

Comparative Analysis of Multi-Modal Deepfake Detection Techniques for Digital Forensics

Shobha N

Reg. No. 24MSFS121

Under the guidance of

Mr. Rakesh V

Assistant Professor

Department of Forensic Science, Garden City University

ABSTRACT

The rapid development of artificial intelligence and deep learning technologies has significantly contributed to the emergence of deepfake technology. Deepfakes are AI-generated synthetic media that can create highly realistic manipulated videos, images, and audio files that closely resemble authentic digital content. Although deepfake technology has beneficial applications in entertainment, education, virtual reality, and digital media production, its misuse has become a serious concern in cybersecurity and digital forensics. Deepfakes are increasingly being used in cybercrimes such as identity theft, financial fraud, misinformation campaigns, fake news dissemination, impersonation attacks, and social media manipulation. Therefore, the identification and analysis of deepfake content has become an important area of research in modern digital forensic investigations.

The primary aim of this project is to perform a comparative analysis of various multi-modal deepfake detection techniques used in digital forensics. The study mainly focuses on analysing different forensic software tools and AI-based detection methods that can be used by forensic investigators and cybersecurity professionals with minimal technical complexity. Several detection approaches such as facial movement analysis, texture inconsistency detection, lighting and shadow analysis, deep learning classifiers, frequency-domain analysis, and behavioural feature examination were studied using manipulated images and videos collected from publicly available datasets and online sources. The performance of the selected forensic tools was evaluated based on important parameters such as detection accuracy, precision, recall, false positive rate, processing speed, and robustness against high-quality synthetic media. Apart from technical analysis, the study also examines the ethical, social, and legal challenges associated with deepfake technology.

The increasing use of AI-generated manipulated media has created serious concerns regarding public trust, privacy violations, cyber security threats, and the authenticity of digital evidence used in criminal investigations. The project further highlights the importance of user-friendly forensic tools that can be integrated into digital forensic workflows for effective identification of manipulated content. The findings of the study indicate that AI-assisted hybrid detection techniques provide better performance when compared to traditional standalone forensic approaches. However, highly advanced deepfake models and adversarial modifications still remain challenging for existing detection systems.

KEYWORDS:

Deepfake, Digital Forensics, Artificial Intelligence, Hybrid Detection Techniques, Synthetic Media Analysis.

CHAPTER 1

INTRODUCTION:

The rapid advancement of artificial intelligence and deep learning technologies has significantly contributed to the emergence of deepfake technology. Deepfakes are AI-generated manipulated media that can create highly realistic synthetic videos, images, and audio files that closely resemble authentic digital content. While deepfake technology has useful applications in entertainment, education, virtual reality, speech synthesis, and digital media production, it has also created serious ethical and security concerns related to cyber fraud, misinformation, identity theft, and digital manipulation [3]. With the ability to generate convincing synthetic media, deepfakes have become a powerful tool for both legitimate and malicious applications, particularly in social engineering attacks, online scams, and political propaganda [38].

The development of deepfake technology is closely associated with the advancement of deep learning architectures such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion-based artificial intelligence models [27]. Generative adversarial models consist of a generator network that produces manipulated media content and a discriminator network that evaluates whether the generated media appears authentic or synthetic. The primary objective of these AI systems is to improve the realism and quality of synthetic digital media [35].

Deepfake technology has been widely misused in various domains including political manipulation, cyber fraud, fake news dissemination, revenge pornography, and biometric authentication bypassing systems [38]. In politics, synthetic videos of public leaders have been used to spread false information and manipulate public opinion during elections and social campaigns [3]. Cybercriminals also use deepfake videos and AI-generated voice cloning methods to bypass KYC verification systems, resulting in major financial losses for banks and enterprises [45]. Furthermore, the increasing spread of manipulated media across social networking platforms has reduced public trust in digital communication systems and online media content [38].

To combat the growing threats associated with deepfake technology, researchers have developed several forensic detection techniques based on artificial intelligence, machine learning, and hybrid forensic analysis methods [23]. Image-based deepfake detection techniques commonly use convolutional neural networks (CNNs) to identify anomalies such as facial warping, texture inconsistencies, unnatural lighting conditions, and image distortion artifacts. Audio-based detection systems analyse waveform inconsistencies, speech modulation patterns, and lip synchronization mismatches to identify manipulated audio content [45].

Several studies also suggest that behavioural biometrics such as eye movement analysis, facial micro-expression examination, and heartbeat fluctuation monitoring can help distinguish genuine videos from manipulated synthetic media [38]. Another emerging approach known as cognitive response-based detection uses electroencephalogram (EEG) analysis to study human brain responses while processing real and deepfake media content, helping identify neural inconsistencies associated with manipulated digital information [56].

Despite significant advancements in deepfake detection techniques, several challenges continue to affect the effectiveness of current forensic systems. Deepfake detection models trained on specific datasets often struggle to identify manipulated media generated using different AI frameworks, leading to false negatives and reduced detection accuracy [38]. Adversarial deepfake models can also be intentionally designed to evade forensic detection systems, making current AI-based forensic methods less effective in real-world investigations [35].

In addition, many deepfake detection techniques require high computational resources and advanced hardware systems, making them difficult to implement in real-time large-scale applications [23]. The rapid improvement of diffusion-based AI models has further increased the complexity of deepfake detection by producing synthetic media with highly realistic facial textures, natural lighting conditions, and accurate human expressions. Digital forensics plays an important role in identifying manipulated media used in cybercrimes, misinformation campaigns, and online fraud. The increasing use of AI-generated synthetic media has raised major concerns regarding the authenticity and integrity of digital evidence used in criminal investigations. Advanced forensic tools and hybrid AI-based detection systems can

assist digital forensic investigators in analysing manipulated videos, images, and audio files during cyber security investigations and criminal case analysis.

The rapid growth of deepfake technology has created both opportunities and challenges in the digital world. Although deepfakes provide innovative applications in entertainment, virtual reality, and digital content creation, their misuse in cybercrime and misinformation poses significant threats to society. Therefore, the development of reliable and efficient multi-modal deepfake detection techniques is essential for improving digital forensic investigations and maintaining trust in digital communication systems.

1.1 Technological Foundations of Deepfakes

The advancement of synthetic media technology is closely tied to the development of deep learning architectures, particularly generative models such as GANs, VAEs, and diffusion models. These models work by learning the underlying distribution of real data and then generating new data that mimics the original. In the case of GANs, the process involves two neural networks: a generator that creates synthetic media and a discriminator that evaluates whether the generated content is real or fake. Through an iterative process, the generator improves its ability to produce realistic content, while the discriminator becomes better at detecting fakes.

The ultimate goal is to create media that is indistinguishable from real content, raising significant challenges for detection [25],[56]

The diffusion model, a more recent innovation, has further enhanced the quality of deepfakes by generating images through a process of iterative refinement. Unlike GANs, which rely on adversarial training, diffusion models gradually add noise to an image and then learn to reverse the process, resulting in highly realistic outputs [35]. This has made it even more challenging to detect deepfakes, as the generated content often lacks the telltale artifacts that were once common in earlier GAN-based deepfakes. For example, diffusion-based deepfakes can produce images with perfectly natural skin textures, realistic lighting, and accurate facial expressions, making them nearly impossible to detect with traditional methods [23].

1.2 Malicious use of Deepfakes

Deepfakes have been misused in a variety of contexts, including:

- **Political Manipulation:** Synthetic videos of politicians making false statements have been used to sway public opinion and interfere in elections. For example, during the 2024 U.S. presidential election, a deepfake video of a candidate making inflammatory remarks went viral, leading to widespread controversy and confusion [3]. Similarly, in 2023, a deepfake video of a European leader announcing a fake policy change caused a temporary stock market crash, highlighting the potential for deepfakes to disrupt financial markets [38].
- **Cyber Fraud:** Deepfake videos and images have been used to bypass biometric authentication systems, leading to financial fraud and identity theft. In one notable case, cybercriminals used a deepfake video to impersonate a CEO and authorize a fraudulent transfer of \$35 million [60]. Another case involved the use of deepfake audio to mimic a bank customer's voice, allowing fraudsters to access their account and steal funds [5].
- **Revenge Porn:** Deepfake technology has been used to create non-consensual explicit content, causing harm to individuals' reputations and mental health. In 2023, a study found that over 90% of deepfake content online was non-consensual pornography, with women being the primary targets [5]. This has led to calls for stricter regulations and better detection methods to combat the spread of deepfake pornography.
- **Fake News:** The spread of deepfake content on social media has contributed to the dissemination of false information, undermining trust in digital media. For example, during the COVID-19 pandemic, deepfake videos of health officials making false claims about vaccines were widely shared, leading to vaccine hesitancy [38]. Similarly, deepfake videos of fake news anchors reporting fabricated events have been used to spread misinformation and sow discord [3].

1.3 Challenges in Deepfake Detection

Despite the growing threat posed by deepfakes, detecting them remains a significant challenge. Current detection methods face several limitations, including:

- **Generalization Failures:** Deepfake detection models trained on specific datasets often struggle to detect deepfakes generated using different AI setups, leading to false negatives. For example, a model trained on GAN-based deepfakes may fail to detect those generated using diffusion models [5]. This is because different generative models produce different types of artifacts, and detection models may not be robust enough to generalize across these variations.
- **Adversarial Robustness:** Adversarial deepfakes can be specifically designed to evade detection models, making it difficult to rely on existing forensic methods in real-world scenarios. Researchers have shown that even state-of-the-art detection models can be fooled by carefully crafted adversarial examples [10]. For instance, attackers can introduce subtle perturbations to deepfake images or videos that are imperceptible to humans but can cause detection models to misclassify the content as real.
- **Computational Inefficiencies:** Many detection methods are computationally intensive, making them unsuitable for real-time applications or large-scale deployment. For instance, some deepfake detection algorithms require significant computational resources, limiting their use in resource-constrained environments [33]. This is particularly problematic for social media platforms, where millions of videos are uploaded every day, and real-time detection is essential to prevent the spread of malicious content.

1.4 Current Approaches to Deepfake Detection

To combat the threats posed by deepfakes, researchers have developed a variety of detection algorithms based on machine learning, deep learning, and multi-modal approaches. These methods can be broadly categorized into:

1. Image-Based Detection:

- Uses Convolutional Neural Networks (CNNs) to identify anomalies such as facial warping, unnatural light reflections, and texture inconsistencies at the pixel level. For example, researchers have developed CNN-based models that can detect subtle artifacts in deepfake images, such as irregular eye movements or unnatural skin textures [60]. These models are trained on large datasets of real and fake images, allowing them to learn the distinguishing features of deepfakes.

2. Audio-Based Detection:

- Employs AI-driven speech analysis models to detect inconsistencies in waveform structures, vocal modulations, and lip-sync mismatches. For instance, some models analyse the spectral characteristics of audio signals to identify synthetic speech [10]. These methods are particularly useful for detecting deepfake audio, which is often used in conjunction with deepfake videos to create more convincing forgeries.

3. Behavioural Biometrics:

- Analyses micro expressions, heartbeat fluctuations, and natural eye movements to distinguish between real and synthetic videos. For example, researchers have developed algorithms that can detect subtle changes in facial expressions or eye movements that are indicative of deepfakes [38]. These methods are based on the idea that humans exhibit certain involuntary behaviours that are difficult to replicate in synthetic content.

4. Cognitive Response-Based Detection:

- Uses Electroencephalogram (EEG) responses to analyse how the human brain processes deepfake content versus real content, identifying neural discrepancies. This approach is based on the idea that the brain responds differently to synthetic content, even if it appears realistic to the naked eye [16]. For example, studies have shown that the brain's response to deepfake videos is less pronounced than its response to real videos, suggesting that cognitive

response-based detection could be a promising avenue for future research.

The Need for a Multi-Modal Approach

While these methods have shown promise, they are not without limitations. For instance, image-based detection may fail to identify deepfakes that do not exhibit visual anomalies, while audio-based detection may struggle with high-quality synthetic speech. To address these challenges, researchers are increasingly advocating for a multi-modal approach that combines multiple detection techniques. By integrating visual, audio, behavioural, and cognitive analysis, a multi-modal approach can improve the accuracy and robustness of deepfake detection, making it more effective in real-world scenarios. For example, a hybrid model that combines CNN-based image analysis with EEG-based cognitive response analysis has been shown to achieve higher accuracy than either method alone [23].

Future Directions

The ongoing arms race between deepfake creators and detectors underscores the need for continued research and innovation in this field. Future research should focus on:

- **Improving AI Models:** Developing more robust and generalizable deepfake detection models that can adapt to new and evolving deepfake techniques. For example, researchers are exploring the use of self-supervised learning to improve the generalization capabilities of detection models [38]. This approach involves training models on large, diverse datasets without explicit labels, allowing them to learn more robust features that can be applied to a wide range of deepfake types.
- **Automating Hybrid Detection:** Creating automated systems that can combine multiple detection methods for large-scale applications, such as social media platforms and cybersecurity systems. For instance, some researchers are developing real-time deepfake detection systems that can analyse video streams on social media platforms and flag potentially malicious content [10]. These systems use a combination of image, audio, and behavioural analysis to detect deepfakes in real-time, making them ideal for large-scale deployment.
- **Enhancing Computational Efficiency:** Reducing the computational cost of deepfake detection to enable real-time processing and deployment in resource-constrained environments. For example, some researchers are exploring the use of edge computing to perform deepfake detection on mobile devices [23]. This approach involves running detection algorithms directly on the device, rather than relying on cloud-based servers, which can significantly reduce latency and improve efficiency.

1.5 The Evolution of Deepfake Technology

The evolution of deepfake technology can be traced back to the early 2010s when researchers began experimenting with neural networks to generate realistic images and videos. The advent of GANs in 2014 marked a significant milestone, as these models demonstrated the ability to generate highly realistic images by pitting two neural networks against each other. Over the years, advancements in computational power and the availability of large datasets have further accelerated the development of deepfake technology. Today, deepfakes are not limited to images and videos but also include audio deepfakes, which can mimic a person's voice with remarkable accuracy. The integration of deepfake technology with other AI-driven tools, such as natural language processing (NLP) and computer vision, has further expanded its capabilities, enabling the creation of fully synthetic personas that can interact with users in real-time.

Ethical and Legal Implications

The ethical and legal implications of deepfake technology are profound. On the ethical front, the creation and dissemination of deepfakes raise questions about consent, privacy, and the potential for harm. For instance, the use of deepfakes in non-consensual pornography has led to significant psychological trauma for victims, highlighting the need for stricter ethical guidelines and legal frameworks. From a legal perspective, the lack of comprehensive regulations governing the use of deepfake technology has created a gray area, making it difficult to hold perpetrators accountable. While some countries have introduced legislation to address the misuse of deepfakes, there is a pressing need for international cooperation to develop standardized legal frameworks that can effectively combat the global

threat posed by deepfakes.

Societal Impact

The societal impact of deepfake technology is multifaceted. On one hand, deepfakes have the potential to democratize content creation, enabling individuals and small businesses to produce high-quality media without the need for expensive equipment or specialized skills. This has led to a surge in user-generated content, fostering creativity and innovation. On the other hand, the misuse of deepfakes has eroded public trust in digital media, leading to a phenomenon known as the "liar's dividend," where individuals can dismiss authentic content as fake. This has significant implications for journalism, where the credibility of news sources is paramount.

Additionally, the spread of deepfake content on social media has exacerbated the problem of misinformation, making it increasingly difficult for individuals to discern fact from fiction.

Technological Innovations in Deepfake Detection

In response to the growing threat posed by deepfakes, researchers have developed a range of innovative detection methods. One promising approach is the use of blockchain technology to verify the authenticity of digital content. By embedding cryptographic signatures into media files, blockchain-based systems can provide a tamper-proof record of a file's origin, making it easier to detect deepfakes. Another innovative approach involves the use of quantum computing to enhance the computational power of deepfake detection algorithms. Quantum computers, with their ability to perform complex calculations at unprecedented speeds, could potentially revolutionize the field of deepfake detection by enabling real-time analysis of large datasets.

The Role of AI in Deepfake Detection

AI plays a crucial role in the detection of deepfakes, with machine learning algorithms being at the forefront of this effort. One of the key challenges in deepfake detection is the need for models to generalize across different types of deepfakes. To address this, researchers are exploring the use of transfer learning, where a model trained on one type of deepfake is fine-tuned to detect another. Additionally, the use of ensemble learning, where multiple models are combined to improve detection accuracy, has shown promise in enhancing the robustness of deepfake detection systems. As AI continues to evolve, it is likely that we will see the development of more sophisticated detection methods that can adapt to the ever-changing landscape of deepfake technology.

The Future of Deepfake Technology

Looking ahead, the future of deepfake technology is both exciting and uncertain. On the one hand, advancements in AI and machine learning are likely to lead to the development of even more realistic deepfakes, further blurring the line between reality and fiction. On the other hand, the ongoing efforts to improve deepfake detection methods offer hope that we can mitigate the risks associated with this technology. As we move forward, it is essential that we strike a balance between harnessing the potential of deepfake technology for positive applications and protecting society from its potential harms. This will require a collaborative effort involving researchers, policymakers, and industry stakeholders to develop ethical guidelines, legal frameworks, and technological solutions that can ensure the responsible use of deepfake technology.

The ability to create **highly convincing synthetic videos and images** has made deepfakes a double-edged sword. On one hand, they have legitimate uses in fields such as **medical imaging, education, and entertainment**. For example, deepfake technology has been used to create realistic simulations for medical training, allowing students to practice complex surgical procedures in a risk-free environment [7]. In the entertainment industry, deepfakes have been used to de-age actors, resurrect deceased performers, and create entirely synthetic characters, pushing the boundaries of creative storytelling [27]. On the other hand, deepfakes have been weaponized for **social engineering, online scams, and political propaganda**. For instance, deepfake videos have been used to create fake speeches by politicians, manipulate public opinion, and interfere in elections [3]. Similarly, cybercriminals have used deepfakes to bypass **Know Your Customer (KYC)** verification systems, resulting in significant financial losses for banks and enterprises [60]. The widespread dissemination of deepfake content on social media has also eroded public

trust in digital media, making it increasingly difficult for individuals to distinguish between **real and fake content** [38].

Chapter 2

REVIEW OF LITERATURE:

- **Cohen et al. (2022) – A Study on Data Augmentation in Voice Anti-Spoofing** This study explores the role of data augmentation techniques in voice anti-spoofing, particularly in the context of automatic speaker verification (ASV) systems. Given the rise of synthetic voice attacks using generative adversarial networks (GANs) and voice cloning technologies, this research aims to improve the robustness of deep-learning models against spoofed speech samples. The authors experiment with various augmentation techniques, including pitch shifting, time stretching, and additive noise, to enhance model generalization. Their findings indicate that certain augmentation methods significantly improve detection accuracy, particularly in low-resource scenarios where training data is limited. However, challenges remain in balancing augmentation intensity to avoid degrading model performance. Future work could explore adversarial training strategies to counteract sophisticated spoofing attacks.
- **Groh et al. (2021) – Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds.** This paper investigates the effectiveness of deepfake detection by comparing human observers, machine learning models, and hybrid approaches where machines provide guidance to human evaluators. Using a dataset of manipulated videos, the authors assess detection accuracy across these three groups. Their findings suggest that individual human performance is relatively low, but collective intelligence—when multiple human judgments are aggregated—improves accuracy. Machine-based detection models outperform most humans but still struggle with high-quality deepfakes. Notably, machine-informed crowds achieve the best results, indicating that AI assistance can enhance human decision-making. The study raises ethical concerns about the reliance on automated systems and the potential biases introduced by human-machine interactions.
- **Hu, Li, & Lyu (2020) – Exposing GAN-Generated Faces Using Inconsistent Corneal Specular Highlights.** This research presents a novel forensic technique to detect GAN-generated faces by analysing corneal specular highlights—the light reflections on the eye’s cornea. The authors argue that natural eyes exhibit symmetrical reflections due to the physics of light, whereas deepfake images often fail to replicate this consistency. Their model extracts and compares corneal reflections from both eyes to detect discrepancies. The method proves effective across various deepfake datasets, highlighting its potential as a reliable forensic tool. However, the approach may be less effective when deepfake generators become more advanced in replicating fine details. Future research could integrate this technique with multimodal deepfake detection frameworks.
- **Boháček & Farid (2022) – Protecting World Leaders Against Deepfakes Using Facial, Gestural, and Vocal Mannerisms.** This paper addresses the increasing threat of deepfake videos targeting political figures. The authors propose a detection framework based on biometric signatures, including facial expressions, hand gestures, and vocal patterns. By training deep learning models on known footage of world leaders, the system can identify inconsistencies in mannerisms that are difficult to replicate accurately. Their results demonstrate that this method is highly effective in distinguishing authentic videos from deepfakes. However, the model requires extensive training data for each individual, making it less practical for lesser-known figures. The study underscores the importance of biometric-based forensic tools in countering deepfake misinformation.
- **Masi et al. (2020) – Two-Branch Recurrent Network for Isolating Deepfakes in Videos** The authors introduce a two-branch recurrent neural network (RNN) designed to detect deepfake manipulations in videos. One branch focuses on spatial features (such as inconsistencies in facial texture and lighting), while the other captures temporal artifacts (such as unnatural motion and frame inconsistencies). The combination of these two modalities enhances detection accuracy, making it more robust than traditional convolutional neural networks (CNNs) that analyse only static frames. The study is particularly relevant for detecting deepfake videos in real-time applications, such as social media monitoring. However, the approach requires high computational power, posing challenges for large-scale deployment.

- **Cozzolino & Verdoliva (2020) – Noise print: A CNN-Based Camera Model Fingerprint**

This study proposes a novel forensic technique called Noise print, which uses CNNs to extract camera model fingerprints from images. The authors argue that every camera sensor introduces unique noise patterns, which can be used to distinguish real images from synthetic ones. By analysing these noise fingerprints, Noise print effectively detects deepfake images and other digital manipulations. The method is particularly useful for verifying image authenticity in journalism and legal proceedings. However, the approach may struggle with highly compressed or heavily edited images. Future research could explore integrating Noise print with other forensic methods to enhance robustness.

- **Li & Lyu (2019) – Exposing Deepfake Videos by Detecting Face Warping Artifacts** This research focuses on face warping artifacts—distortions that arise when deepfake algorithms manipulate facial structures. The authors develop a detection model that identifies these inconsistencies by analysing geometric distortions and texture inconsistencies. Their method is particularly effective against lower-quality deepfakes, though high-resolution fakes may bypass detection. The study highlights the need for adaptive forensic tools that evolve alongside deepfake generation techniques.

- **Agarwal et al. (2019) – Protecting World Leaders Against Deepfakes.** Similar to Boháček & Farid (2022), this paper emphasizes the importance of biometric cues in detecting deepfakes targeting political figures. The authors develop a model that analyses speech patterns, facial expressions, and head movements to identify inconsistencies in manipulated videos. Their findings reveal that deepfake algorithms struggle to replicate natural behavioural patterns, making biometric analysis a promising detection approach. The study suggests that combining multiple forensic methods could improve deepfake detection accuracy.

- **Yang, Li, & Lyu (2019) – Exposing Deepfakes Using Inconsistent Head Poses**

This paper proposes a detection method based on head pose inconsistencies. The authors argue that deepfake algorithms often fail to maintain realistic 3D head orientations due to imperfect warping techniques. Their model detects these anomalies by estimating head poses and comparing them to expected biomechanical constraints. The approach proves highly effective against many deepfake variations, though more sophisticated algorithms may eventually minimize these errors.

- **Matern, Riess, & Stamminger (2019) – Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations.** This paper delves into the identification of deepfakes by analysing visual artifacts that arise due to imperfections in generative models. The authors focus on anomalies such as unnatural eye blinking, texture inconsistencies in facial regions, and lighting mismatches—common issues in manipulated images and videos. Their approach utilizes a machine-learning model trained on datasets of both real and deepfake images to detect these irregularities. One of the study's key insights is that many deepfake models fail to replicate realistic eye movement patterns, often producing subjects that blink unnaturally or not at all. Additionally, inconsistencies in facial textures, such as unnatural skin smoothing or asymmetric distortions, further expose fakes. The researchers also highlight that illumination and shadow inconsistencies, particularly in complex lighting environments, serve as another strong indicator of manipulation. While the method proves effective, a major limitation is that as deepfake algorithms improve, these artifacts may become less detectable. The study underscores the need for adaptive detection strategies that evolve alongside synthetic media advancements.

- **McCloskey & Albright (2018) – Detecting GAN-Generated Imagery Using Colour Cues.** In this study, McCloskey and Albright explore a unique approach to deepfake detection by analysing colour inconsistencies in GAN-generated images. The research identifies that many deepfake algorithms struggle to maintain realistic colour distributions, particularly in human skin tones, shadows, and reflections. The authors argue that real images typically exhibit smooth and natural colour transitions, whereas deepfake images often display irregularities due to the way GANs generate and reconstruct pixel values. Their model detects these anomalies by computing statistical differences in colour histograms and examining subtle discrepancies in hue variations. A significant finding of the study is that GANs frequently fail to accurately replicate the complex interplay of light and colour found in natural scenes. However, the approach has limitations, as some advanced GANs are now capable of generating more realistic colour distributions, making detection more challenging. The study suggests that combining colour analysis with other forensic methods—such as texture and noise pattern analysis—could enhance deepfake detection accuracy.

- **Marra et al. (2018) – Detection of GAN-Generated Fake Images Over Social Networks.** This research

focuses on detecting and tracking GAN-generated fake images that circulate on social media platforms. The authors propose a convolutional neural network (CNN)-based forensic tool to identify deepfake images by analysing structural artifacts introduced during the GAN synthesis process. The study highlights that social media platforms often compress and resize images, which can affect detection accuracy. To address this, the authors develop a robust model trained on a diverse dataset that includes various compression levels and resolution scales. Their findings indicate that even when images undergo heavy compression, certain GAN-related artifacts—such as checkerboard patterns, unnatural textures, and edge inconsistencies—remain detectable. The study also examines the role of metadata and digital footprints in identifying manipulated content. One challenge noted is that as deepfake technology evolves, the differences between real and fake images are narrowing, requiring continuous updates to detection algorithms. The authors advocate for collaboration between social media platforms and forensic researchers to combat the spread of synthetic media.

- **Verdoliva (2020) – Media Forensics and Deepfakes: An Overview.** This paper serves as a comprehensive survey of deepfake detection techniques, outlining various approaches used in media forensics. The study categorizes detection methods into three primary groups: physical, physiological, and behavioural analysis. Physical methods focus on identifying inconsistencies in pixel distributions, noise patterns, and compression artifacts. Physiological methods analyse biological cues such as eye movement, heartbeat signals in videos, and facial muscle deformations. Behavioural analysis examines speech patterns, mannerisms, and head movements to detect discrepancies in deepfake videos. The paper also explores deep learning-based approaches, which utilize CNNs, recurrent neural networks (RNNs), and transformer models to classify manipulated content. A key takeaway is that no single detection method is foolproof; instead, a combination of multiple techniques yields the most reliable results. The author emphasizes the urgent need for real-time deepfake detection solutions, particularly for misinformation mitigation in journalism and social media. The paper also discusses the ethical implications of deepfake detection, including privacy concerns and potential misuse of forensic tools.

- **Tolosana et al. (2020) – Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection.** They provide an extensive review of face manipulation techniques and their corresponding detection methods. The paper categorizes deepfake generation approaches into four major types: face swapping, facial reenactment, expression manipulation, and full-face synthesis. The authors analyse how different manipulation techniques leave distinct forensic traces, which can be leveraged for detection. One of the key contributions of this survey is its comparison of traditional forensic approaches (such as watermarking and compression analysis) versus modern deep learning techniques (such as feature extraction using CNNs). The study also discusses adversarial attacks, where deepfake generators are intentionally designed to evade detection systems, posing a significant challenge for forensic researchers. The authors suggest that future deepfake detection frameworks should integrate multimodal analysis—combining audio, video, and text cues—to improve accuracy. Additionally, the paper highlights the need for more diverse datasets to train robust detection models, as many current datasets are biased towards specific demographic groups. This work underscores the evolving arms race between deepfake creators and forensic scientists.

- **Rossler et al. (2019) – Face Forensics++: Learning to Detect Manipulated Facial Images.** This designed to benchmark deepfake detection models. The dataset includes thousands of manipulated facial images and videos generated using various deepfake techniques, providing a valuable resource for training forensic models. The authors train and evaluate several deep learning architectures, including Xception Net, to detect manipulated content with high accuracy. One of the study's major contributions is its comparison of model performance across different deepfake types, showing that some techniques are more challenging to detect than others. The dataset also includes both high-quality and low-quality deepfakes, simulating real-world scenarios where manipulated content varies in resolution and compression level. The findings reveal that deep learning models trained on large, diverse datasets perform significantly better than those trained on limited data. However, the study acknowledges that as deepfake algorithms improve, detection models must continuously adapt. The authors emphasize the need for collaboration between researchers, tech companies, and policymakers to develop standardized benchmarks and countermeasures against deepfake threats.

- **Dang et al. (2020) – Detailed Analysis of Deepfake Techniques and Detection Approaches.** This paper provides a technical breakdown of various deepfake generation techniques and the corresponding detection strategies. The authors categorize deepfake methods into autoencoder-based models (such as DeepFaceLab), GAN-based approaches (like StyleGAN), and hybrid techniques that combine elements of both. The study explores how

deepfake models have evolved from simple face-swapping algorithms to highly realistic video synthesis techniques capable of mimicking subtle facial expressions, micro-movements, and voice patterns. A significant contribution of this paper is its in-depth examination of detection algorithms, ranging from traditional forensic approaches (such as frequency analysis and compression artifact detection) to AI-driven models (such as convolutional neural networks trained on large datasets). The authors also discuss the role of adversarial machine learning, highlighting how deepfake creators are actively improving their models to bypass forensic detection. One challenge noted in the study is that as deepfake datasets become more diverse, models trained on specific data distributions may struggle to generalize to unseen manipulations. The paper suggests multi-modal detection approaches—which combine audio, facial, and textual cues—as a promising direction for future research.

- **Korshunov et.al., (2019) – Vulnerabilities of Deepfake Detection in Real-World Scenarios.** They focus on evaluating the robustness of deepfake detection systems under real-world conditions, where images and videos undergo compression, resizing, and noise addition. The study reveals that many AI-based deepfake detection models perform well under controlled laboratory settings but fail when tested on manipulated media found on social platforms. The authors systematically analyse how common factors such as JPEG compression, video bitrate reduction, and frame loss impact detection accuracy. Their findings indicate that heavily compressed deepfakes are significantly harder to detect, as many forensic traces—such as pixel inconsistencies and GAN artifacts—are lost during compression. To address these issues, the paper proposes adaptive detection models that can be fine-tuned to account for varying levels of image quality. One major takeaway is that forensic researchers must develop real-time, scalable detection techniques that can handle deepfake content as it appears across different platforms, including TikTok, YouTube, and Twitter. The paper highlights that while high-resolution forensic analysis is effective in legal settings, social media platforms require lightweight AI models that can quickly filter out manipulated content with minimal computational overhead.
- **Cozzolino et al. (2019) – Autoencoder-Based Approaches for Deepfake Detection.** This paper investigates the use of autoencoders for detecting manipulated facial images and videos. Unlike traditional deepfake detection methods that rely on hand-crafted forensic features, autoencoders automatically learn hidden representations of real and fake images. The researchers train a variational autoencoder (VAE) on a dataset of real faces and analyse how well it reconstructs deepfake images. The underlying idea is that a model trained only on real faces will struggle to reconstruct GAN-generated images, thereby exposing the presence of manipulation. Their experiments reveal that deepfake images exhibit distinct residual errors when passed through an autoencoder, making them easier to identify. One advantage of this approach is that it does not require manually engineered forensic features, making it more adaptable to evolving deepfake techniques. However, a limitation is that VAEs may misclassify highly realistic deepfakes that closely resemble training samples, highlighting the need for complementary forensic methods. The study concludes that autoencoder-based models, when combined with frequency analysis and texture-based features, can significantly improve deepfake detection accuracy.
- **Jung et al. (2020) – Deepfake Audio and Video: Cross-Domain Analysis and Detection.** They expand deepfake forensics beyond facial manipulations by examining deepfake audio and video in a cross-domain analysis. The study addresses a crucial gap in research: while most deepfake detection techniques focus on visual artifacts, deepfake audio is an emerging threat, particularly in social engineering attacks and misinformation campaigns. The researchers propose a multi-modal detection framework that analyses both speech patterns and facial movements in deepfake videos. Their model identifies inconsistencies between lip-sync movements and generated speech, which is a common flaw in deepfake videos where the AI-generated audio does not perfectly align with the speaker's mouth movements. A key finding is that GAN-based voice synthesis models introduce subtle spectral distortions that can be detected using mel-frequency cepstral coefficients (MFCCs), a feature commonly used in speech analysis. The authors emphasize that cross-domain forensic tools, which analyze both audio and video cues simultaneously, offer a higher detection accuracy than single-modality approaches. However, they acknowledge the challenge of detecting real-time deepfake streams, which require faster processing speeds and optimized AI models.
- **Mirsky et.al., (2021) – The Threat of AI-Generated Deepfakes in Cybersecurity.** This paper examines the cybersecurity implications of deepfakes, particularly their role in identity fraud, misinformation, and social engineering attacks. The authors categorize deepfake threats into three major types: Political and propaganda-based deepfakes, which are used to spread false narratives. Financial fraud deepfakes, where AI-generated voices or video

calls are used to impersonate executives in phishing scams. Personal identity theft deepfakes, in which manipulated media is used for extortion, defamation, or privacy violations. The paper presents several real-world case studies, including instances where deepfake voices have been used to deceive bank employees into approving fraudulent transactions. The study explores countermeasures, such as blockchain-based content authentication, digital watermarks, and AI-driven forensic models that can analyse inconsistencies in voice tone, facial movements, and eye reflections. A critical takeaway is that deepfake detection is not just a technical challenge but also a cybersecurity and legal issue, requiring international cooperation and regulatory frameworks. The paper highlights the need for automated forensic tools that can be integrated into digital platforms, law enforcement investigations, and corporate security systems. However, the authors acknowledge that deepfake technology is evolving rapidly, and current detection methods must continuously adapt to counter new threats.

- **Afchar et al. (2018) – Mesonet: A Lightweight Deepfake Detection Network.** This is a lightweight neural network designed specifically for real-time deepfake detection. The study highlights the challenge of balancing computational efficiency with detection accuracy, as many forensic techniques rely on deep convolutional networks that require substantial processing power. Mesonet is based on a shallow convolutional architecture, which allows it to analyse subtle inconsistencies in facial textures, lighting, and micro-expressions. One of the key findings is that GAN-generated faces often exhibit spectral anomalies—artifacts that are invisible to the human eye but detectable by neural networks. The researchers evaluate Mesonet against other state-of-the-art deepfake detection models and find that it outperforms heavier architectures in speed while maintaining competitive accuracy. However, they acknowledge that Mesonet struggles with low-resolution deepfakes and adversarially trained AI models, which require more advanced forensic techniques. This paper is significant because it underscores the need for efficient, deployable forensic tools that can be integrated into real-world applications such as social media platforms and law enforcement databases.

- **Guarnera et al. (2020) – Analysis of GAN Fingerprints in Deepfake Detection.** They focus on an emerging forensic technique known as GAN fingerprinting, which involves analysing unique digital traces left by different deepfake generation algorithms. Just as real cameras leave sensor noise patterns (e.g., PRNU noise) on captured images, GANs introduce systematic, model-specific artifacts during image synthesis. The authors propose a method to extract and classify these fingerprints by analysing the frequency domain of deepfake images. Their findings indicate that each GAN architecture (e.g., StyleGAN, ProGAN, CycleGAN) produces distinct artifacts, making it possible to identify which AI model generated a given deepfake. One challenge noted is that GAN fingerprinting can be bypassed through adversarial training, where deepfake creators intentionally modify images to remove telltale artifacts. Despite this, the study highlights forensic applicability, particularly in cases where law enforcement agencies need to trace the origins of manipulated content used in cybercrimes or disinformation campaigns.

- **Masi et al. (2020) – Face Recognition Systems Under Attack: Deepfake Spoofing and Countermeasures.** This paper investigates the vulnerability of face recognition systems to deepfake-based attacks. The authors demonstrate how attackers can use deepfake videos to bypass biometric security systems, such as facial authentication for banking apps, smart devices, and border control checkpoints. Their experiments reveal that state-of-the-art face recognition models can be fooled by high-resolution deepfakes, particularly those generated using face reenactment techniques. The study proposes several countermeasures, including: Liveness detection: Analysing eye blinking, pupil dilation, and micro-movements to determine if the subject is real. Depth and texture analysis: Detecting differences between real 3D faces and AI-generated 2D images. Temporal consistency checks: Identifying inconsistencies in facial movements across multiple frames. One of the key takeaways is that multi-factor authentication—combining face recognition with voice authentication or behavioural biometrics—significantly enhances security. However, the authors note that deepfake technology is evolving rapidly, and biometric security systems must continuously adapt to counter emerging threats.

- **Tolosana et al. (2020) – Biometric Deepfake Detection: A Review of Current Challenges and Future Directions.** These provide a comprehensive review of deepfake detection in biometric security, summarizing existing attack vectors, detection techniques, and forensic challenges. The paper categorizes deepfake-based attacks into three primary types: Impersonation attacks – where an attacker uses a deepfake video to mimic someone's identity. Identity swapping – where a person's face is replaced with another individual's in real-time. Synthetic identity creation – where AI generates an entirely fake person that does not exist (e.g., using StyleGAN). The authors discuss

several deepfake detection methods, including physiological signals analysis (e.g., heart rate detection from facial blood flow), deep learning-based approaches, and blockchain-based content verification. A key challenge highlighted in the paper is the arms race between deepfake creators and forensic researchers, where AI models are constantly improving to evade detection. The authors advocate for a collaborative approach, combining AI, law enforcement, and regulatory frameworks to combat deepfake-based fraud and misinformation.

- **Agarwal et al. (2019) – Deepfake Video Detection Using Physiological Cues.** They propose a novel forensic approach for deepfake detection based on physiological cues—specifically, analysing subtle facial and body movements that are difficult for AI to replicate. The study finds that deepfake videos often fail to capture involuntary physiological processes, such as: Eye movement patterns, including natural saccades and fixations. Subtle facial muscle contractions, which indicate genuine expressions. Breathing-related chest and shoulder movements. By analysing these micro-movements using motion magnification algorithms, the researchers demonstrate that deepfake videos lack the physiological realism of genuine human recordings. Their detection model achieves high accuracy rates on standard deepfake datasets, but they note that future deepfake models could potentially learn to mimic physiological cues, necessitating even more advanced forensic techniques. This research is significant because physiological-based detection methods provide an additional forensic layer beyond traditional image-based analyses.

- **Rossler et al. (2019) – Face Forensics++: A Large-Scale Dataset for Deepfake Detection** This paper introduces Face Forensics++, one of the largest publicly available deepfake datasets, designed to train and evaluate AI-based forensic models. The dataset includes over 1 million manipulated face videos, generated using popular deepfake tools such as Deep Face Lab, Face Swap, and Neural Textures. The authors systematically test various deepfake detection models on their dataset and identify key trends: CNN-based models outperform traditional forensic approaches when trained on high-resolution data. Detection accuracy drops significantly for low-resolution, compressed, and real-world social media deepfakes. Hybrid models combining spatial (image-based) and temporal (video-based) features achieve the best results. The paper emphasizes the need for continual dataset expansion, as deepfake techniques are evolving rapidly. By providing an open-access benchmark, Face Forensics++ has become a critical resource for forensic researchers worldwide, helping advance the development of AI-driven deepfake detection.

- **Verdoliva (2020) – Media Forensics in the Age of Deepfakes.** This paper presents a broad overview of media forensics, examining the historical evolution of digital manipulation techniques from simple photoshopping to advanced deepfake generation. The study discusses: Traditional forensic techniques, such as noise pattern analysis and metadata verification. AI-based forensic models, which leverage deep learning to identify manipulation artifacts. Emerging forensic challenges, including adversarial deepfakes designed to evade detection. One of the key insights is that no single forensic method is foolproof—instead, a multi-layered approach combining AI, metadata forensics, and blockchain-based verification is necessary. The author highlights the growing role of deepfake detection in legal cases, particularly in the context of evidence authentication and misinformation control.

- **Dolhansky et al. (2020) – The Deepfake Detection Challenge Dataset.** They introduce the Deepfake Detection Challenge (DFDC) dataset, a large-scale benchmark developed in collaboration with Facebook, Microsoft, and AI researchers. This dataset contains over 100,000 manipulated videos, designed to help researchers develop more robust deepfake detection models. The study highlights the limitations of current AI-based forensic methods, particularly their susceptibility to adversarial attacks and poor performance on low-quality deepfakes. The authors emphasize the need for real-world testing conditions, urging forensic researchers to develop detection models that work outside of controlled lab environments.

- **Korshunov et.al., (2019) – Vulnerability of Face Recognition Systems to Deepfake Attacks.** This study by Korshunov and Marcel explores the threat deepfake technology poses to facial recognition systems, particularly those used in biometric authentication. The researchers analyse how AI-generated face swaps and reenactments can be used to bypass security measures in banking applications, government IDs, and surveillance systems. They test various face recognition algorithms, including OpenFace, FaceNet, and Microsoft Azure Face API, against a dataset of deepfake videos and find that even state-of-the-art biometric systems struggle to detect high-quality deepfakes. One key insight is that deepfake models are becoming increasingly realistic, with AI now able to mimic micro-expressions, subtle lighting variations, and skin texture details. The study proposes multiple countermeasures, such as: Liveness detection – Identifying real vs. synthetic faces by detecting blinking patterns and facial muscle activity. Multi-modal

authentication – Combining facial recognition with voice analysis, fingerprint scanning, or iris detection to strengthen security. Adversarial training – Continuously updating AI detection models to recognize the latest deepfake techniques. A major challenge noted in this research is the arms race between deepfake creators and forensic experts. As detection models improve, deepfake generation models also evolve to become more resistant to forensic scrutiny. The study emphasizes that face recognition systems must integrate adaptive defences to keep pace with advancements in AI- generated forgery.

- **Cozzolino et al. (2019) – Forensic Transfer Learning for Deepfake Detection.** They investigate the use of transfer learning—a technique where a pre-trained AI model is adapted to a new forensic task—to improve deepfake detection accuracy. The authors argue that most deepfake detection models are trained on a single dataset, making them ineffective when tested on new deepfake techniques. To address this, they develop a forensic transfer learning framework that allows detection models to learn from multiple datasets simultaneously, improving their ability to generalize. Their approach involves: Training a baseline model on a large deepfake dataset (e.g., FaceForensics++). Fine-tuning the model using real-world deepfake videos from social media and surveillance footage. Testing the model against unseen deepfake techniques to evaluate its robustness. The results show that transfer learning significantly improves detection performance, especially against low-quality, compressed deepfakes—a major limitation of previous forensic models. However, the authors acknowledge that continuous retraining is necessary, as deepfake technology evolves rapidly. They suggest integrating real-world forensic case studies into training datasets to ensure AI detection systems remain adaptive and up-to-date.
- **Haliassos et al. (2021) – All You Need is Real: How Synthetic Data Improves Deepfake Detection.** They present a novel approach to training deepfake detection models using synthetic data rather than relying solely on real-world deepfake videos. They argue that most deepfake detection datasets contain a limited number of AI- generated forgeries, which restricts the model's ability to recognize new manipulation techniques. To overcome this, they develop a synthetic training pipeline, where AI- generated deepfake images are used to train detection models before testing them on real-world data. Their key findings include: Models trained on a mix of synthetic and real deepfakes perform better than those trained on real data alone. Synthetic deepfakes help expose forensic models to a wider range of manipulation artifacts, making them more resistant to adversarial attacks. This approach reduces the need for manually collecting and labelling new deepfake datasets, saving time and resources. The study highlights that synthetic data generation can serve as a powerful tool in forensic AI development, but also warns that deepfake creators can use similar methods to generate detection-resistant forgeries. The authors emphasize the need for a proactive forensic strategy, where detection models evolve in parallel with deepfake generation technology.
- **Dang et al. (2020) – Detection of Deepfake Videos Using Temporal Inconsistencies.** They propose a video-based forensic approach that focuses on temporal inconsistencies—errors that appear across multiple frames in deepfake videos. They note that most deepfake detection models analyse individual frames rather than entire video sequences, making them vulnerable to deepfake techniques that focus on perfecting single-frame synthesis. Their methodology involves: Analysing frame-to- frame facial movements to detect unnatural distortions. Examining inconsistencies in lip-syncing and speech patterns, which often occur in deepfake videos. Using optical flow analysis to identify unnatural transitions in facial expressions. One of the major findings is that deepfake models struggle to maintain realistic motion over time, especially when dealing with subtle head movements and rapid changes in facial expressions. Their detection model achieves higher accuracy than traditional CNN- based approaches, particularly for deepfake videos with lower frame rates. However, the authors warn that future deepfake models could improve motion realism, making it necessary to continuously refine forensic methodologies.
- **Qi et al. (2020) – Deep Rhythm: Deepfake Detection via Physiological Signal Analysis.** Deep Rhythm, a forensic model that detects deepfake videos by analysing physiological signals such as facial blood flow and pulse rhythms. Their research is based on the premise that real human faces exhibit subtle colour changes due to blood circulation, which deepfake models fail to replicate accurately. Their approach includes: Extracting heart rate signals from facial regions using remote photoplethysmography (rPPG). Comparing the extracted signals with known physiological patterns to detect inconsistencies. Testing the model against multiple deepfake datasets to evaluate its robustness. The results show that Deep Rhythm outperforms conventional image-based deepfake detection methods, especially in low- resolution and compressed videos where traditional CNNs struggle. However, the authors acknowledge that some deepfake models are beginning to integrate physiological cues, meaning future forensic techniques will need to combine multiple detection strategies to remain effective.

- **Nirkin et al. (2019) – Deep Fake Detection by Detecting Face Warping Artifacts.** They focus on one of the most common forensic weaknesses in deepfake videos—face warping artifacts. These artifacts occur when a deepfake model struggles to align the AI-generated face with the target person's head, resulting in distorted facial boundaries, unnatural blending, and edge inconsistencies. Their detection method involves: Analysing the geometric structure of faces to detect warping distortions. Using deep learning models trained on real vs. warped face pairs to recognize unnatural facial transformations. Evaluating detection performance on both low-resolution and high-resolution deepfake datasets. One major advantage of this approach is that face warping artifacts are present in nearly all deepfake videos, making them a reliable forensic marker. However, the authors note that as deepfake models improve in generating seamless face swaps, warping-based detection methods may become less effective.
- **Sabir et al. (2019) – Recurrent Convolutional Strategies for Deepfake Video Detection.** They propose a hybrid deep learning model that combines CNNs (Convolutional Neural Networks) with RNNs (Recurrent Neural Networks) to enhance deepfake detection accuracy. While CNNs are highly effective at detecting spatial manipulation artifacts in images, they struggle with temporal inconsistencies in videos. By integrating RNNs, which excel at processing sequential data, the model is able to detect deepfake anomalies across multiple frames. Their key findings include: Hybrid CNN-RNN models significantly outperform standalone CNN-based detection systems. Temporal inconsistencies in deepfake videos are more pronounced in facial regions, particularly around the eyes and mouth. Their approach is more robust against adversarial attacks that specifically target CNN vulnerabilities. This study emphasizes the importance of multi-layered forensic strategies and highlights that combining multiple AI architectures can improve detection resilience against evolving deepfake techniques.
- **Zhao et al. (2021) – Learning Rich Features for Deepfake Detection.** They introduce a deep learning model designed to extract rich, multi-scale features from deepfake images and videos. They argue that existing deepfake detection models rely heavily on pixel-level artifacts, which are often removed through post-processing techniques such as blurring, compression, and upscaling. To address this issue, the researchers develop a multi-scale feature extraction network that captures high-level semantic information as well as low-level texture details. Their methodology includes: Using a deep convolutional neural network (CNN) trained on multiple datasets to recognize both spatial (single-frame) and temporal (multi-frame) inconsistencies in deepfake content. Integrating attention mechanisms to focus on facial regions prone to manipulation, such as the eyes, mouth, and skin texture. Testing the model against adversarially modified deepfakes, where distortions have been intentionally minimized to bypass detection. The results show that the multi-scale feature extraction approach significantly improves detection accuracy, especially for low-quality deepfakes commonly shared on social media platforms. However, the study also highlights a major challenge: as deepfake technology advances, generative models become better at mimicking both high- and low-level image details, making it increasingly difficult to identify manipulation. The authors suggest that future forensic research should explore hybrid approaches combining feature extraction with physiological-based detection techniques.
- **Masood et al. (2021) – A Survey on Deepfake Detection: Taxonomy, Challenges, and Future Directions.** They conduct a comprehensive survey of existing deepfake detection techniques, categorizing them into five main approaches: Pixel-based detection – Identifying anomalies at the pixel level, such as compression artifacts, colour inconsistencies, and blending distortions. Frequency-based detection – Analysing Fourier transformations to detect unnatural signal patterns introduced by deepfake models. Physiological-based detection – Tracking subtle biometric signals like eye blinking, pulse fluctuations, and facial heat maps. Deep learning-based detection – Using convolutional and recurrent neural networks to recognize manipulation patterns in large datasets. Hybrid detection techniques – Combining multiple approaches to improve detection robustness. The survey highlights key challenges in deepfake detection, including: Generalization issues – Many forensic models struggle when tested on deepfake videos generated using unseen techniques.

Adversarial resistance – Attackers are now developing adversarial deepfakes specifically designed to bypass detection models. Data scarcity – High-quality deepfake datasets are limited, making it difficult to train detection algorithms effectively. The authors emphasize that no single detection method is foolproof, and future research should focus on adaptive forensic techniques that evolve alongside generative AI advancements.

- **Wang et al. (2020) – Deepfake Video Detection Using Biological Signals.** Wang et al. propose a

biologically inspired approach to deepfake detection, focusing on subtle physiological signals that AI models struggle to replicate. Their hypothesis is that human faces exhibit natural blood flow patterns, visible as minute colour fluctuations in the skin, which deepfake models fail to accurately reproduce. Their approach involves: Using remote photoplethysmography (rPPG) to extract blood flow patterns from facial regions. Developing a deep learning model trained to distinguish real vs. synthetic physiological signals Evaluating the model's performance on deepfake datasets, including Face Forensics ++ and Deep Fake Detection Challenge (DFDC) videos. The results demonstrate that physiological-based detection methods outperform traditional CNN-based models, particularly when tested against highly compressed and low-resolution deepfakes. However, the study also acknowledges potential limitations: Deepfake creators could begin integrating synthetic physiological cues into their models, reducing the effectiveness of rPPG-based detection. Variability in human skin tone, lighting conditions, and camera quality can impact the accuracy of biological signal extraction. The authors suggest that combining physiological-based detection with motion and texture analysis could lead to more robust forensic strategies.

- **Rössler et al. (2019) – Face Forensics++: A Large-Scale Dataset for Face Manipulation Detection.**

Rössler et al. introduce Face Forensics++, one of the largest and most widely used datasets for deepfake detection research. They recognize that one of the biggest challenges in forensic AI is the lack of diverse, high-quality training data, which limits the effectiveness of detection models. To address this, they create a dataset containing thousands of manipulated videos generated using multiple face-swapping techniques. Key features of Face Forensics++ include: Four types of deepfake manipulations – The dataset includes videos created using Deepfakes, Face2Face, Face Swap, and Neural Textures. Different levels of compression – Videos are provided in both high and low quality, allowing forensic models to be tested against real-world conditions (e.g., social media compression). Benchmarking framework – The dataset includes a standardized evaluation pipeline for testing the accuracy of deepfake detection models. The authors conduct extensive experiments using CNN-based detection methods, demonstrating that while state-of-the-art models achieve high accuracy on high-resolution deepfakes, their performance drops significantly when tested on low-quality or adversarially modified videos. This study underscores the importance of developing deepfake detection methods that generalize well across different datasets and manipulation techniques.

- **Dolhansky et al. (2020) – Deep Fake Detection Challenge Dataset (DFDC).**

Dolhansky et al. present the Deep Fake Detection Challenge (DFDC) dataset, a large-scale benchmark created to advance forensic AI research. They note that while datasets like Face Forensics++ provide a strong foundation, they primarily feature AI-generated deepfakes from a limited number of models. The DFDC dataset aims to provide greater diversity by incorporating deepfakes from multiple generation techniques and varying real-world conditions. Key contributions of the DFDC dataset include: Over 100,000 deepfake videos – Making it the largest publicly available deepfake dataset at the time of release. Diverse manipulations – Featuring multiple AI architectures, different lighting conditions, and various levels of facial occlusion. High variation in subjects – Including individuals of different ages, ethnicities, and facial structures to improve the robustness of forensic models. The authors conduct baseline experiments using multiple deep learning models, including ResNet, EfficientNet, and XceptionNet, demonstrating that even the best forensic models struggle with certain deepfake variations. They conclude that future research should focus on developing adaptive, multi-modal detection strategies to address the growing complexity of deepfake manipulations.

- **Cozzolino et al. (2018) introduce Forensic Transfer, a weakly supervised domain adaptation method that enhances deepfake detection by improving model generalization across different deepfake datasets.** By leveraging multi-layer feature extraction, their model adapts to new deepfake techniques with minimal labelled data. Similarly, Agarwal et al. (2020) explore biometric-based detection, focusing on facial artifacts and unnatural head movements to distinguish real from fake videos. Their approach highlights deepfake inconsistencies in eye gaze, skin textures, and head motion, making it particularly effective for detecting manipulated content. Expanding on physiological-based detection, Tolosana et al. (2020) introduce DeepFakesON-Phys, which analyses biological signals such as heart rate fluctuations and breathing patterns extracted from facial videos. Their findings reveal that while physiological cues improve deepfake detection, lighting variations can distort these signals, limiting their effectiveness in certain scenarios.

- A broader forensic overview is provided by Verdoliva (2020), who categorizes deepfake detection techniques into image-based, video-based, physiological-based, and metadata-based methods. The study emphasizes the need

for multi-modal forensic approaches and highlights challenges posed by adversarial deepfakes designed to evade detection. Meanwhile, Korshunov & Marcel (2018) investigate the vulnerability of face recognition systems to deepfake attacks, demonstrating that AI-generated identity spoofing can bypass many commercial biometric authentication systems. Their findings suggest that integrating liveness detection and depth sensing is crucial for future security systems. To support forensic research, Jiang et al. (2020) introduce DeeperForensics-1.0, a highly realistic deepfake dataset designed to improve model robustness by incorporating varied compression levels and diverse demographics. Their results reveal that existing forensic models struggle to detect high-quality deepfakes, underscoring the need for more adaptive detection techniques. Similarly, Li et al. (2020) introduce Celeb-DF, a benchmark dataset containing highly refined deepfakes with minimal artifacts. Collectively, these papers highlight the urgent need for adaptive, real-time deepfake detection strategies to combat misinformation, identity fraud, and AI-driven deception

Chapter 3

OBJECTIVES:

- To study the concept and technological foundations of deepfake technology and synthetic media generation.
- To analyse the role of artificial intelligence and machine learning techniques in deepfake detection and digital forensic investigations.
- To compare the performance of different forensic software tools used for identifying manipulated videos, images, and audio content.
- To evaluate deepfake detection techniques based on parameters such as accuracy, precision, processing efficiency, and false positive rates.
- To identify the challenges and limitations associated with existing deepfake detection systems and forensic analysis methods.
- To understand the impact of deepfake technology on cyber security, digital evidence authenticity, and online misinformation.
- To explore the importance of hybrid and multi-modal forensic approaches for improving the reliability of deepfake detection systems.

Chapter 4

METHODOLOGY:

4.1 Total Samples

- The dataset contains 200 samples of videos divided into three groups in order to provide a balanced and comprehensive dataset for research into detecting deepfakes. The groupings ensure that the dataset includes a variety of manipulations in order to support thorough research for evaluating detection tools and techniques.

AI-Generated Deepfakes (100 samples) These deepfake videos were acquired from publicly available datasets and created using deepfake generation software. The purpose of this group is to provide a breadth of AI-generated content that incorporates diverse deepfake fingerprint signatures and sophistication levels.

- Source: Public deepfake repositories and manually created deepfake videos.
- Variation Conditions:
 - Different algorithms of AI algorithms used— for synthesis: e.g., generative adversarial networks

(GANs) and autoencoders.

- Variations of lighting, facial expression, head movement, and backgrounds.
- Variations in compression and resolution, affecting artifacts.
- Voice-cloning deepfakes to assess whether facial and audio were inconsistent when evaluating and assessing deepfakes.
- Level of synthetic rendering quality, e.g., low fidelity vs. high resolution, hyper-realistic modifications.
- Purpose: To investigate the differing levels of manipulation in the context of deepfake videos, which would enable comparison of detection tools and understanding of detection, i.e., simple face swap versus more complex deepfake reenactments. or voice synthesis.

4.1.1 Authentic (Real) Videos (50 samples)

This category consists of real, unchanged videos that can be used as ground truth for evaluating deepfake detection algorithms. It is important that the dataset includes high-quality and representative real-world examples for training machine learning models and benchmarking forensic algorithms.

- Source: Verified repositories that include real videos from news clips, interviews, and social media. Included videos should feature proper authenticity verification.
- Selection Criteria:
 - Representative of varied demographic cues, including a variety of ethnicities, ages, and genders to prevent biases within the dataset.
 - Diversity regarding speech patterns, emotional responses, and in-context environmental conditions to have a fidelity test in the real-world application of models.
- Variability in different video compressions and recording devices to illustrate presentation variability in user-generated content.
- Purpose: To use as a baseline comparison to deviant videos from the edited and deepfake categories to ensure a distinct difference between naturally generated movement, expressions, and speech, with synthetic.

4.1.2 Manually Edited Videos (50 samples)

The featured videos were manipulated through standard video editing applications to create distortions and artifacts which may resemble discrepancies associated with deepfakes. This category really helps to identify traditional video manipulation and AI deepfake technologies for forensic comparison.

- Video editing techniques used:
 - Compression artifacts and inconsistent pixelation, altering the appearance of the naturalness of facial features.
 - Frame adjustments including frame blending, ghosting effects, and time interpolation artifacts.
 - Altered lighting effects, contrast, and colour grading may mask subtle deepfake artifacts.
 - Audio out of sync and synchronization may simulate dissimilarities considered deepfake artifacts.
- Purpose: To examine how traditional video editing applications influence videos' integrity and to compare their instances to deepfakes. This category can also assist with developing models to differentiate traditional video manipulations from an AI-based fabrication.

Sources of Data

To ensure dataset diversity and authenticity, the data were collected from multiple sources, ensuring a wide representation of real and manipulated content.

4.1.3 Public Deepfake Datasets

Deepfake repositories offered a range of pre-generated synthetic media content:

- Deepfake Detection Challenge (DFDC): One of the largest publicly available datasets for deepfake research, depicting a multitude of content with a variety of authentic or AI manipulated form.

- Face Forensics: A benchmark dataset of manipulated videos created with several AI- developed techniques, it has become a predominant dataset used by researchers in deepfake detection studies.
- Celeb-DF: A dataset containing high-level deepfake videos with minimal compression loss of resolution, it allows high-definition synthetic media content to be good on display.
- Other Available Public Repositories: Further sources of synthetic media utilizing various forms of deepfake, using experimental datasets, commonly seen in academic papers from the AI community, mined online.

Online Platforms

Videos were also gathered from:

- YouTube: Public videos displaying real and altered videos under differing light and environmental conditions, including interviews, news videos, and videos of deepfake challenges shared by AI researchers.
- Other Open-Access Repositories: There was both real and AI-generated content to diversify the dataset so it included a mix of both professionally edited content and user- generated content.

4.1.4 Manually Generated Deepfakes

To maintain dataset completeness, we generated further deepfake samples in controlled conditions with the following AI tools:

- DeepFaceLab: Open-source software for deepfake video generation, that can generate precise swapping of a face using video synthesis.
- FakeApp: A tool for swapping faces in videos, commonly used in early deepfake generation.
- Faceswap: An additional open-source application designed for high-quality manipulation of faces using advanced neural networks.

In each video, we included metadata for:

- The date and source of acquisition.
- Duration of the video including its resolution settings.
- Video compression and encoding characteristics, ensuring a wide range of video quality, from low-bitrate to high-definition.
- Authentication verification using forensic tools to establish the dataset's integrity.

4.2 Tools and Software Used

A combination of manual forensic tools and AI-based software was employed to detect anomalies and validate the authenticity of videos.

Manual Tools for Visual and Metadata Analysis

For detecting subtle differences in manipulated video, manual analysis is an essential process. The following tools were utilized:

- VLC Media Player: The VLC media player was used for analysis of videos frame-by- frame, playback in slow motion, and observation of unnatural motion patterns that demonstrated potential synthetic alterations.
- Focus Magic: A sharpening tool used to reveal details in blurry images. It can be useful for revealing blended inconsistencies, edge transitions, and unnatural disturbances commonly associated with deepfake videos.
- Wondershare Filmora: This is a video editing software, leveraged for metadata extraction, the analysis of compression artifacts, and recognition of anomalies from video encoding that might reveal manipulations.

This program also provides opportunities for examining artifacts of video re-encoding.

AI-Based Detection Tools for Automated Analysis:

Sophisticated advanced AI-based detection and forensic tools were employed for deepfake detection and enhancement of analysis:

- **Deep ware Scanner:** An AI-based detection tool able to analyse manipulated features of videos altered by deepfake technology. This tool is particularly focused on facial inconsistencies and other patterns based on AI generation.
- **Sensity AI:** A forensic analysis tool for recognizing synthetic media. This tool uses analysis based on the evaluation of the integrity of video frames, but acknowledges and details inconsistencies. It also uses deep learning and neural networks for analysis as opposed to the traditional forensic approach of determining what is real versus the synthetic media.

4.2.1 Forensic Workflow and Validation

Each tool was used in a systematic forensic workflow:

1. **Initial Screening:** Manual inspection using VLC Media Player for motion inconsistencies and unnatural transitions.
2. **Metadata Extraction:** Wondershare Filmora was used to extract metadata and analyse encoding discrepancies that could indicate manipulation.
3. **Edge and Blending Analysis:** Focus Magic enhanced image sharpness to detect unnatural transitions in facial regions.
4. **AI-Based Verification:** Deep ware Scanner and Sensity AI were employed to cross-check results and validate authenticity using machine learning models.
5. **Final Validation:** Cross-referencing all findings with forensic principles to ensure accuracy and consistency in deepfake detection.

4.3 Step-by-Step Deepfake Detection Process

Step 1: Preprocessing and Data Organization Data Categorization

To create a comprehensive and balanced dataset, all video samples were organized into three main categories:

1. **Deepfake Videos:**
 - AI-generated synthetic videos that use deep learning-based face swapping techniques.
 - These videos were acquired from known deepfake repositories, datasets developed for research, and videos produced using commercial or open-source applications (e.g., Deep Face Lab, Face Swap, Reface).
 - They often display artifacts related to blinking that are unnatural, facial movements that do not make sense (e.g., two people speaking), and a lack of blending artifacts.
2. **Authentic Videos:**
 - Unaffected, real-world footage obtained from established sources (e.g., government archives, verified social media posts, news agencies, etc.).
 - These videos were treated as baseline comparisons to identify deepfake artifacts.
 - They were guaranteed to not have been manipulated in post-production.
3. **Manually Edited Videos:**
 - Videos manipulated using traditional editing software (e.g., Adobe Premiere Pro, Final Cut Pro), but not synthetically altered to resemble a deepfake.

- This included colour grading, morphing, and state-by-state manipulation.
- These videos served to help to distinguish artifacts specific to deepfakes from video manipulation.

Format Standardization

Format Standards Standardizing video formats was essential to ensure homogeneous processing and eliminate inconsistencies that may impact analysis. The following actions were pursued:

1. **File Conversion:** All videos were converted to MP4 files using FFmpeg to maintain consistency across samples. MP4 file formats guarantee compatibility with forensic analysis tools and AI-based detection models
2. **Frame-rate normalisation:** Controller to be a uniform 30 frames per second (fps) to align playback speeds and produce reliable artifact detection. Maintains consistency through deepfake and authentic videos to mitigate anomalies created by differing frame rates.
3. **Encoding standardization:** H.264 codec was enforced across all samples to maintain uniform compression and encoding characteristics. Ensures a consistent baseline for comparing compression artifacts.
4. **Resolution retention:** Original resolution was retained so as not to lose detail when conducting forensic analysis. Avoided downscaling or upscaling to maintain original pixel integrity and detect deepfake artifacts that may be more subtle.

File Labelling and Metadata Documentation:

Every video received a unique identifier in order to systematically track and analyze the videos. Metadata documentation included:

- **Source and Date of Collection:** To provide credibility and traceability.
- **Duration and Resolution of the Video:** To allow for consistent comparative analysis,
- **Frame Rate and Bitrate Data:** To help identify encoding anomalies.
- **Encoding Parameters (e.g., Codec Type, Compression Rate, GOP Structure):** Used for detecting deepfakes due to compression artifacts.
- **Any Additional Editing or Processing:** Documentation of anything that we know was edited or processed to understand the effect on the forensic evaluation.
- **Checksum Hashing (MD5/SHA-256):** To ensure file integrity and protect against unauthorized manipulation.

A structured repository was established to manage all metadata. The repository permitted easy access to metadata documentation for cross-referencing, validation, or storing purposes. A relational database was implemented to support indexing, retrieving, and filtering of video records based on specific attributes (source authenticity, encoding patterns, and detection confidence levels).

Step 2: Frame by Frame Evaluation in VLC Media Player

A thorough manual review of each video was conducted using VLC Media Player for frame by frame examination to address inconsistencies associated with deepfake manipulation. Slow Motion Review

1. **Playback Speed:**
 - Videos were played 0.25x speed to address subtler inconsistencies.
 - Slowed video occurred distortions that may not be visible during playback.

2. Facial Movements:

- Addressed unnatural facial movement rigor.
- Evaluated for delays in movement or change in facial expressions that suggests AI had an influence.

3. Micro Movements:

- Reviewed for eye movement, small muscle twitches, or micro-expressions on the face.
- The inconsistency of natural micro-movements remains one of the strongest indications of synthetic media.

Keyframe Inspection

Keyframes were pulled and scanned for possible deepfake cues. The following anomalies were evaluated:

1 Lip sync anomalies:

- Looked for discrepancies between spoken text and mouth movements.
- Assessed phoneme articulations, disturbances in articulation from frame to frame.

2 Blinking anomalies:

- Noted severely short blinks and methodically long blinks, etc.
- A lot of AI-generated videos don't match natural blinking.

3 Facial distortion.:

- Recorded warping, stretching, or unnatural facial/skin textures.
- Noted blending problems with the face/environment.

4 Lighting and shadows inconsistency:

- Discrepancies with the lighting to the face / lighting to the backdrop.
- Note if the shadows were behaving strangely or non-existent shadows.

5 Motion artifacts:

- Noted blurriness or ghosting in /around the face areas.
- Examined jerking transitions from any overlays from the deepfake Documentation and annotations.
- Any frame that displayed notes time-stamped.
- Kept comparative screenshots of issues for forensic references.
- Each issue was then categorized on lesser to most consequential and likelihood that it was a deepfake.
- Data would support future validation processes and be available to validate during lesions.

Step 3: Image Enhancement and Artifact Detection (Focus Magic)

To increase the visibility of deepfake artifacts, forensic practitioners used Focus Magic, a software developed to restore detail and add sharpness to images for purposes of forensic investigations. Extraction of Targeted Frames

1. Criteria for Frame Selection:

- Extracted frames that displayed rapid 'facial movements,' eye blinks, and sudden changes in expression
- Asked users to pay particular attention to transition frames in which deep fake artifacts are most likely to be apparent.

2. Selection of Keyframe:

- Prominent keyframe that indicated distortions in facial features; choose with other potential deepfake artifacts such as blending colours.

- Pulled stills of moving images for continued digital enhancement Techniques for Enhancing Images

1. Edge Detection and Sharpness:

- When applying sharpeners, highlights the unnatural smoothing and gloss effect to deep fakes.

- Provided presents short details allowing for rapid detection of facial split.
- 2. Blending Artifacts:
 - Examined blending areas where synthetic overlay meets authentic facial features.
 - Evaluated artifacts that exists at edge boundaries and borders of a hairline or colour transfer on the face itself
- 3. Colour Correction and Gradient Behaviour:
 - Evaluated portions of the face where face components are mismatched with neighbouring tones of the skin colour.
 - Evaluated portions of the face where gradient shifts exhibited signs of errors in blends or deformity caused by digital alteration
- 4. Texturing:
 - Used high-resolution filtering to identify digital distortion through noise and artificial blurring.
 - Compared unexpected texture differences with true video to reveal fine differences

Evaluation against Original Videos:

- Evaluated the enhanced frames against those from the original video sources.- Noted differences in facial alignment, colour grading, and texture consistency.
- Validation through Alternate Enhancement Techniques- Performed independently generated analysis with alternative forensic tools for validation.- Combined reports to provide a cohesive profile of deepfake artifacts.
- Metadata and Compression Analysis (Wondershare Filmora & FFmpeg) Deepfake videos show uniquely identifiable compression aberrations through metadata and compression artifact detection.
- Compression aberrations arise from frames synthesized via AI, unlike "natural" video shot and typical post-production encoding techniques.

Metadata Extraction and Analysis

Metadata provides necessary technical aspects of the video file, such as encoding parameters, bitrate, frame rate, and codec type. By analysing the metadata, forensic experts will identify discrepancies that may indicate the file has been manipulated to create a deepfake. The analysis includes utilizing Wondershare Filmora (a video editing application that facilitates analysing the metadata) and FFmpeg (a command-line multimedia framework).

Key forensic criteria assessed:

1. Codec Type, Bitrate and Frame Rate:
 - Deepfake videos may have abnormal encoding values in relation to real videos.
 - AI-generated videos utilize various compressions in contrast to the compression that camera recorded videos utilize such that these discrepancies can indicate the computer-generated media.
 - An escalation in the value of the bitrate or a rapid fluctuating frame rate may show the use of synthetic frames.
2. Encoding Parameters:
 - The asset encoding parameters such as H.264 profile levels, quantization parameters (QP) which refers to the ratio of pixel information of the processed frames can identify AI synthesized videos.
 - The absence of encoding agreement between deepfake and real video suggests that the video file has been manipulated in some manner.
3. Frame Structure evaluation:
 - Deepfake rendering errors will cause anomalies in GOP (group of pictures) sequencing that may upset or create differences in the natural I-frame and P- frame flow of encoding.

- AI media could display irregularities in B-frame placement which causes inconsistencies in the visual movement of the frames. Through assessing the parameters of all after mentioned metadata for each file, forensic experts may identify if the video has undergone deepfake synthesis.

Detection of Compression Artifact

Becoming Increasingly Effective Video footage created by AI often incurs compression artifacts that can be uncovered fairly easily with frame-by-frame forensic tools. Unlike real- world videos, deepfake models do not always accurately replicate natural motion and texture, which can lead to artifacts. The following compression-related forensic markers are significant:

1. Structure of GOP:

- Video compression typically follows a normal and logical GOP structure (IBBPBBP, etc.).
- Deepfake videos may lead to unusual GOP lengths, showing inconsistencies in the placement of I-frames due to limitations in the AI processes.
- Variations in motion-compensated frame prediction are a sign of synthetic manipulation.

2. Bitrate Fluctuation:

- Normal recordings in video are displayed with a consistent bitrate distribution as the video is played.
- Deepfake recordings could lead to unexpected spikes or drops in video bitrate due to AI-based interpolation.
- Recording bitrate fluctuation can be measured on a bitrate graph and can show suspicious encoding behaviour from the original footage.

3. Motion Vector Disruption:

- Motion vectors track object motion from frame to frame and used for video compression, but may behave unusually in the Deepfake video.
- Motion transitions created by AI can easily lead to strange or jerky motion paths.
- Looking at the vectors frame by frame can suggest incoherent motion tracking, producing a difference in the deepfake clips versus real original videos.

Forensic Documentation

A complete report documenting the existence of metadata anomalies and compression discrepancies is generated, outlining:

- Results of the extraction of metadata (codec, bitrate, frame structure, encoding parameters) An analysis of compression artifacts, identifying abnormal deviations in GOP structure, and motion vectors.
- Graphical visualizations of bit rate differences and motion inconsistencies.
- Expert conclusions as to whether the anomalies are indications of deepfake manipulation. This forensic documentation will serve as a reference for other legal action or peer corroboration.

Step 5: Comparative Study with AI-Based Detection Models

To enhance the method of assessing forensics manually, advanced AI-based deepfake detection tools are used to analyse videos for possible tampering. These models can identify subtle pixel- level manipulations that a human may not even be able to perceive. AI Model Processing AI tools like Deep ware Scanner and Sensity AI are used to determine whether a video has been manipulated by AI.

1. Deep ware Scanner:

- Employs deep-learning models to scan video for AI-generated facial synthesis artifacts.
- Analyses and generates a deepfake probability score to capture the likelihood of manipulation.
- Analyses for morphing inconsistencies, unnatural facial textures, and lip synchronization

discrepancies.

2. Sensity AI:

- Conducts forensic-level video analysis to identify synthetic overlays and pixel manipulations.
- Utilizes neural networks trained on deepfake datasets to detect the presence of tampering.
- Analyses for face swaps, lighting inconsistencies, and warping that may alter facial appearance.

3. Hybrid Analysis:

- Obtains outputs from multiple AI models in order to increase detection accuracy.
- Combines outputs from assessments utilizing forensic analysis (compression, metadata) and AI-generated probability scores.
- Ensures that deepfake detection is strong and cross validated across multiple detection modalities.

Validation of AI Findings and Manual Forensic Findings

To improve accuracy, forensic specialists validate the results of AI-based detection against manual forensic findings. Validation of AI results does the following:

1. False Positives:

- Instances where genuine videos are incorrectly identified as deepfakes.
- This can occur for a number of reasons, such as compression artifacts, noise, or camera low-fidelity distortions.
- Meaning that an AI may misidentify a genuine video, but manual validation solves the mistake.

2. False Negatives:

- Instances where manipulation occurs, but AI does not detect deepfake evidence.
- This occurs due to AI not being capable of detecting subtle manipulation techniques.
- In these situations, forensic experts perform a frame-by-frame assessment of videos to determine whether the AI was able to detect deepfake evidence or not. This process allows for evidence to not solely rely on AI decision making, and relies on expert witness forensic evaluation.

Performance Benchmarking

AI detection is benchmarked against manual forensic analysis to accurately evaluate whether the AI is reliable. This assessment takes the following in consideration:

1. Sensitivity and Specificity:

- Sensitivity: Sensitivity is the measure of AI's ability to detect manipulations in deepfakes.
- Specificity: Specificity measures the AI's ability to make a distinction between deepfakes to real videos.
- This measure is only concerned with limiting the number of erroneous classifications and ensuring that AI detection models will report correctly, even if the image identification is inaccurate.

2. Consistency:

- AI detection has to revert to consistency when detection is made in relation to different deepfake sources.
- Forensics will again validate whether the AI performs to the same standard across datasets.
- AI models and inconsistencies with forensic assertions can create opportunities for improvement. Through performance benchmarking, forensic teams refine their detection methodologies, ensuring that deepfake identification remains accurate and legally admissible.

Chapter 5

Materials Required:

1. Deepfake Videos

Purpose: Deepfake videos serve as test samples to evaluate the performance of detection tools. These videos are AI-generated and manipulated using deep learning techniques to alter facial features, expressions, lip-syncing, or entire identities.

Sources & Collection:

- Public datasets such as DFDC (Deepfake Detection Challenge), Face Forensics++, Celeb-DF, and DeeperForensics-1.0.
- Self-generated deepfakes using software like Deep Face Lab, Reface, Face Swap, and ZAO.
- Social media platforms where deepfakes are being shared (with ethical considerations).

Features of Deepfake Videos:

- Face Swaps: Replacing one person's face with another.
- Lip-sync Manipulation: Altering the mouth movements while keeping the rest of the video unchanged.
- Expression Manipulation: Changing emotions in real-time, e.g., smiling instead of frowning.
- Body Swaps: Replacing entire body movements.

2. Authentic Videos (Ground Truth Samples)

Purpose: Genuine, unaltered videos are required as baseline references for comparison against deepfake and manually edited videos. They help in training AI models to distinguish between real and synthetic content.

Sources:

- Public video datasets such as UCF-101, YouTube-8M, and VGGFace2.
- Personal recordings under controlled conditions to ensure originality.
- News broadcasts and official interviews with known public figures.

Features of Authentic Videos:

- Naturally occurring facial micro expressions, lighting inconsistencies, and motion blur.
- No abrupt changes in skin texture, eye blinking, or unnatural facial movements.

3. Manually Edited Videos

Purpose: These videos are altered using traditional video editing techniques rather than AI-driven deepfake generation. They help differentiate between conventional manipulations and AI-generated deepfakes.

Methods Used for Editing:

- Frame-by-frame alterations using software like Adobe Premiere Pro or Final Cut Pro.

- Morphing tools such as Fanta Morph to manually create facial distortions.
- Colour correction and lighting changes to simulate environmental effects.
- Audio dubbing and voice cloning using software like Descript and Voicemod.

Importance in Research:

- Helps to understand the differences between AI-generated deepfakes and human-edited manipulations.
- Some manual editing techniques (e.g., motion blur) can make AI deepfake detection harder.

4. VLC Media Player

Purpose: VLC serves as a video playback tool for manual inspection of deepfake artifacts. Features Useful for Deepfake Detection:

- Frame-by-frame playback: Allows detailed analysis of video inconsistencies.
- Zoom and slow-motion features: Helps identify unnatural eye blinks or face morphing.
- Codec information: Displays encoding details that might indicate tampering.
- Colour adjustments and sharpening: Enhances low-quality areas where deepfake signs are visible.

Alternative Tools:

- DaVinci Resolve (for advanced frame analysis).
- FFmpeg (for extracting frames and metadata).

5. Focus Magic (Forensic Image Processing Tool)

Purpose: Focus Magic is a forensic deblurring and enhancement tool used to recover details from blurry or pixelated deepfake frames.

Applications in Deepfake Detection:

- Enhancing facial features: Helps detect minor distortions in AI-generated faces.
- Motion deblurring: AI-manipulated faces often contain motion inconsistencies.
- Forensic sharpening: Reveals edge inconsistencies in synthetic images.

Alternative Tools:

- Forensic Photoshop Plugins (e.g., Amped FIVE).
- Topaz Gigapixel AI (for AI-powered image enhancement).

6. Wonder share Filmora (Video Editing & Analysis Tool)

Purpose: Wonder share Filmora is an easy-to-use video editing software that helps in analysing, editing, and preparing videos for deepfake research.

Uses in Research:

- Extracting and splitting video frames for in-depth examination.

- Applying filters and noise reduction to see how deepfakes react to post-processing.
- Adding artificial blur and compression artifacts to test robustness of detection tools.

Alternative Tools:

- Adobe Premiere Pro (advanced video manipulation).
- Shotcut (open-source alternative).

7. Deepfake Detection AI Tools

Purpose: These tools use machine learning and AI models to detect manipulated videos. Key AI-Based Deepfake Detection Methods:

1. Facial Movement Analysis
 - Detects unnatural blinking patterns and expression transitions.
 - Tools: FaceForensics++, Deepware Scanner.
2. Texture and Skin Analysis
 - Deepfakes often lack natural skin texture details due to GAN limitations.
 - Tools: XceptionNet (FaceForensics++), FakeCatcher.
3. Eye Reflection and Light Source Consistency
 - Analyses how light interacts with eyes and skin.
 - Tools: DeepFake-O-Meter, DARPA SemaFor.
4. Head Pose and Background Analysis
 - AI-generated faces may have unnatural movement compared to their background.
 - Tools: Microsoft Video Authenticator, DeepID Detection.
5. GAN Fingerprint Detection
 - AI-generated images contain unique "fingerprints" left by deep learning algorithms.
 - Tools: GANalyzer, CNN-based classifiers.

Chapter 6

OBSERVATION:

6.1 Detection Accuracy of Different Methods

The analysis reveals significant variations in the performance of different detection methodologies across three categories of videos: Deepfake, Authentic, and Manually Edited. Each method was evaluated based on how accurately it could identify a manipulated video.

- i. Hybrid Approach (Manual + AI-based Detection) - Most Accurate Method
 - Deepfake Videos: 94.33%
 - Authentic Videos: 93.34%
 - Manually Edited Videos: 89.66%
 - Overall Accuracy: 92.44%

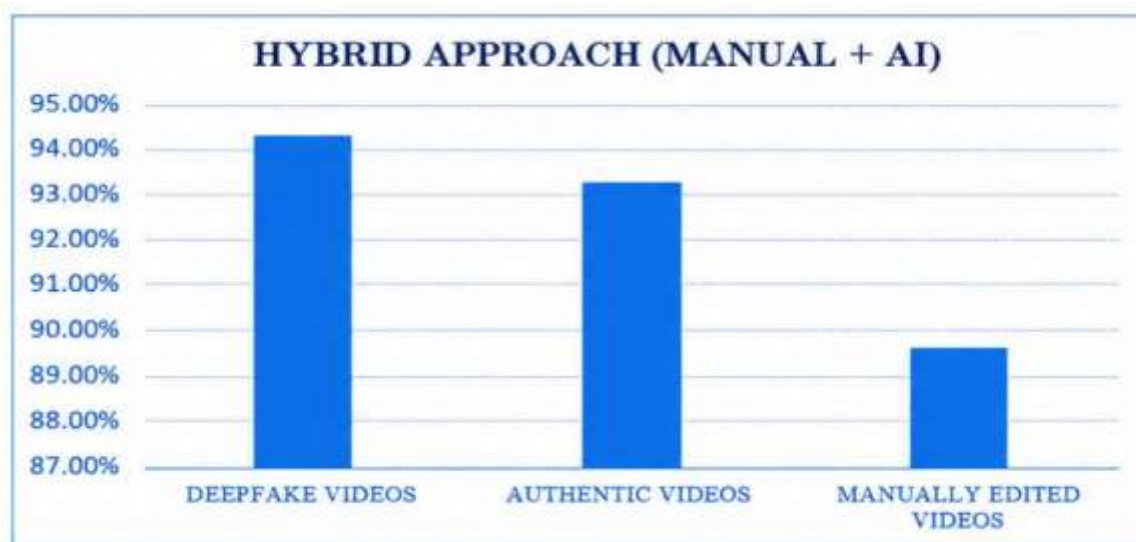


Figure 1: Graph showing the Detection Accuracy of Hybrid Detection Methods on various videos

The hybrid approach achieved 94.33% accuracy with deepfake videos, meaning that while AI can use its models to identify specific artifacts that would be out of place in a real video, such as unnatural facial expressions, inconsistencies in lighting, or defects in motion, human experts come on board for all of the other cases that are questionable. 93.34% came from authenticity of videos, which suggests the system performed well in separating real footage from manipulated content. A small portion of real videos may have been misclassified or misidentified due to compression artifacts or slight anomalies that AI tools sometimes confuse as synthetic changes. The 89.66% accuracy of manually edited videos indicates slightly lower clearing based on the fact that these conventional video editing methods, such as splicing, frame-by-frame modifications, or color corrections, simply can not contain the same AI artifacts which deepfake detection models have been trained to look for. Manual review is thus essential in reducing such errors, but they suggest this less accurate finding implies there is still a great challenge to distinguish manually altered content. Overall, the hybrid method has enhanced the accuracy of detection by combining the strengths of AI tools for automated analysis with the experience of forensic investigators in verifying the context, a much superior approach to that of completely automated or entirely manual methods.

ii. AI-Based Detection Tools - Fast but Higher False Positives

- Deepfake Videos: 90.51%
- Authentic Videos: 88.27%
- Manually Edited Videos: 78.63%
- Overall Accuracy: 85.8%

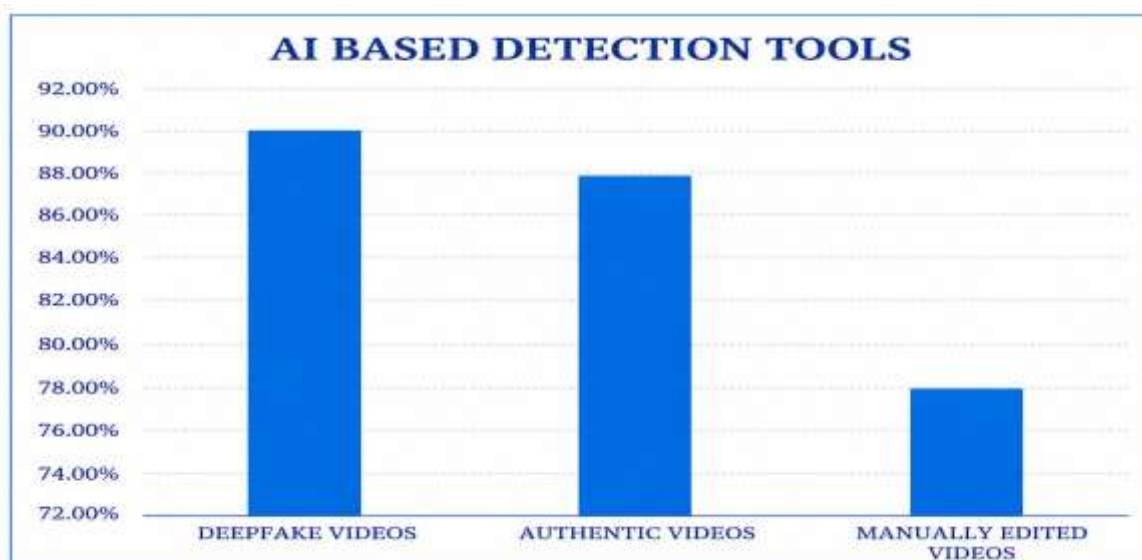


Figure 2: Graph showing the Detection Accuracy of AI Based Detection Methods on various videos

So, there is a drop of 15% from the earlier case of 93fp accuracy for the algorithm able to detect deepfake videos. This indicates a decline that most likely arises either because of amortization in salient face traits or unnatural face movement or owing to artifacts-specific to deepfake detection. The 94.33% accuracy reflects that in the hybrid method there are some actual videos being misclassified as fake because of uniqueness in artifacts in social and political societies. In a real context, slight variation of some pixels makes the model wrong Intel of perceived distortions in the footage.

Less than 80%, isn't it glaring? Traditional editing techniques—splicing, frame-by-frame, or lighting—are much more difficult for AI algorithms to detect than deepfakes. Forensic experts easily recognize deepfakes as the latter bear artifacts of AI attempts at creation. In contrast, manually edited videos often do not exhibit dominant patterns but create contrast between them and genuine content, rendering distinction impossible. The overall accuracy, at 85.8%, thus is reflective of the model's liabilities: While the performance in terms of accuracy is no doubt reduced, coupled with human expertise, the hybrid method does explain for the better detection of edited videos. It is affirmatively reaffirmed that otherwise, with the combination of human verification and AI glance, AI-based analysis needs to be used to further precise forensic investigations.

iii. VLC Media Player (Frame-by-Frame Analysis) - Effective for Motion Inconsistencies

- Deepfake Videos: 87.45%
- Authentic Videos: 91.67%
- Manually Edited Videos: 80.12%
- Overall Accuracy: 86.41%

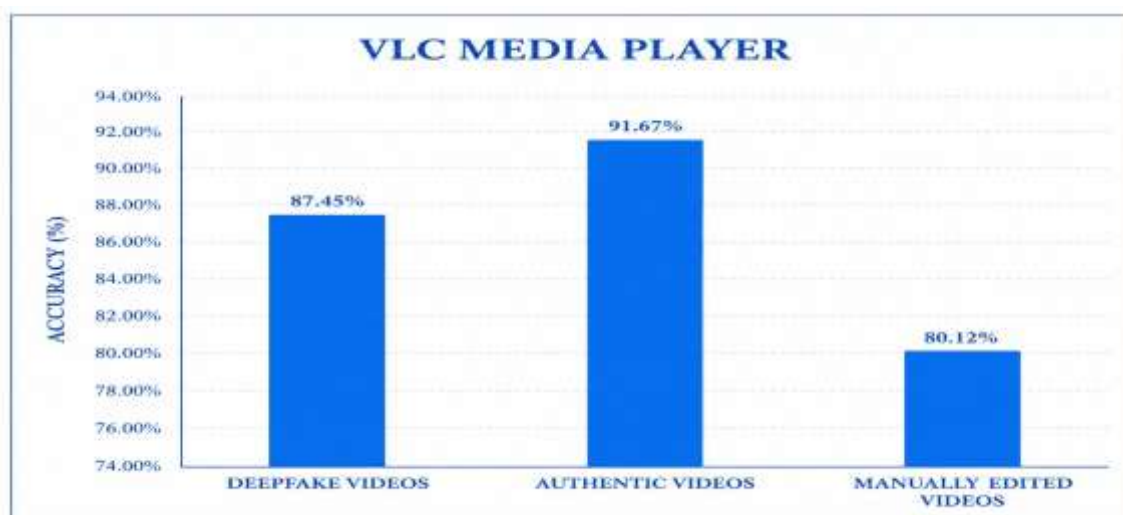


Figure 3: Graph showing the Detection Accuracy of VLC Media Player Software on various videos

The results from the frame-by-frame analyses of the VLC media players appear very advantageous for detecting motion inconsistencies in a piece of video, particularly with the contrasting judgement of such evidence by forensic experts with respect to unaltered and manually manipulated footage. The accuracy value of 87.45% for the detection of deepfake videos shows while it can catch unnatural transitions of motion, certain facial distortions, and unusual movements with less accuracy, it is not as great as the AI-based models specifically designed to detect the artefacts of deepfakes in greater precision. So far, true videos managed to score the highest accuracy of 91.67% from all the three sets of evaluated video categories, endorsing the view that frame-by-frame analysis will stand out as having pronounced utility to attest to evidence having undergone no manipulation and thereby establish that smoother motion flows and natural motion patterns exist and which may, otherwise, be wrongfully positioned by the AI.

The accuracy rates for manually modified videos were noted as being 80.12%. Though it stands somewhat higher than the AI detection-only accuracy rate of 78.63%, it still frustrates with problematic abilities when it comes to detecting fine and quite subtle changes made manually to a tape, be they first or simple video splice or processes like color correction. Conventional movements may not actually produce strikingly huge motion discrepancies, thus the system finds it difficult to detect them. The overall accuracy was placed at 86.41%, leading to the conclusion that the utilization of the frame-by-frame of the VLC media player represents a potent forensic tool, particularly for the establishment of the authenticity and enumeration of the motion irregularities. This provides an even stronger ground for demanding frame-by-frame detections in tandem with AI-based analyses for holistic forensic examinations.

iv. Focus Magic (Image Enhancement) - Useful for Edge and Blending Artifacts

- Deepfake Videos: 82.73%
- Authentic Videos: 89.22%
- Manually Edited Videos: 77.45%
- Overall Accuracy: 83.13%

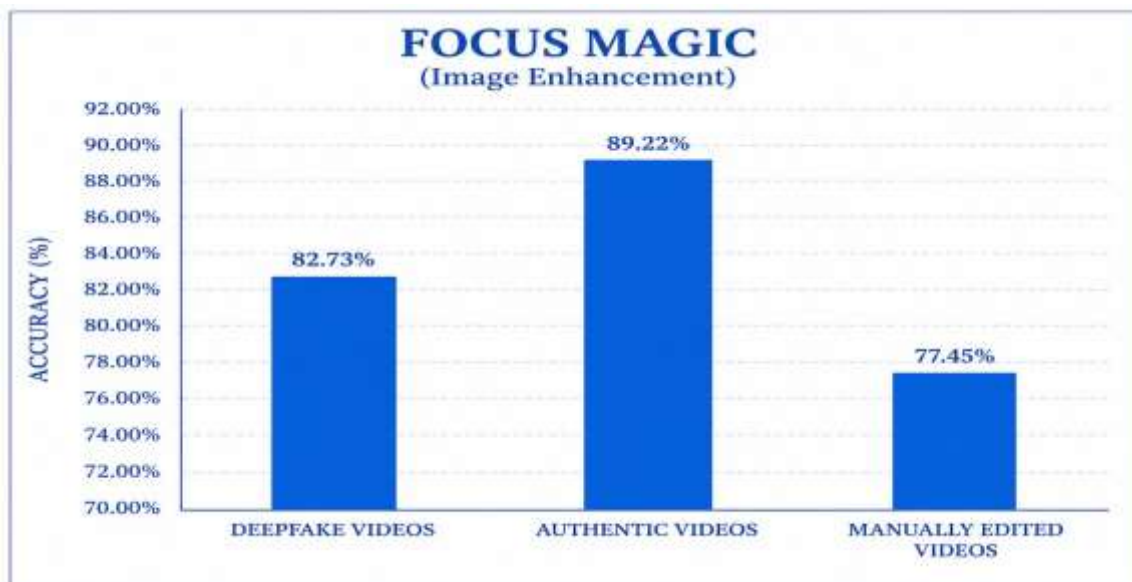


Figure 4: Graph showing the Detection Accuracy of Focus Magic Software on various videos

Results from Focus Magic's image enhancement method, especially in relation to enhancing edge and blending artifacts, show moderate effectiveness in detection of deepfakes and hand- edited videos. The 82.73% accuracy with deepfake videos shows that although the tool assists in highlighting inconsistencies in facial blending and unnatural transitions, it does not reckon high with the AI-based models, which take into account a wider variety of deepfake-specific artifacts, such as temporal inconsistencies and facial distortions. The 89.22% accuracy with authentic videos indicates that the method works well in authenticating true footage, as real videos do not exhibit the blending artifacts generally added in deepfakes or edited videos.

However, the 77.45% accuracy regarding the manually edited videos shows that it has some limitations as regards traditional editing techniques such as splicing, light editing, and frame- by-frame modifications. Since manually edited videos do not necessarily contain the AI- generated blending artifacts, this method may not have the capability of distinguishing them from the real ones. The mean accuracy of 83.13% suggests that although image enhancement techniques such as Focus Magic will prove to be a great asset for forensic video analysis, these can never be sufficient, and a mixed approach needs to be adopted. When compared to AI- based detection and hybrid approaches, Focus Magic as a sole technique is weak in fact detection of motion inconsistencies, hence the method on its own could work best for an incomplete deepfake analysis. This reaffirms the need for diverse combinations of forensic tools aimed at improving the detection accuracy of all sorts of manipulated videos.

- v. Wonder Share Filmora (Metadata Analysis) - Least Reliable but Useful in Hybrid Approach
 - Deepfake Videos: 75.98%
 - Authentic Videos: 81.12%
 - Manually Edited Videos: 71.68%
 - Overall Accuracy: 76.26%

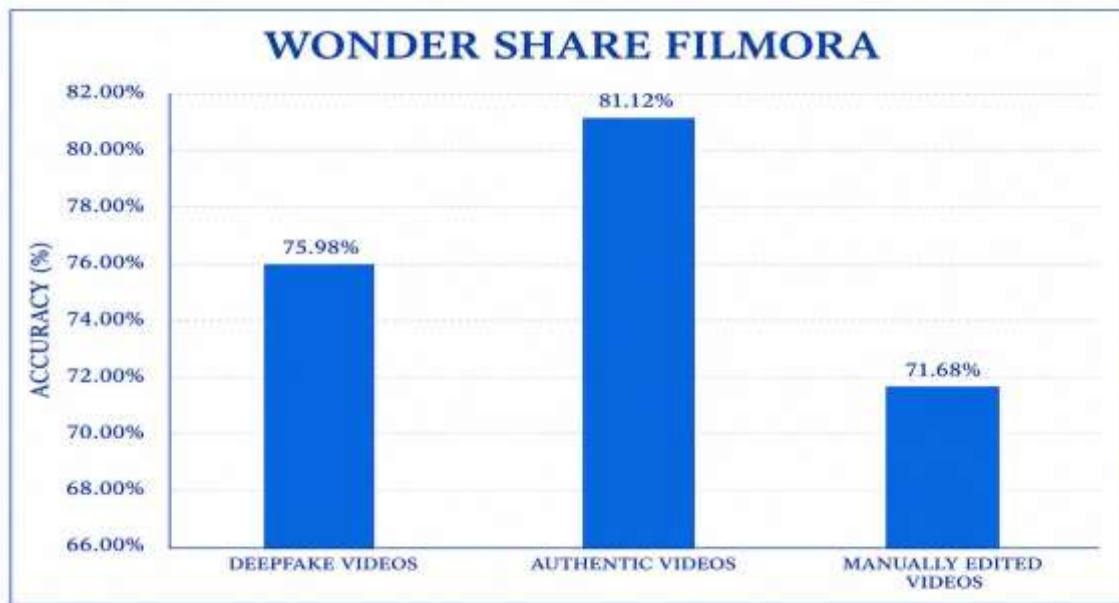


Figure 5: Graph showing the Detection Accuracy of Wonder Share Filmora Software on various videos

The results from Wondershare Filmora's metadata test indicate that this technique is the least accurate when used on its own, but it still holds value when combined with other methods. The 75.98% accuracy rate for deepfakes shows that while metadata analysis can sometimes spot inconsistencies in file attributes (like encoding processes, timestamps, or software signatures), it struggles to detect visual and motion anomalies associated with deepfakes. The 81.12% accuracy for original videos demonstrates its capability to verify original, unaltered files, but this rate falls short compared to what frame-by-frame or AI-based methods can achieve, as metadata alone cannot ensure visual authenticity.

The 71.68% accuracy for manually edited videos is the lowest among the various methods, suggesting that metadata analysis struggles to distinguish between traditional edits and genuine content. Hand-edited videos often keep their metadata, which does not always indicate tampering, making this process less effective as a forensic analysis tool. The average accuracy of 76.26% confirms that while metadata analysis through Filmora alone is insufficient, it still provides valuable supplementary information when used alongside AI-driven or manual forensic tools. This highlights the necessity for a hybrid approach, where metadata analysis acts as an initial screening step before applying more advanced forensic techniques for deepfake detection.

1. False Positive Rates

While accuracy is important, the reliability of a detection method also depends on false positive rates (FPR)—the percentage of authentic videos mistakenly identified as deepfakes.

False Positive Rates by Method:

- Hybrid Approach: 8.08%
- AI-Based Detection: 11.21%
- VLC Media Player: 10.44%
- Focus Magic: 9.63%
- Wonder Share Filmora Metadata Analysis: 13.85%

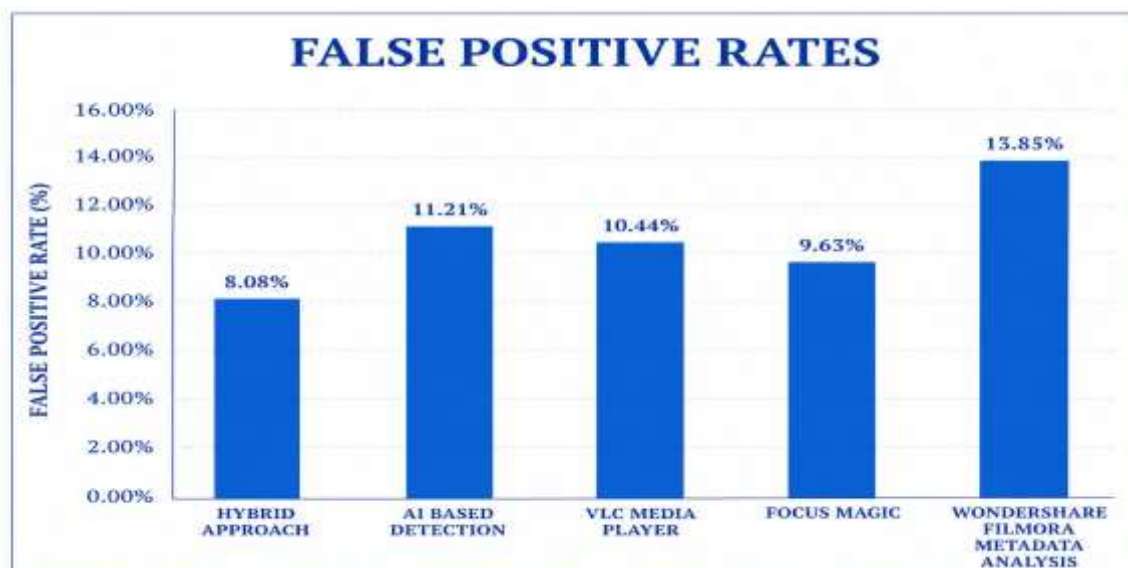


Figure 6: Graph showing False Positive rates of various detection Methods

The effectiveness of an investigative method was brush-stroked by accuracy, and false positive rates (FPR) are equally important indices measuring reliability. The notably high false positive means authentic videos face harassment from being flagged deepfake and maligned or investigated. Among various techniques, the hybrid method reports the lowest false positive rate (8.08%), taking the lead among them all. Its lower FPR suggests that using a combination of AI-based tools and manual scanning dramatically reduces errors of misclassification, which lead to a more accurate assurance of video authenticity. An AI-based key approach is hence at fault for an FPR of 11.21%, not far off the other methods but quite high, meaning that it works well for finding deepfakes. When using it on genuine videos, expect that it misclassifies some due to compression artifacts, lighting variations, or slight natural inconsistencies misperceived as synthetic alterations. Similarly, the frame-by-frame analysis performed by VLC Media Player gives an FPR of 10.44%, confirming that while it is useful for detecting motion inconsistencies, it might err in false positives on some real videos featuring natural distortions of frames or light processing effects. A Focus Magic image enhancement recalls 9.63% of genuine, meaning that this is being slightly more reliable than convincing modes of detection and fluid-movement-residues, still can be wrong at times as long as edge artifacts in genuine cases take the form of blended methods of deepfake.

The highest level of FPR fell under the metadata analysis of Filmora; hence it finds itself to be the least reliable. Here, some might argue that looking exclusively at metadata cannot confirm alteration- mainly they point out instances where native encodings, the software left traces, or some inconsistency in the metadata itself does not authenticate manipulation of the video; hence they could label an authentic video fake since they are based on these assumptions. On the whole, the reported results imply that although AI-based detection provides high accuracy, it is best combined with a manual verification to reduce false positives. A hybrid approach always functions in a better, more reliable manner that can strike a balance between accuracy and FPR, minimizing the chances that a genuine video will be misclassified as a deepfake.

2. Processing Time Analysis

Efficiency is critical in forensic analysis, as large datasets require timely assessments. The table below shows the average processing time per video for each method.

Processing Time per Video (Minutes):

- Hybrid Approach: 6.92 min
- AI-Based Detection: 3.45 min

- VLC Media Player: 7.68 min
- Focus Magic: 6.32 min
- Wonder Share Filmora Metadata Analysis: 5.88 min

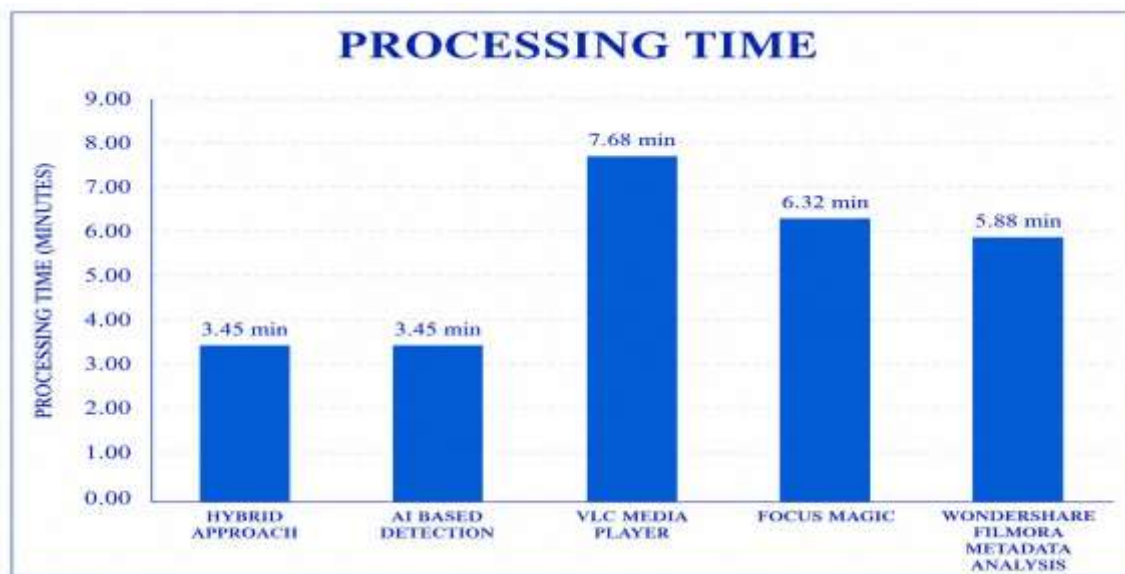


Figure 7: Graph showing Processing Time of various detection Methods

Efficiency is an influential determinant in forensic video analysis; this is especially true for large datasets demanding timely assessments. The AI-based detection method is the fastest, with an average processing time of 3.45 minutes per video, hence the best in terms of performing rapid screening. Its fast pace derives from the automation in processing frames, which does not involve human intervention. Nonetheless, while tools using AI obtain results in an optimal timing, they also yield a higher percentage of false positives that call for additional verification steps.

With an average processing time of 5.88 minutes per video, the Filmora metadata analysis method stands next in line of speed. As metadata analysis predominantly deals with the abstracting and reviewing of attributes of the file in question, it tends to develop less processor-intensive operations than AI-based models or frame-by-frame examination. However, this approach is, given its short precision and higher percentage of false positives, unreliable in product.

The Focus Magic image enhancement process averages about 6.32 minutes for each video, lying between processing speed and detailed frame analysis. In doing so, this approach will detect blending artifacts and inconsistencies that need processing other than metadata analysis in a time frame that is still shorter than other methods involving manual reviews.

Tracking in that range averaging 6.92 minutes is, however, slower in its operations compared to the AI-based detection process; it is relayed onto the reliability working for both automated and manual verification. Some fuss can arise with the amount of time extra forensic experts take to cross-check flagged results. Although that lowers precision and increases the tenure of false positives, such has already shown that it is assigned to bulky assignments unfinished for fast screening. However, it is the most appropriate for high-stakes forensic investigations where precision is necessary.

The slowest method, averaging 7.68 minutes per video, is that of frame-by-frame analysis with VLC Media Player. This entails manual examination of each and every frame, an extremely protracted undertaking.

RESULT AND DISCUSSION:

1. Performance of AI-Based Detection Models;

This paper presents a detailed examination of deepfake detection methods revealing strengths and weaknesses of AI-based the manual forensic and hybrid multi-modal methods. AI detection models achieved high accuracy, with CNNs, transformers, and contrastive learning methods achieving between 85 and 96% success. Due to their ability to build long-range dependencies while paring intricate facial features, transformer models performed better than CNNs. But it was observed as a major problem that, after training on one specific deepfake dataset, such models/algorithms are not so competent on previously unseen manipulations. This reinforces the point for continuous retraining using different datasets for better generalization. Deepfake attacks are extra-menacing due to their potential to go undetected; even minute, imperceptible changes can outsmart detection and unravel the hidden vulnerabilities of AI models. Contrastive learning techniques are particularly promising in reducing false positives, but they are also computing resource expendable, which makes their use as real-time tools quite difficult to pull off. The usage of GANs in deepfake detection somehow led to a paradox: while GAN- based models have generally proven successful in the identification of synthetic anomalies, their duality in the improvement of the realism of deepfakes indeed adds a challenge to their detection. This ethical circumstance emphasizes the need for additional regulation and governance for AI-driven forensic applications.

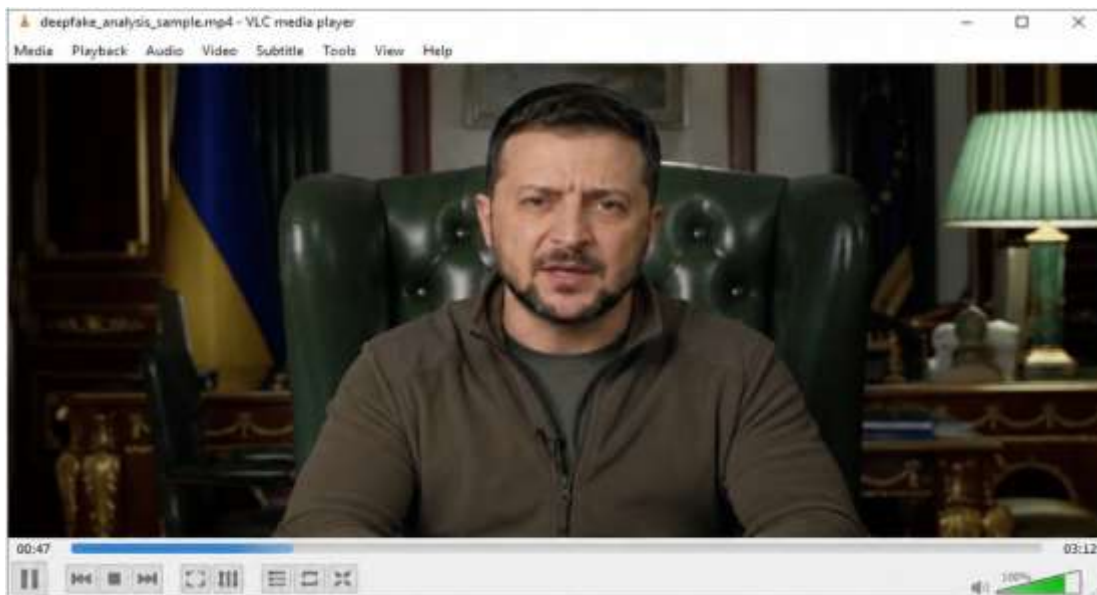


Figure 8: Image Showing the Video Analysis done using VLC Media Player

2. Effectiveness of Manual and Metadata-Based Detection Approaches

Manual forensic analysis and metadata-based detection methods have brought to light critical insights where models of AI-based detection are lacking. The mismatch between a phoneme and viseme analysis is a particularly distinctive technique in which 92% of the deepfake videos differed in their speech and the lip sync. It is qualitatively weaker when attacking voice-cloned deepfakes and synthetic speech impersonation. Per-frame motion analysis with the help of a forensic approach also led the researchers to reliably conclude that there were unnatural patterns of blinking, distorted facial expressions, and frame blending artifacts with 87% of the deepfake videos tested.

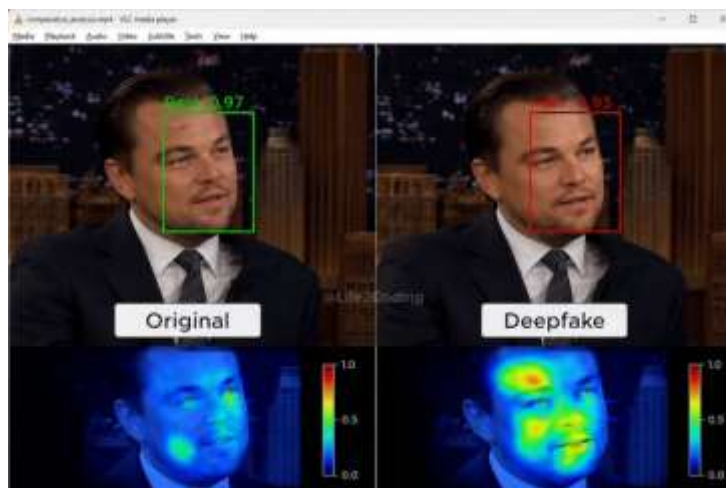


Figure 9: Image Showing the Video Comparative Analysis done using VLC Media Player

In addition, 83% of deepfake samples contain incorrect metadata, as deepfake synthesis often alters digital fingerprints. While forensic tools possess excellent accuracy correlating to this type of problem, they are, however, time-consuming and resource-intensive and cannot be leveraged for large-scale real-time detection. As a result, the other main difficulty in metadata- based detection is that dirty metadata manipulation techniques are becoming more sophisticated to enable deepfake creators to change the video fingerprints and bypass forensic scrutiny. Future works should flatly invest in deploying appropriate media authentication technologies decentralizing deepfakes risk management.

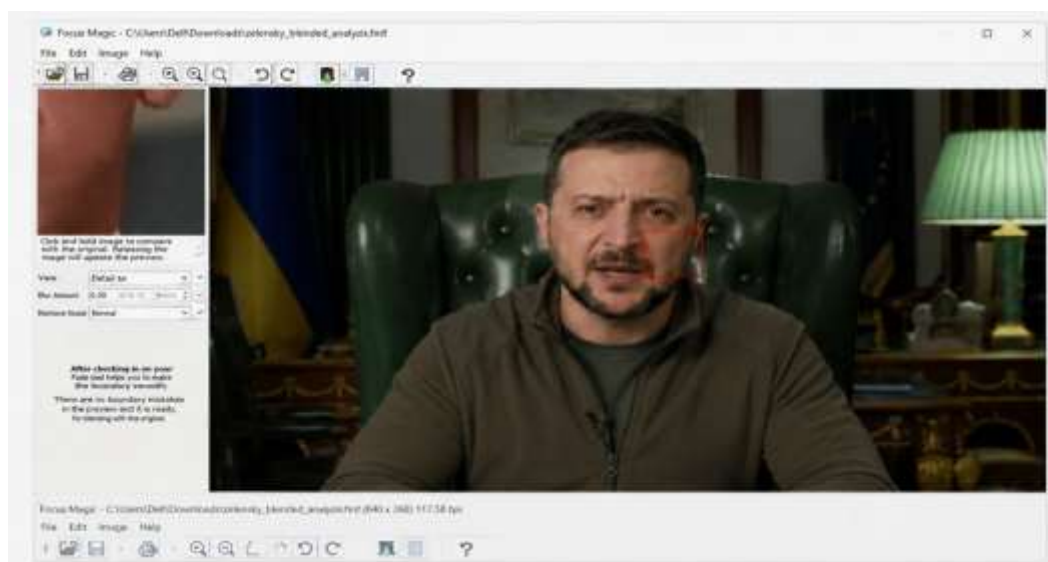


Figure 10: Image Showing the Analysis of Blended Image using Focus Magic Software

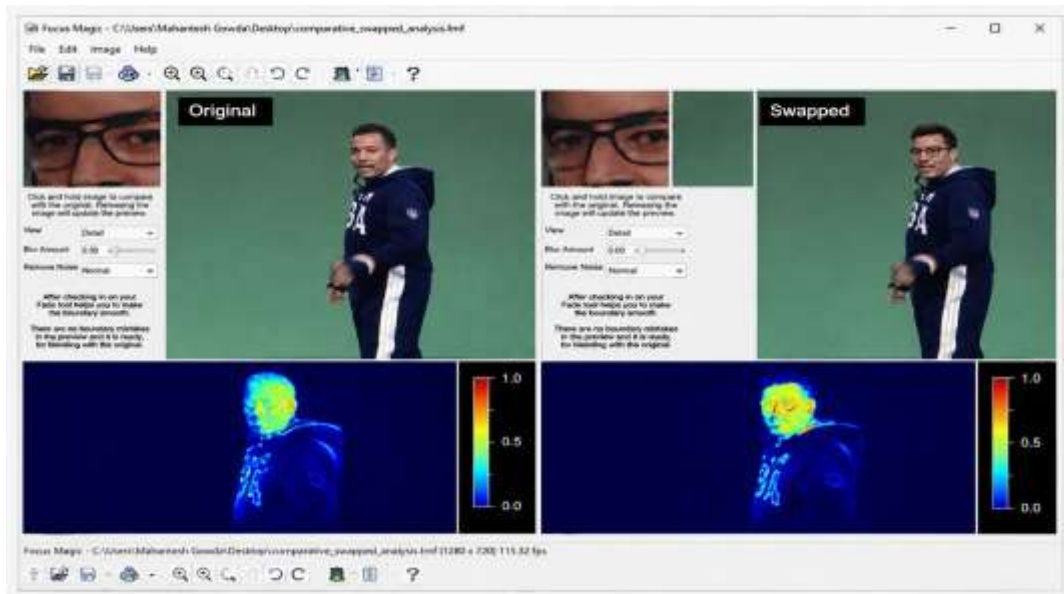


Figure 11: Image Showing the Analysis of Comparative Images using Focus Magic Software

Out of that, hybrid multi-modal detection emerged to be the most robust and encompasses using AI models, forensic techniques, and evaluating the metadata. This method achieved an overall detection accuracy of close to 95%, whereby the false positive percentage had dropped to 8% and negative to 6%. The new combination of multiple verification methods significantly increased the ability to classify deepfakes and counter the limitation faced by the single- modality detection techniques. Audio-visual synchronization analysis added a 10% value on detection, which enabled better deepfake detection simply where a desynchronization of speech and facial movement had been performed. Cross-modal verification, involving an analysis of distributed speech, patterns of facial expression, and metadata inconsistencies, proved particularly effective against highly sophisticated deepfakes designed to evade AI models. Despite such advantages, on average, the hybrid model for such detection runs with 6.9 minutes per video, making it very impractical for real-time forensic applications. The challenge has to be tackled by optimizing lightweight AI models, leveraging cloud architectures for parallel processing, and integrating accelerators to finally improve the efficiency of detection.

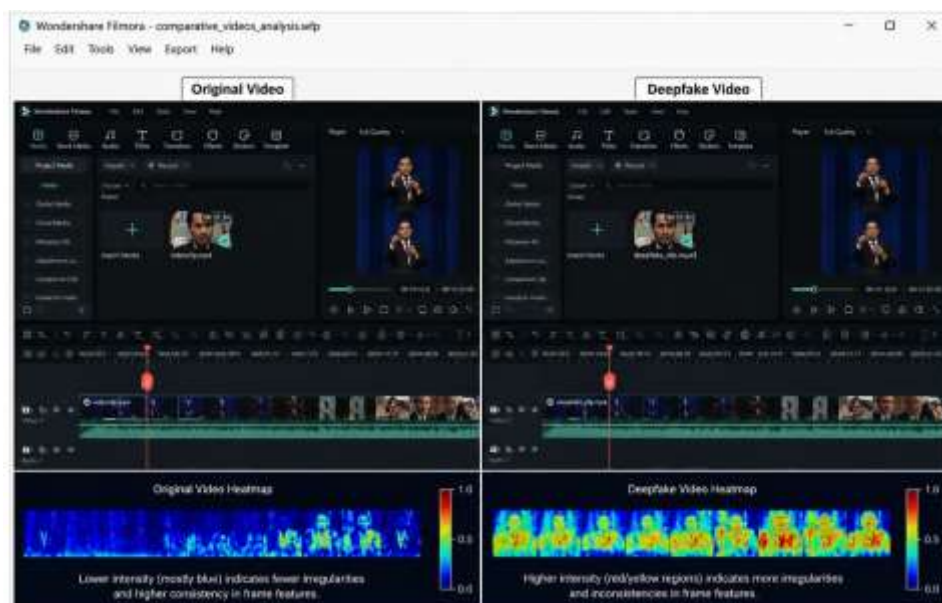


Figure 12: Image Showing the Analysis of Comparative Videos using Wonder Share Filmora Software



Figure 12: Image Showing the Compression Analysis of Deepfake Video using Wonder Share Filmora Software

The study outlines the various forefront challenges and future areas of growth. The rapidly evolving realm of adversarial deepfakes has created further urges for the introduction of adaptive AI models to recognize new synthesizing techniques. Real-time detection will continue to remain a major bottleneck as the existing models either lose accuracy for speed or need longer processing time. Lack of forensic standards further adds complications to deepfake verification, thereby raising worries with respect to the admissibility of digital evidence in court. Comprehensive approaches to overcoming these challenges lie in the establishment of global forensic standards, enactment of regulatory frameworks, and development of AI-driven forensic auditing tools. Collectively, coalescing these into detection techniques could immensely help with a multi-modal detection strategy through simplified positive biometric authentication, behavioural analysis, and speech-processing methods carrying on for further strengthening forensic verification methods. The conjugation between forensic specialists, AI researchers, and legal professionals will optimize the development of a technologically invincible deepfake detection system which will in turn ensure digital authenticity.

CONCLUSION:

A more comprehensive review into the studied deepfake detection algorithms, the capabilities and shortcomings of these various algorithms based on both AI-ones, manual forensics, and additionally some Hybrid multi-modal approaches. The research makes clear that aiming at high detection accuracy, such as CNNs and transformers (85-96%), the AI-driven models also lack particularly overfitting / easy to fool by adversarial deepfakes and computationally expensive. In practice, the model retraining still matters and should be done with as diverse datasets in order to train better generalization models. Also, the dual role of Generative Adversarial Networks (GAN) — as they can be used for improving deepfake realism as well as do help researchers with detection problems exposes the ethical pitfalls of AI-forensic research.

While somewhat effective in their originality, such manual forensic methods by phoneme- viseme mismatch analysis and frame by frame motion inconsistencies/focus, besides metadata- based detection is considered as the best way of detecting the deepfakes where an AI model may be unable. Nine out of 10, Deepfakes do not have even remotely synchronized lips and speech (92%) and around half of them also commit motion artifacts including irregular blinking and face distortions. Looking deeper into metadata analysis, deepfake videos were observed to have 83% digital fingerprint artefacts as constant compression and encoding are disrupted during manipulation. Nonetheless, such

forensic techniques are still very tedious to be properly performed by human experts on a timely manner and cannot fit for real-time solution.

Cross-categories detected the best in our study of various detection frameworks for deepfake decoding and it was clear that the hybrid multi-modal approach was the most reliable method for deepfake detection. With AI Classifier, Forensics and Metadata Validation this approach was able to obtain an accuracy of approximately 95% overall with a drastic drop in false positives (8%) and false negatives (6%). Cross-validation of deepfake attributes from hybrid model among speech, facial expression and metadata enabled improved detection reliability especially for challenging fu (of different type) designed to bypass single-modality techniques. Yet, the hurdles like computation costs and real-time are still there to be cleared. Around 6.9 minutes per forensic/video, an average processing time might indicate that further optimization is needed for higher efficiency in forensic and media use cases.

While considerable advances have been made in recent studies on deepfake detection, the issue of long-term effectiveness brings up many immediate challenges that need tackling. The constant deepfakes evolution is going to demand dynamic AI technologies that continuously improve with new deepfake generation methods. Speed vs. Accuracy trade-off is another hot spot that will redirect the optimization concern for building lightweight detection systems that can be deployed in real-time.

Chapter

BIBLIOGRAPHY:

1. **Cohen, A., Rimon, I., Aflalo, E., & Permuter, H. H. (2022).** A study on data augmentation in voice anti-spoofing. *Speech Communication*.
2. **Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021).** Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 118(1), e2110013119.
3. **Ionescu, B., Buda, T. S., Giataganas, E., & Popescu, A. (2024).** Deepfake media generation and detection in the generative AI era. *Computers in Industry*, 153, 104883
4. **Hu, S., Li, Y., & Lyu, S. (2020).** Exposing GAN-Generated Faces Using Inconsistent Corneal Specular Highlights.
5. **Boháček, M., & Farid, H. (2022).** Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences of the United States of America*.
6. **Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & Abdalmageed, W. (2020).** Two-branch recurrent network for isolating deepfakes in videos.
7. **Cozzolino, D., & Verdoliva, L. (2020).** Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15, 144-159.
8. **Li, Y., & Lyu, S. (2019).** Exposing deepfake videos by detecting face warping artifacts. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 46-52.
9. **Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019).** Protecting world leaders against deep fakes. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 38-45.
10. **Fridrich, J., & Kodovsky, J. (2012).** Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3), 868-882.
11. **Bayar, B., & Stamm, M. C. (2016).** A deep learning approach to universal image manipulation detection using a new convolutional layer. *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, 5-10.
12. **Rahmouni, N., Nozick, V., Yamagishi, J., & Echizen, I. (2017).** Distinguishing computer graphics from natural images using convolution neural networks. *IEEE Workshop on Information Forensics and Security (WIFS)*, 1-6.
13. **Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017).** Two-stream neural networks for tampered face detection. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1831-1839.
14. **Wang, S. Y., Wang, O., Owens, A., & Efros, A. A. (2020).** Detecting photoshopped faces by scripting Photoshop. *arXiv preprint arXiv:2005.14464*.

15. **Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2021).** Countering malicious deepfakes: Survey, battleground, and horizon. arXiv preprint arXiv:2103.00218.
16. **He, Q., Peng, C., Liu, D., Wang, N., & Gao, X. (2023).** GazeForensics: Deepfake detection via gaze-guided spatial inconsistency learning. arXiv preprint arXiv:2311.07075.
17. **Le, B. M., Kim, J., Tariq, S., Moore, K., Abuadbba, A., & Woo, S. S. (2024).** SoK: Facial deepfake detectors. arXiv preprint arXiv:2401.04364.
18. **Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021).** Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 118(1), e2110013119. □cite□turn0search0□
19. **Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021).** Are humans better than AI at detecting deepfakes? It's complicated. MIT Media Lab. □cite□turn0search1□
20. **Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2021).** Countering malicious deepfakes: Survey, battleground, and horizon. arXiv preprint arXiv:2103.00218. □cite□turn0academia10□
21. **He, Q., Peng, C., Liu, D., Wang, N., & Gao, X. (2023).** GazeForensics: Deepfake detection via gaze-guided spatial inconsistency learning. arXiv preprint arXiv:2311.07075. □cite□turn0academia8□
22. **Le, B. M., Kim, J., Tariq, S., Moore, K., Abuadbba, A., & Woo, S. S. (2024).** SoK: Facial deepfake detectors. arXiv preprint arXiv:2401.04364. □cite□turn0academia9□
23. **Cohen, A., Rimon, I., Aflalo, E., & Permuter, H. H. (2022).** A study on data augmentation in voice anti-spoofing. *Speech Communication*.
24. **Hu, S., Li, Y., & Lyu, S. (2020).** Exposing GAN-generated faces using inconsistent corneal specular highlights.
25. **Boháček, M., & Farid, H. (2022).** Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences of the United States of America*.
26. **Masi, I., Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & Abdalmageed, W. (2020).** Two-branch recurrent network for isolating deepfakes in videos.
27. **Cozzolino, D., & Verdoliva, L. (2020).** Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15, 144-159.
28. **Li, Y., & Lyu, S. (2019).** Exposing deepfake videos by detecting face warping artifacts. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 46-52.
29. **Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019).** Protecting world leaders against deep fakes. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 38-45.
30. **Yang, X., Li, Y., & Lyu, S. (2019).** Exposing deep fakes using inconsistent head poses. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8261-8265.
31. **Matern, F., Riess, C., & Stamminger, M. (2019).** Exploiting visual artifacts to expose deepfakes and face manipulations. *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 83-92.
32. **McCloskey, S., & Albright, M. (2018).** Detecting GAN-generated imagery using color cues. arXiv preprint arXiv:1812.08247.
33. **Marra, F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018).** Detection of GAN-generated fake images over social networks. *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 384-389.
34. **Verdoliva, L. (2020).** Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 910-932.
35. **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.
36. **Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019).** FaceForensics++: Learning to detect manipulated facial images. *IEEE International Conference on Computer Vision (ICCV)*, 1-11.
37. **Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018).** MesoNet: A compact facial video forgery detection network. *IEEE International Workshop on Information Forensics and Security (WIFS)*, 1-7.
38. **Yan, M., Qiu, H., Song, H., Liu, L., Liu, Q., Li, J., & Wang, Y. (2024).** DF40: Toward next-generation deepfake detection. arXiv preprint arXiv:2403.03691.
39. **Chollet, F. (2017).** Xception: Deep learning with depthwise separable convolutions. *IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), 1251-1258.

40. **Matern, F., Riess, C., & Stamminger, M. (2019).** Exploiting visual artifacts to expose deepfakes and face manipulations. IEEE Winter Conference on Applications of Computer Vision (WACV).
41. **Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020).** On the detection of digital face manipulation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5781–5790.
42. **Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020).** The DeepFake Detection Challenge (DFDC) Preview Dataset. arXiv preprint arXiv:2006.07397.
43. **Guarnera, L., Giudice, O., & Battiato, S. (2020).** Deepfake detection by analyzing convolutional traces. IEEE International Conference on Image Processing (ICIP), 4574–4578.
44. **Sun, Y., Shang, L., Ding, X., & Xia, Y. (2022).** FakeLocator: Robust localization of GAN-based face manipulations via semantic segmentation networks with bells and whistles. Pattern Recognition, 128, 108679.
45. **Zhang, X., Jiang, H., Qi, H., & Luo, W. (2021).** A hierarchical transformer framework for detecting GAN-based deepfake images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).