are improving at a pace that often exceeds the development of reliable detection methods, creating a gap between generation and detection capabilities (Hasan et al., 2026, arXiv).

A key issue lies in the growing realism of deepfakes. Modern generative models can produce images and videos with minimal visible artifacts, making them difficult for both humans and automated systems to identify. Detection models often struggle because the inconsistencies they rely on are subtle and inconsistently distributed across the content (Gong & Li, 2024, MDPI Electronics). As a result, traditional detection approaches are becoming less effective over time.

Another major concern is the lack of generalization in existing detection systems. Many models perform well under controlled conditions but fail when applied to new datasets or unseen manipulation techniques. This limitation significantly reduces their effectiveness in real-world scenarios, where deepfakes are created using a wide range of tools and settings (Guarnera et al., 2022, Journal of Imaging). Variations in resolution, compression, and editing further complicate the detection process.

In addition, the accessibility of deepfake creation tools has increased significantly. Open-source software and user-friendly applications have made it possible for individuals with limited technical knowledge to produce convincing fake content. This widespread availability accelerates the creation and dissemination of deepfakes, while scalable detection solutions remain limited (Patel et al., 2025, IJSRCSEIT).

Furthermore, many current detection techniques rely on identifying visual or statistical anomalies such as inconsistencies in lighting, facial expressions, or motion. However, as generative models continue to improve, these anomalies are becoming less noticeable or are eliminated entirely, reducing the effectiveness of such approaches (Springer Review, 2025, Springer). This leads to an ongoing arms race between deepfake generation and detection technologies.

In conclusion, the problem is defined by three key challenges: the rapid advancement and realism of deepfake generation, the limited robustness and generalization of detection models, and the lack of scalable solutions suitable for real-world deployment. Addressing these challenges is essential to ensure the integrity of digital media and to mitigate the risks associated with deepfake technology.

1.4 Research Objectives

The primary objective of this study is to systematically examine existing approaches to deepfake detection, with a particular focus on the comparative effectiveness of traditional and deep learning–based techniques. The research aims to analyse a range of detection methodologies, including feature-based approaches and data-driven models, in order to assess their strengths and limitations in identifying manipulated media.

Furthermore, this study seeks to evaluate the performance of machine learning and deep learning models such as convolutional neural networks and recurrent neural networks—in detecting deepfake content across diverse datasets and real-world scenarios. Prior research has demonstrated that deep learning models can achieve high detection accuracy when trained on large-scale datasets, though their generalizability remains a concern ["DeepFake Detection Using Deep Learning: A Survey", IEEE Access].

In addition, the research intends to identify existing gaps in current detection frameworks, particularly in terms of scalability, robustness against adversarial techniques, and adaptability to rapidly evolving deepfake generation methods. Studies have highlighted that many detection systems struggle to maintain performance when exposed to unseen manipulation techniques ["FaceForensics++: Learning to Detect Manipulated Facial Images", ICCV].

Finally, the study aims to propose potential directions for future improvement, including the integration of hybrid detection models and the development of more generalized detection systems capable of operating effectively across multiple platforms and media formats ["A Survey on Deepfake Detection Methods", ACM Computing surveys "]

1.5 Research Question

The central research question guiding this study is: To what extent do deep learning–based detection models outperform traditional detection techniques in accurately identifying deepfake video content on social media platforms?

2. Literature Review

The study of deepfake technology has evolved rapidly in recent years, driven by advancements in artificial intelligence and increased computational capabilities. Early research primarily focused on understanding the mechanisms behind synthetic media generation, while more recent work has shifted toward the development of robust detection techniques. The growing body of literature reflects a dual progression: the improvement of deepfake generation models and the parallel advancement of detection frameworks designed to counter them.

Scholars have examined the transition from conventional media manipulation techniques to AI-driven synthesis, emphasizing the role of generative adversarial networks and deep neural architectures in producing highly realistic content ["DeepFake Detection Using Deep Learning: A Survey", IEEE Access]. These developments have significantly reduced the presence of visual artifacts that were once commonly used as indicators of tampering, thereby complicating detection efforts.

In addition, several studies have explored the limitations of traditional detection methods, which often rely on handcrafted features and domain-specific knowledge. Such approaches, while effective against earlier forms of manipulation, have shown limited adaptability to modern deepfake techniques that continuously evolve in response to detection strategies ["A Survey on Deepfake Detection Methods", ACM Computing Surveys].

Recent literature has increasingly focused on machine learning and deep learning–based detection models, highlighting their ability to learn discriminative features directly from data. These models, particularly convolutional neural networks and recurrent architectures, have demonstrated promising results in identifying both spatial and temporal inconsistencies in manipulated videos ["FaceForensics++: Learning to Detect Manipulated Facial Images", ICCV]. However, concerns remain regarding their generalization capability, computational requirements, and vulnerability to adversarial manipulation.

Overall, the literature indicates that while significant progress has been made in both the creation and detection of deepfakes, the field remains highly dynamic. Continuous advancements in generation techniques necessitate equally adaptive and scalable detection solutions, underscoring the importance of ongoing research in this domain.

2.1 Evolution of Deepfake Generation Techniques

The development of deepfake generation techniques has undergone a significant transformation, progressing from basic digital editing methods to highly sophisticated artificial intelligence–driven systems. Early forms of manipulated media relied on traditional video editing tools and computer-generated imagery (CGI), which required substantial manual effort and technical expertise. These methods often produced detectable inconsistencies, limiting their effectiveness and realism.

With the advancement of deep learning, particularly the introduction of generative adversarial networks (GANs), the landscape of synthetic media creation changed considerably. GAN-based architectures enable the generation of highly realistic images and videos by training two neural networks in opposition, allowing for continuous refinement of output quality [“Generative Adversarial Networks”, NeurIPS]. This approach has significantly enhanced the ability to replicate facial features, expressions, and movements with a high degree of accuracy.

Subsequent innovations have further improved the realism and accessibility of deepfake technology. Techniques such as autoencoders and neural rendering have enabled efficient face swapping and facial reenactment, making it possible to manipulate video content with minimal computational resources. As a result, deepfake creation tools have become increasingly accessible to the general public, contributing to their widespread use across various domains, including entertainment and social media [“Deep Learning for Deepfakes Creation and Detection: A Survey”, IEEE Access].

Moreover, continuous improvements in model training, dataset availability, and processing power have led to the reduction of visual artifacts that were once common in earlier deepfake outputs. This evolution has made it increasingly difficult to distinguish between authentic and manipulated content, thereby posing significant challenges for detection systems and raising concerns about the potential misuse of such technologies [“Face2Face: Real-time Face Capture and Reenactment of RGB Videos”, CVPR].

2.2 Traditional Fake Media Generation Methods

Before the emergence of deep learning–based synthesis techniques, manipulated media was primarily created using traditional editing approaches. These methods included manual video editing, image splicing, and computer-generated imagery (CGI), which required skilled operators and significant time investment. Such techniques were widely used in film production and visual effects industries, where realism was achieved through controlled environments and post-production refinement.

Traditional manipulation methods typically involved frame-by-frame editing or compositing multiple visual elements to create altered content. While effective in controlled settings, these approaches often introduced noticeable inconsistencies in lighting, shadows, texture alignment, and facial continuity. These artifacts made it relatively easier for both human observers and early detection systems to identify tampered content [“Media Forensics and Deep Learning”, IEEE Signal Processing Magazine].

In comparison to modern AI-based generation techniques, traditional methods lacked the ability to produce fully automated and highly realistic outputs. The absence of learning-based models meant that these systems could not adapt or improve based on data, limiting their scalability and sophistication. As a result, most outputs remained visually distinguishable from authentic media, particularly under close inspection.

Furthermore, the reliance on manual intervention meant that traditional fake media generation was not easily scalable for mass production or real-time applications. This limitation significantly reduced its potential for widespread misuse compared to contemporary deepfake technologies, which can generate high-quality synthetic media in a matter of seconds using trained neural networks ["A Survey on Image and Video Forgery Detection", ACM Computing Surveys].

2.3 Deep Learning-Based Deepfake Creation

The introduction of deep learning has fundamentally transformed the process of synthetic media generation, enabling the creation of highly realistic and increasingly indistinguishable manipulated content. Unlike traditional editing techniques, deep learning–based approaches rely on large-scale data-driven models that learn complex patterns in facial structure, movement, and expression.

Generative adversarial networks (GANs) are among the most widely used architectures in deepfake creation. In this framework, a generator network produces synthetic outputs while a discriminator network evaluates their authenticity, resulting in iterative improvements in realism ["Generative Adversarial Nets", NeurIPS]. This adversarial process allows the system to progressively refine facial details, making generated images and videos increasingly convincing.

In addition to GANs, autoencoders have played a significant role in face-swapping and identity manipulation tasks. These models compress facial features into latent representations and reconstruct them in different contexts, enabling the transfer of facial characteristics between individuals. Variants of autoencoder-based systems have been widely applied in real-time face replacement applications ["DeepFakes: A New Threat to Face Recognition?", arXiv].

Recent advancements also include the integration of neural networks for voice cloning and multimodal synthesis, which combine audio and visual manipulation to produce more coherent and realistic outputs. These developments have contributed to the rapid expansion of deepfake capabilities, making it possible to generate synchronized facial expressions and speech with minimal human intervention ["Speech Synthesis and Voice Cloning Using Deep Learning", IEEE Access].

As a result of these advancements, distinguishing between real and synthetic content has become increasingly difficult. Deep learning models are capable of replicating subtle facial micro-expressions and temporal dynamics, significantly reducing the effectiveness of traditional detection methods and increasing the overall realism of generated media.

2.4 Why Are Deepfakes Hard to Detect ?

Deepfakes have become increasingly difficult to detect because modern artificial intelligence models are now capable of generating highly realistic images, videos, and audio with very few visible errors. Earlier forms of manipulated media often contained obvious visual inconsistencies such as unnatural facial movements, poor lighting alignment, blurred edges, or irregular blinking patterns. However, recent deep learning techniques, especially Generative Adversarial Networks (GANs), have significantly reduced these artifacts and improved the overall realism of synthetic media (Nguyen et al., 2022, Monash University Research Repository).

One major reason deepfakes are challenging to identify is their ability to accurately imitate human facial expressions and micro-details. Advanced neural networks can now replicate eye movement, skin texture, lip synchronization, and subtle emotional expressions with impressive precision. As a result, both humans and automated systems often struggle to distinguish authentic media from manipulated content (Korshunov & Marcel, 2020, arXiv). Studies have shown that some deepfake videos appear so realistic that even trained observers fail to identify them consistently.

Another important challenge is the continuous evolution of deepfake generation models. Detection systems are generally trained to recognize known manipulation patterns or artifacts. However, creators of deepfakes constantly improve their techniques to remove these detectable traces. This creates an "arms race" between deepfake generation and detection technologies, where detection models quickly become outdated when exposed to newer and more advanced fake content (Le et al., 2023, arXiv).

Furthermore, modern deepfakes can bypass many traditional forensic detection methods because they adapt to the weaknesses of existing systems. Some generation models intentionally reduce compression artifacts, improve frame consistency, and enhance temporal smoothness in videos, making detection even harder. Researchers have noted that detection accuracy often decreases significantly when models are tested on previously unseen deepfake datasets or low-quality social media videos (Wired, 2020, Wired Magazine).

In addition, social media platforms contribute to the detection challenge. Videos uploaded online are frequently compressed, resized, or filtered, which removes many of the subtle indicators that detection systems rely on. This degradation reduces the effectiveness of feature-based and machine learning detection techniques. Consequently, researchers are now focusing on more robust approaches that combine spatial, temporal, biological, and audio-visual signals for improved accuracy (Saif & Tehseen, 2021, Journal of Information and Communication Technology).

3. Deepfake Detection Techniques

Deepfake detection techniques have evolved rapidly in response to the increasing sophistication of AI-generated media. As deepfake videos become more realistic and difficult to identify with the human eye, researchers have developed various detection methods that analyse visual, audio, and temporal inconsistencies within manipulated content. These techniques generally fall into three major categories: traditional detection methods, machine learning–based approaches, and deep learning–based detection systems.

The primary goal of deepfake detection is to distinguish authentic media from manipulated content by identifying patterns or abnormalities introduced during the generation process. Early detection approaches focused mainly on visible artifacts such as unnatural facial expressions, inconsistent lighting, distorted shadows, or irregular blinking patterns. However, modern deepfakes generated using advanced neural networks and GANs have significantly reduced these detectable flaws, making traditional approaches less effective (Verdoliva, 2020, IEEE Journal).

To address these limitations, researchers introduced machine learning techniques capable of learning manipulation patterns from large datasets. These systems use classification algorithms to differentiate between real and fake media based on extracted features. Although machine learning methods improved detection accuracy, their effectiveness still depends heavily on dataset quality and feature engineering (Dolhansky et al., 2020, Facebook AI Research).

More recently, deep learning–based detection models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have become the dominant approach in deepfake detection research. These models automatically learn spatial and temporal features from videos, enabling them to identify subtle inconsistencies that may not be visible to humans. Deep learning approaches have achieved high detection accuracy in many benchmark datasets, but they also require significant computational resources and large amounts of training data (Nguyen et al., 2022, Monash University Research Repository).

Despite these advancements, no single detection technique is completely reliable against all types of deepfakes. As generation technologies continue to evolve, researchers must continuously improve detection systems to handle new manipulation methods and real-world social media content. The following sections discuss the major deepfake detection techniques, their working principles, strengths, and limitations in greater detail.

3.1 Traditional Detection Methods

Traditional deepfake detection methods were among the earliest approaches developed to identify manipulated media before the widespread adoption of advanced deep learning techniques. These methods mainly focus on detecting visible inconsistencies and digital artifacts that appear during the editing or generation process. Unlike modern AI-based detection systems, traditional techniques rely heavily on manually engineered features and forensic analysis to determine whether a video or image has been altered.

One common traditional detection approach involves analysing visual inconsistencies within facial regions. Early deepfake videos often contained unnatural eye blinking, irregular lip synchronization, distorted facial boundaries, or inconsistent head movements. Researchers discovered that many AI-generated videos failed to accurately reproduce normal human blinking patterns, making blinking frequency an important detection feature (Li et al., 2018, arXiv). Similarly, mismatches between facial expressions and speech movements were also used to identify manipulated videos.

Another widely used method is image forensics analysis, which examines lighting, shadows, reflections, and texture inconsistencies in digital media. In authentic videos, lighting conditions generally remain consistent across the face and surrounding environment. However, deepfake generation models sometimes produce unrealistic shadows or mismatched illumination that can reveal manipulation (Verdoliva, 2020, IEEE Journal). Texture analysis can also identify abnormalities in skin smoothness, color blending, or image compression artifacts.

Traditional detection methods also include frequency-domain analysis, where researchers examine pixel distributions and hidden noise patterns within images and videos. Since manipulated media often undergoes multiple processing stages, it may contain irregular statistical patterns that differ from authentic recordings. These methods attempt to detect changes introduced during face swapping or image synthesis (Fridrich & Kodovsky, 2012, IEEE Transactions on Information Forensics and Security).

Although traditional methods were initially effective against low-quality deepfakes, they have several limitations when dealing with modern AI-generated content. Advanced deepfake models now produce highly realistic outputs with fewer visible artifacts, reducing the effectiveness of feature-based forensic analysis. In addition, traditional methods often struggle to adapt to new manipulation techniques because they depend on predefined rules and manually selected features (Nguyen et al., 2022, Monash University Research Repository).

Another major limitation is that traditional detection systems are sensitive to video quality and compression. Social media platforms frequently compress uploaded videos, which can remove important forensic indicators and reduce detection accuracy. Variations in lighting, camera angles, and facial expressions may also affect the reliability of these methods in real-world environments.

Despite these weaknesses, traditional detection techniques still play an important role in digital media forensics. They are often used alongside machine learning and deep learning systems to improve overall detection performance. In many cases, combining forensic analysis with AI-based models provides more accurate and reliable results for identifying manipulated media.

3.2 Machine Learning–Based Detection

Machine learning–based detection methods were developed to overcome some of the limitations of traditional forensic approaches in identifying deepfake media. Unlike traditional methods that rely mainly on manually observing visual inconsistencies, machine learning models learn patterns and characteristics of manipulated content from large datasets. These systems are trained using algorithms that classify media as either real or fake based on extracted features and statistical patterns.

One of the main advantages of machine learning–based detection is its ability to recognise hidden manipulation patterns that may not be visible to human observers. Researchers commonly use features such as facial landmarks, texture irregularities, blinking frequency, head movement, and image noise patterns as inputs for machine learning classifiers (Agarwal et al., 2020, MIT Media Lab). Algorithms including Support Vector Machines (SVM), Random Forest, Decision Trees, and Logistic Regression have been widely applied in early deepfake detection research.

Feature extraction plays an important role in machine learning–based systems. Before classification, important information is extracted from images or video frames to help the model identify manipulated content. For example, facial feature analysis may examine inconsistencies in eye movement, mouth alignment, or skin texture. Some studies also use frequency-domain analysis and color distribution patterns to improve classification accuracy (Verdoliva, 2020, IEEE Journal).

Machine learning models are generally trained using large datasets containing both authentic and manipulated videos. During training, the algorithm learns differences between real and fake media and later applies this knowledge to classify unknown content. The performance of these systems depends heavily on the quality, diversity, and size of the dataset used for training. Well-balanced datasets with different ethnicities, lighting conditions, camera angles, and video qualities usually produce more reliable detection results (Dolhansky et al., 2020, Facebook AI Research).

Compared to traditional forensic methods, machine learning approaches provide improved adaptability and automation. They can process large amounts of data efficiently and identify subtle manipulation patterns with higher accuracy. These systems also reduce the need for manual feature analysis, making them more suitable for large-scale social media monitoring and automated content verification.

However, machine learning–based detection methods still face several challenges. One major limitation is their dependence on handcrafted feature extraction. If important features are not selected correctly, the model's performance may decrease significantly. In addition, many machine learning models struggle when exposed to new types of deepfakes that differ from the training dataset (Nguyen et al., 2022, Monash University Research Repository).

Another issue is overfitting, where a model performs well on training data but poorly on real-world videos. Social media compression, low-resolution uploads, and video editing filters can further reduce detection accuracy. Researchers have also found that machine learning models are vulnerable to adversarial attacks designed to intentionally fool detection systems (Goodfellow et al., 2015, arXiv).

Despite these limitations, machine learning–based detection methods remain an important stage in the evolution of deepfake detection research. They provide a foundation for more advanced deep learning approaches and continue to be used in combination with other techniques to improve detection reliability and efficiency.

4. Modalities and Detection Frameworks

4.1 Deepfake Image vs Video Detection

The detection of deepfake content differs substantially depending on whether the target media is a static image or a video sequence. In image-based detection, the primary focus falls on pixel-level inconsistencies that arise during the synthesis process, including unnatural noise distributions, blending artefacts along face boundaries, and irregular skin texture patterns that differ from authentic photographs (Tolosana et al., 2020). These inconsistencies, while often imperceptible to the human eye, can be identified by convolutional neural networks trained to distinguish the statistical fingerprints of generated imagery from those of real photographs. Because images present a single spatial frame, detection models can be applied in a relatively straightforward manner without needing to account for motion or temporal coherence.

Video-based detection introduces a substantially higher degree of complexity. Beyond the spatial analysis required for still images, video detection must account for temporal consistency across frames, examining whether facial movements, blinking patterns, and expression transitions remain coherent over time (Guera & Delp, 2018). Authentic videos contain natural micro-movements and physiological signals that deepfake generation pipelines frequently fail to replicate accurately. Temporal analysis using recurrent architectures or 3D convolutional networks can exploit these inconsistencies, but the computational demands are considerably greater, and the variety of deepfake generation techniques makes developing a universally effective video detector a persistent challenge.

4.2 Video and Image Modality Fusion

A promising direction in deepfake detection research involves combining spatial and temporal data streams into unified multi-modal frameworks. Modality fusion approaches recognise that neither pixel-level spatial analysis nor temporal sequence modelling is sufficient on its own to detect the full range of deepfake manipulation techniques. By combining these two streams, fusion architectures can exploit complementary cues that are unavailable to single-modality detectors (Zhao et al., 2021). For example, a model might process individual frames through a spatial branch while simultaneously analysing optical flow or frame-difference signals through a temporal branch, combining the outputs at a decision or feature level.

Hybrid models built on this principle have demonstrated measurable accuracy improvements over single-modality baselines in several benchmark evaluations. Multi-modal architectures that incorporate both appearance and motion features have proven particularly effective at detecting face-swap deepfakes, where the inserted face may appear visually plausible in any single frame but exhibits subtle temporal discontinuities across the sequence. The integration of audio modality presents a further opportunity, as mismatches between lip movements and speech signals provide an additional detection channel that is difficult for current deepfake generation systems to eliminate entirely (Mittal et al., 2020).

4.3 Facial Tampering Detection

Facial tampering detection focuses specifically on identifying manipulations applied to the face region of an image or video, which accounts for the vast majority of deepfake content encountered in practice. Landmark-based approaches analyse the geometric arrangement of facial keypoints such as the eyes, nose, and mouth, identifying deformations or alignment inconsistencies that arise when a synthesised face is composited onto a target identity (Li et al., 2020). These geometric irregularities are particularly evident in regions where the synthesised face meets the background or the target subject's neck and hair, areas that current generation algorithms handle with the least precision.

Biometric inconsistency analysis provides a complementary detection pathway by examining physiological signals that are inherently difficult to fabricate. Remote photoplethysmography, which detects subtle changes in skin colour caused by blood flow, has been explored as a deepfake detection signal on the basis that synthesised faces lack genuine circulatory responses (Ciftci, et al., 2020). Similarly, the analysis of unnatural expression dynamics, such as the absence of expected asymmetry in spontaneous facial movements, offers a signal that is grounded in known properties of authentic human facial behaviour rather than in the specific artefacts of any particular generation algorithm, making it more robust to novel generation techniques.

4.4 Conceptual Framework for Detection Systems

A conceptual framework for deepfake detection systems can be described as a sequential pipeline moving from raw media input through preprocessing, feature extraction, and classification to an output decision. In the preprocessing stage, incoming images or video frames are normalised, faces are detected and cropped using standard face detection algorithms, and any necessary resolution adjustments are applied to prepare the input for feature extraction (Rossler et al., 2019). This stage is critical because the quality and consistency of preprocessing directly affects the reliability of downstream analysis.

The feature extraction and classification stages form the core of the detection pipeline. During feature extraction, the preprocessed input is passed through a trained neural network that produces a compact representation encoding the patterns most informative for distinguishing authentic from manipulated media. This representation is then passed to a classification layer that outputs a probability score indicating the likelihood that the input has been manipulated. Integrating such pipelines into scalable systems suitable for deployment on social media platforms requires careful attention to throughput requirements, since platforms must process millions of pieces of content daily, and to the management of false positive rates, given the reputational and legal consequences of incorrectly flagging authentic content as manipulated.

5. Algorithms and Implementation

5.1 Algorithms Used in Deepfake Detection

Convolutional neural networks (CNNs) have been the foundational architecture for deepfake detection since the field emerged as a research priority. Among CNN architectures, ResNet variants have been widely adopted due to their use of residual connections that allow effective training of very deep networks without the vanishing gradient problem, enabling the extraction of high-level semantic features while preserving low-level textural information relevant to detecting synthesis artefacts (He et al., 2016). VGG-based architectures, characterised by their use of small convolutional filters stacked in deep sequences, have similarly been applied as baseline detectors in numerous benchmark studies due to their effectiveness at capturing local texture patterns. EfficientNet has emerged as a particularly strong performer on deepfake detection benchmarks, offering strong accuracy with substantially lower computational cost than older architectures.

Transformer-based models have more recently been applied to deepfake detection, with vision transformers demonstrating competitive or superior performance to CNN baselines on several benchmarks (Wodajo & Atnafu, 2021). The self-attention mechanism in transformers enables the model to capture long-range spatial dependencies across an image, potentially identifying global inconsistencies that local convolutional filters may miss. Hybrid detection techniques that combine convolutional feature extraction with transformer-based attention mechanisms represent the current frontier of the field, attempting to

exploit the complementary strengths of both architectural paradigms while managing the substantial computational cost associated with training large transformer models from scratch.

5.2 Implementation of Detection Systems

Implementing an effective deepfake detection system begins with the assembly and preparation of training data. Systems are typically trained on large benchmark datasets such as FaceForensics++, which contains thousands of manipulated and authentic video sequences produced using multiple generation methods, or the Deepfake Detection Challenge dataset released by Facebook in collaboration with academic partners (Dolhansky et al., 2020). Data augmentation strategies including random cropping, horizontal flipping, colour jitter, and compression artefact simulation are applied during training to improve the robustness of the trained model to the variation encountered in real-world deployment conditions.

Feature extraction and classification are typically implemented as an end-to-end trained pipeline in which the convolutional or transformer backbone and the classification head are optimised jointly on the training data. Model evaluation uses a standard suite of metrics including accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC), with AUC-ROC being particularly informative for deepfake detection tasks because it captures the trade-off between sensitivity and specificity across different decision thresholds (Rossler et al., 2019). Cross-dataset evaluation, where a model trained on one benchmark is tested on a different benchmark, is widely used to assess generalisation capability, which remains one of the central challenges in the field.

5.3 Real-Time Video Processing

Real-time deepfake detection on live video streams presents challenges that are qualitatively different from those encountered in offline analysis. Processing latency must be low enough to support synchronous analysis of video content without introducing perceptible delays, yet the deep learning models that achieve the highest detection accuracy are also among the most computationally expensive (Li & Lyu, 2019). On standard consumer hardware, running a full inference pass of a large detection model on every frame of a high-definition video stream typically exceeds the available computational budget, necessitating strategies such as processing every nth frame, using lighter-weight model architectures, or offloading computation to dedicated hardware accelerators.

Model optimization techniques including quantisation, knowledge distillation, and pruning have been explored as means of reducing the computational cost of detection models without proportionate losses in accuracy. Quantisation reduces the precision of model weights from 32-bit floating point to 8-bit or even lower representations, dramatically reducing memory requirements and enabling faster inference on hardware with integer arithmetic acceleration. Knowledge distillation trains a smaller student model to replicate the outputs of a larger, more accurate teacher model, producing a compact detector that retains much of the teacher's performance at a fraction of the inference cost. These techniques are likely to be essential components of any real-world real-time deepfake detection deployment.

6. Security, Ethical, and Societal Implications

6.1 Security Risks Posed by Deepfakes

Deepfakes pose a significant and growing set of security risks that extend well beyond the domain of media authenticity. Identity theft enabled by deepfake technology allows malicious actors to impersonate specific individuals with a degree of visual fidelity that renders traditional identity verification methods unreliable. Video-based authentication systems, which have been deployed in financial services and government applications, are particularly vulnerable, since a sufficiently convincing deepfake video of a

legitimate account holder could potentially bypass liveness detection protocols (Mirsky & Lee, 2021). The accessibility of deepfake generation tools has lowered the technical barrier to executing such attacks to the point where sophisticated impersonation is no longer restricted to well-resourced adversaries.

Financial fraud represents one of the most immediately consequential applications of deepfake technology in the security domain. Several documented cases have involved the use of AI-generated audio or video to impersonate senior executives and instruct employees or financial institutions to transfer funds, a form of attack sometimes referred to as business email compromise elevated to business video compromise (Chesney & Citron, 2019). Beyond individual financial attacks, the potential use of deepfakes to compromise critical infrastructure security, manipulate stock markets by fabricating statements from corporate leaders, or undermine the integrity of legal proceedings through fabricated video evidence represents a systemic vulnerability that existing legal and technical frameworks were not designed to address.

6.2 Impact on Privacy

The privacy implications of deepfake technology are particularly severe because the harm caused by the non-consensual use of a person's likeness can be immediate, widespread, and extremely difficult to reverse. Non-consensual intimate deepfakes, in which the likeness of a real individual is placed into sexually explicit content without their knowledge or consent, represent the most prevalent and psychologically damaging form of privacy violation enabled by this technology (Henry, et al., 2020). Victims frequently report severe psychological harm, reputational damage, and difficulties in obtaining removal of content once it has been distributed across multiple platforms and jurisdictions. The effectiveness of current platform moderation responses to this form of harm is widely regarded as insufficient.

Beyond the most acute forms of harm, deepfake technology creates longer-term risks associated with the erosion of control individuals have over their own digital representation. The digital footprint created by social media activity, including photographs and videos posted publicly over many years, constitutes the training data from which a convincing deepfake of any individual with a sufficient online presence can be generated. This means that the consent violation inherent in deepfake creation is not limited to the moment of creation but extends retrospectively to every piece of content the individual has shared online, fundamentally altering the privacy calculus of social media participation in ways that users could not reasonably have anticipated when that content was first posted.

6.3 Social and Political Risks

The use of deepfakes to spread political misinformation represents one of the most concerning applications of the technology from a societal perspective. Fabricated video content depicting political figures making statements they never made, or appearing in situations that never occurred, has the potential to influence public opinion and electoral outcomes in ways that are difficult to counter even when the deception is eventually exposed (Chesney & Citron, 2019). Research on misinformation more broadly has demonstrated that initial false impressions persist even after correction, a phenomenon known as the continued influence effect, suggesting that the damage caused by a convincing political deepfake may not be fully reversible even with rapid and effective debunking.

The aggregate societal effect of widespread deepfake circulation may extend beyond the impact of any individual fabricated video. The mere awareness that convincing video fabrications are technically possible and widely circulated creates what researchers have termed the liar's dividend: an environment in which authentic video evidence can be dismissed as potentially fabricated by anyone wishing to avoid

accountability for their real actions (Chesney & Citron, 2019). This erosion of the epistemic status of video as a reliable record of events represents a challenge to journalism, legal proceedings, and democratic discourse that is difficult to address through purely technical means.

6.4 Ethical Considerations

The ethical responsibilities associated with deepfake technology are distributed across developers, platform operators, policymakers, and researchers. Developers of deepfake generation tools bear a primary responsibility for considering the foreseeable harms their technologies enable, and for implementing technical safeguards such as watermarking or provenance tracking that make it possible to identify synthetically generated content after distribution. The argument that general-purpose generation technologies cannot be held responsible for malicious uses is undermined by the extent to which deepfake harms were foreseeable from the outset of the technology's development, and by the existence of specific use cases, such as non-consensual intimate imagery, that have no legitimate application.

Platform operators face the ethical challenge of balancing the legitimate uses of synthetic media, which include entertainment, satire, accessibility applications, and artistic expression, against the harms enabled by the same technical capabilities. The development of regulatory frameworks capable of making these distinctions in ways that are consistent, enforceable, and resistant to capture by the most powerful actors in the technology industry is an urgent priority. Several jurisdictions have enacted or are developing legislation specifically targeting non-consensual deepfakes, but a globally coherent regulatory approach remains absent, creating arbitrage opportunities for malicious actors and inconsistent protection for potential victims (Westerlund, 2019).

7. Challenges in Deepfake Detection

7.1 Limitations of Current Detection Models

Despite significant advances in detection accuracy on benchmark datasets, current deepfake detection models exhibit several important limitations that constrain their real-world utility. The most widely documented limitation is overfitting to the specific characteristics of the training data, resulting in models that achieve high accuracy on the dataset they were trained on but generalise poorly to deepfakes generated using techniques not represented in the training set (Tolosana et al., 2020). This is a fundamental problem because the deepfake generation landscape is evolving rapidly, with new and improved generation methods being released on a timescale faster than the research community can assemble, annotate, and release new training datasets.

The difficulty of detecting unseen deepfakes is closely related to the problem of generalisation, but extends it in an important way. A model that performs well on a held-out test set drawn from the same distribution as its training data is not the same as a model capable of detecting deepfakes produced by generation methods it has never encountered. Studies that have systematically evaluated cross-dataset generalisation have consistently found dramatic performance drops when models trained on one benchmark are applied to a different one, with some models falling to near-random performance under these conditions (Rossler et al., 2019). Addressing this limitation requires either substantially more diverse training data, more theoretically grounded feature extraction approaches that capture fundamental properties of synthetic media rather than the idiosyncratic artefacts of specific tools, or both.

7.2 Adversarial Attacks and Evasion Techniques

The development of detection systems creates an incentive for adversaries to develop deepfake generation techniques specifically designed to evade those systems, creating a dynamic that researchers have

characterised as an arms race between creation and detection. Adversarial attacks on deepfake detectors can take several forms. In the most direct form, adversarial noise optimised to cause a specific detector to misclassify a deepfake as authentic can be added to a manipulated video at a level below the threshold of human visual perception, substantially reducing the detector's effectiveness without degrading the apparent quality of the deepfake (Neekhara et al., 2021). The existence of this attack vector means that detectors whose internal workings are publicly known or easily inferred can be directly targeted.

Beyond targeted adversarial noise attacks, the broader trajectory of deepfake generation research itself functions as a continuous evasion effort, since improved generation quality reduces the artefacts that detection models rely on. The introduction of diffusion model-based face synthesis has produced imagery of a quality that older detection models, trained primarily to recognise artefacts from GAN-based generation, struggle to identify reliably. This suggests that the arms race dynamic is not merely a feature of deliberate adversarial optimisation but is intrinsic to the relationship between generation and detection research, and that maintaining effective detection capabilities requires sustained and proactive investment rather than a one-time technological solution.

7.3 Dataset and Training Challenges

The quality and diversity of training data represents one of the most significant practical constraints on deepfake detection performance. High-quality annotated datasets are labour-intensive to produce: they require the collection of authentic video content with appropriate consent and rights clearances, the systematic application of multiple deepfake generation methods to that content, and careful quality control of the resulting annotations. The datasets that have been released publicly, while valuable, have tended to be limited in the diversity of identities, recording conditions, and generation methods they represent, leading to models trained on them that perform poorly on content exhibiting demographic or environmental variation not captured in the training data (Dolhansky et al., 2020).

Bias in training data is a related concern with both technical and ethical dimensions. If detection models are trained predominantly on data featuring particular demographic groups, they may develop lower detection accuracy for groups underrepresented in the training data, effectively providing less protection to those individuals. This is not a hypothetical concern: disparities in face recognition performance across demographic groups have been extensively documented, and similar disparities are plausible in deepfake detection models given that they share the same general technical approach and data challenges. Obtaining real-world deepfake samples for training presents an additional difficulty, since the most harmful categories of deepfake content cannot be ethically collected and used for training purposes, leaving a systematic gap between the distribution of training data and the distribution of content that detection systems most need to identify.

8. Future Directions and Research Gaps

8.1 Need for Robust Detection Models

The development of detection models capable of generalising reliably across the full diversity of deepfake generation techniques, recording conditions, and demographic groups is the central unsolved technical problem in the field. Progress toward this goal requires both advances in model architecture and approaches to training data that prioritise diversity and representativeness over sheer scale. Some researchers have argued that detection models should be designed to identify fundamental properties of synthetic media that are generation-method-agnostic, such as the statistical regularities introduced by any learned generative process, rather than the specific artefacts of particular tools (Tolosana et al., 2020).

Pursuing this approach requires a deeper theoretical understanding of what distinguishes synthetically generated from authentically captured imagery at a fundamental level.

Cross-platform detection capability represents a further dimension of robustness that has received insufficient attention in the research literature. Content shared on social media platforms typically undergoes multiple rounds of compression, format conversion, and resolution reduction before reaching a detector, and these transformations can degrade or eliminate the artefacts that models trained on pristine data rely on. Detection models intended for real-world deployment must be evaluated on, and ideally trained with, data that reflects the degradation pipeline of the platforms on which they will operate. Federated learning approaches, in which models are trained collaboratively across multiple platforms without centralising user data, may offer a technically and ethically viable path to developing more robust cross-platform detectors.

8.2 Real-Time and Scalable Solutions

The integration of deepfake detection into the content moderation infrastructure of major social media platforms represents both a significant technical challenge and a practical imperative given the scale at which manipulated content currently circulates. Platforms process billions of pieces of content daily, making the per-item computational cost of detection a critical constraint. Research into efficient neural architecture design, hardware-aware model optimisation, and tiered detection systems that apply expensive analysis only to content flagged as suspicious by lightweight preliminary filters will be essential to making real-time platform-scale detection feasible (Li & Lyu, 2019).

Deployment challenges extend beyond computational efficiency to include the integration of detection systems into existing platform workflows, the management of appeal processes for falsely flagged content, and the communication of detection outcomes to users in ways that are informative without being misleading about the reliability of the underlying technology. The last of these is particularly important given the risk that overconfident communication of detection results could undermine trust in authentic content or create a false sense of security about the completeness of platform-level detection. Collaborative frameworks in which platform operators, researchers, civil society organisations, and regulators jointly govern detection deployment decisions may be more likely to achieve equitable and effective outcomes than approaches driven by platform operators alone.

8.3 Policy and Regulatory Frameworks

The technical development of deepfake detection systems cannot on its own address the harms enabled by deepfake technology without complementary policy and regulatory frameworks. Global coordination is needed because the internet does not respect national borders: content created in one jurisdiction and hosted in another can cause harm to individuals in a third, and regulatory approaches that are not coordinated internationally create opportunities for regulatory arbitrage. International bodies including the United Nations and regional organisations such as the European Union have begun to engage with AI governance questions in ways that could eventually encompass deepfake-specific regulation, but progress remains slow relative to the pace of technological development (Westerlund, 2019).

The role of technology companies in shaping and implementing regulatory frameworks is both inevitable and contested. Platforms possess the technical infrastructure, the data, and the operational capacity to implement detection at scale in ways that governments cannot, but they also have commercial interests that may not align with the most effective or equitable approaches to harm prevention. Regulatory models that establish minimum detection and transparency standards while leaving implementation details to platform operators, subject to independent auditing and public reporting requirements, may offer a

workable balance between regulatory prescriptiveness and operational flexibility. The development of provenance standards such as the Content Authenticity Initiative's C2PA specification, which embeds cryptographic records of content history into media files at the point of capture, represents a promising complementary approach that addresses the challenge at the creation stage rather than the detection stage.

9. Conclusion

9.1 Summary of Findings

This paper has examined the landscape of AI-based deepfake detection across its technical, security, ethical, and policy dimensions. A central finding is that while machine learning-based detection approaches, and deep learning models in particular, have achieved substantially higher accuracy on benchmark datasets than traditional signal processing methods, the gap between benchmark performance and real-world deployment effectiveness remains wide. CNN architectures such as ResNet and EfficientNet, transformer-based models, and hybrid approaches combining both have each demonstrated strong results under controlled conditions, but consistent generalisation across generation methods, platforms, and demographic groups has not yet been achieved by any existing system (Rossler et al., 2019; Tolosana et al., 2020).

The key challenges identified in this paper include the problem of cross-dataset generalisation, the ongoing arms race between generation and detection capabilities, the limitations imposed by existing training datasets, and the practical difficulties of deploying computationally intensive detection systems at platform scale. These challenges are interconnected: the generalisation problem is exacerbated by dataset limitations, the arms race dynamic makes static solutions inherently temporary, and computational constraints limit the complexity of models that can be deployed in real-time contexts. Addressing any one of these challenges in isolation is unlikely to produce durable improvements in detection capability; progress requires coordinated advances across all of them simultaneously.

9.2 Implications of the Study

The findings of this paper have implications for several intersecting fields. For cybersecurity, they highlight the inadequacy of current identity verification systems in the face of deepfake-enabled impersonation, and the need for multi-factor verification approaches that do not rely solely on visual or auditory identity signals. For media authenticity and journalism, they underscore the urgency of developing and adopting provenance standards that make it possible to verify the chain of custody of digital media, reducing the evidential burden placed on detection algorithms alone. For social media regulation, they support the case for mandatory minimum standards of deepfake detection capability, transparency reporting, and user recourse mechanisms, as voluntary platform-led approaches have not produced consistent or adequate outcomes.

The contribution of this paper to future AI research lies in its synthesis of technical and non-technical dimensions of the deepfake detection problem. Purely technical treatments of detection tend to focus on benchmark accuracy improvements without adequate attention to the deployment context, adversarial dynamics, or equity implications of the systems being developed. Conversely, policy-focused treatments of deepfake harms sometimes underestimate the genuine technical constraints that limit what detection systems can currently achieve. A research agenda that takes both dimensions seriously, and that treats the relationship between technical capability and social context as a core object of study rather than an afterthought, is likely to produce more durable and equitable contributions to the problem than either approach pursued independently.

9.3 Final Remarks

The challenge of detecting deepfake media is not a problem that will be solved once and then set aside. The generation capabilities that make convincing deepfakes possible will continue to improve, new application contexts and harms will continue to emerge, and the social and political environment in which deepfake content circulates will continue to evolve in ways that are difficult to predict. Maintaining effective detection capabilities in this environment requires sustained institutional commitment to research funding, data infrastructure, and cross-sector collaboration, rather than the treatment of deepfake detection as a discrete engineering problem to be solved and deployed.

Ethical AI development is a necessary condition for detection systems that are not merely technically effective but genuinely beneficial. Systems that achieve high accuracy on average while failing systematically for particular demographic groups, that are deployed in ways that undermine legitimate speech or are exploited for mass surveillance, or that create false confidence in the completeness of platform-level moderation are not ethical merely by virtue of their stated purpose. The long-term project of maintaining digital trust, which is ultimately what deepfake detection is in service of, requires that the development, deployment, and governance of detection systems be held to standards of transparency, accountability, and equity that match the seriousness of the harms they are designed to prevent. Technical innovation and ethical responsibility are not competing priorities in this domain; they are jointly necessary conditions for progress that serve the public interest.

References

1. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J
2. Ciftci, U. A., Demir, I., & Yin, L. (2020). FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Advance online publication. https://doi.org/10.1109/TPAMI.2020.3009287
3. Fridrich, J., & Kodovský, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security, 7*(3), 868–882. https://doi.org/10.1109/TIFS.2012.2190402
4. Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–6). IEEE. https://doi.org/10.1109/AVSS.2018.8639163
5. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. https://doi.org/10.1109/CVPR.2016.90
6. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5001–5010). https://doi.org/10.1109/CVPR42600.2020.00505
7. Li, Y., Chang, M.-C., & Lyu, S. (2018). In ictu oculi: Exposing AI-created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1–7). IEEE. https://doi.org/10.1109/WIFS.2018.8630787
8. Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3207–3216). https://doi.org/10.1109/CVPR42600.2020.00327
9. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing*

Surveys, 54(1), Article 7, 1–41. https://doi.org/10.1145/3425780

10. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 2823–2832). Association for Computing Machinery. https://doi.org/10.1145/3394171.3413570
11. Neekhara, P., Dolhansky, B., Bitton, J., & Ferrer, C. C. (2021). Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 923–932). https://doi.org/10.1109/CVPRW53098.2021.00103
12. 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*, Article 103525. https://doi.org/10.1016/j.cviu.2022.103525
13. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1–11). https://doi.org/10.1109/ICCV.2019.00009
14. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2387–2395). IEEE. https://doi.org/10.1109/CVPR.2016.262
15. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014
16. Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1). https://doi.org/10.1177/2056305120903408
17. Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101
18. Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 40–53. https://doi.org/10.22215/timreview/1282