FORENSIC SCIENCE JOURNAL SINCE 2002

Forensic Science Journal
2025;24(1):13-24
fsjournal.cpu.edu.tw
DOI:10.6593/FSJ.202512_24(1).0002

The Research on the Trends, Challenges, and Misuse-Prevention of Deepfake Technologies

Wen-Chao Yang, Ph.D.

Department of Forensic Science, Central Police University, Taoyuan City, Taiwan (ROC)

Received: 11 October 2025; Received in revised form 2 November 2025; Accepted: 12 November 2025

Abstract

With the rapid advancement of artificial intelligence (AI), deepfake technologies—driven primarily by deep learning architectures, such as autoencoders, generative adversarial networks (GANs), and diffusion models—have enabled the creation of highly realistic synthetic audio-visual content. While these innovations have legitimate applications in entertainment and education, their misuse poses a serious risk to privacy, democracy, and national security. This study provides a comprehensive examination of the evolving threats, challenges, and preventive strategies associated with deepfake technology. It examines the impacts on media integrity, criminal justice, and public trust, alongside the emergence of legal and regulatory responses in Taiwan, the United States, the European Union, and China. Furthermore, the paper explores state-of-the-art deepfake detection and forensic methodologies, including metadata verification, frequency-domain analysis, noise profiling, and AI-assisted detection models. A standardized forensic workflow is proposed to enhance reproducibility and evidentiary reliability in judicial contexts. Ultimately, the study recommends a multifaceted framework that encompasses legal reform, technological innovation, public education, platform accountability, and international cooperation to mitigate the misuse of deepfake technologies while promoting the responsible development of AI.

Keywords: ***forensic science, deepfake, artificial intelligence (AI), legal regulation, digital evidence, information security***

Introduction

With the rapid advancement of artificial intelligence (AI) technologies, it has become possible to generate highly realistic synthetic audio-visual content using deep learning models. The term “deepfake” is derived from the combination of “deep learning” and “fake,” referring to the use of AI algorithms—particularly generative adversarial networks (GANs)—to synthesize hyper-realistic images or voices. Deepfake technology first emerged around 2017 on online forums, where users employed AI models to blend the faces of celebrities into other videos, creating compelling face-swapped clips. Initially, such technology was often used for entertainment parodies or non-consensual pornographic content. Recently, the barriers to access have been lowered, the quality has improved significantly, and the scope of applications has expanded. Today, deepfake videos can make a target individual appear to be speaking or acting in fabricated contexts, while in reality, the content is entirely synthetic. Since the human eye struggles to distinguish authenticity, this technology has raised widespread concerns [1].

Deepfake technology brings multifaceted threats. On a personal level, it may gravely infringe privacy and reputation rights when misused—for instance, by creating and distributing non-consensual pornographic deepfake videos, which cause both psychological harm

* Corresponding author: Wen-Chao Yang, Ph.D. Department of Forensic Science, Central Police University, No.56, Shuren Rd., Guishan Dist., Taoyuan City 333322, Taiwan (ROC)
Tel: 886-3-3282321 ext 5040
E-mail: una135@mail.cpu.edu.tw

and reputational damage to victims [2]. In the political sphere, deepfakes can be exploited to produce fabricated speeches of political figures or fake news events, thereby misleading voters, manipulating public opinion, and undermining democratic elections. Moreover, criminals can use deepfakes for fraudulent purposes, such as impersonating company executives to instruct unauthorized financial transfers (also known as "CEO fraud") or mimicking the voices of relatives and friends in telephone scams [3].

A further concern is the so-called "liar's dividend" effect [4]: as synthetic media proliferates, the credibility of authentic audio-visual evidence is eroded. Wrongdoers may deny accountability by claiming that genuine video evidence has been fabricated. Thus, the misuse of deepfakes not only generates disinformation but also destabilizes the long-standing belief that "seeing is believing," ultimately undermining media ecosystems and public trust.

To address these challenges, governments, industry, and academia worldwide have recently implemented a range of countermeasures, including legal frameworks, technical detection methods, preventive measures, and civic education. This study aims to consolidate these developments by providing a comprehensive analysis of the threats posed by deepfake technologies and exploring potential preventive strategies. Specifically, this paper discusses legal and policy frameworks, technical detection approaches, the impact on media and public trust, and implications for national security and criminal justice, concluding with integrated recommendations across law, technology, education, social platforms, and international cooperation. Furthermore, the paper explores state-of-the-art deepfake detection and forensic methodologies.

Impacts, Challenges, and Policies

Impacts on Media and Society

Deepfake technology has a profound influence on the media ecosystem and public trust. Its misuse lowers the barrier for producing disinformation while simultaneously causing the public to question authentic information. Misinformation risks include political propaganda, commercial defamation, and extortion. During the 2022 Russia–Ukraine war, a fabricated video of President Zelensky calling for surrender illustrated the role of deepfakes in information warfare. Beyond misinformation, the 'liar's dividend' effect undermines the credibility of genuine evidence, weakening public trust and increasing the costs of verification.

Social media platforms have adopted policies against manipulated media (see Table 1). While enforcement varies, platforms generally prohibit deceptive content, require labeling of AI-generated media, and apply penalties. However, challenges remain in distinguishing satire from harm and striking a balance between moderation and free speech.

Table 1. Social media platforms' policies against manipulated media

Platform	Policies Abstract
Facebook/ Meta	Deepfake videos created using AI or similar technologies that are indistinguishable from authentic content by the average person are prohibited. Content that has been artificially edited and could potentially mislead viewers will be removed, with the exception of obvious parodies or art forms. Furthermore, Facebook will downgrade the recommendation weight of videos identified as fake by third-party verification and place a warning label. On the eve of the 2020 US election, Meta announced a complete ban on AI-generated political fakes to prevent election interference.
Twitter (X)	Users are prohibited from sharing synthetic or manipulated media in a deceptive manner that could lead to actual harm, and violators will face content removal. For synthetic media that may mislead but does not pose a serious risk of harm, the platform will label it as "manipulated media" to alert viewers [5]. This policy took effect in February 2020. During the election, Twitter also labeled suspected deepfakes (such as edited and manipulated clips of politicians' speeches) "Edited Video" and provided a link to a fact-checker.

Platform	Policies Abstract
YouTube	YouTube's policy on deepfakes focuses on elections and seriously misleading topics. YouTube explicitly prohibits the uploading of fake content that "could seriously mislead users and cause significant harm," including AI-generated videos of politicians speaking. Specifically regarding elections, the platform has removed videos since 2020 that clearly fabricate candidate statements or actions and could potentially influence election results. For deepfakes not generally related to elections, YouTube primarily addresses the fraud and hate content standards in its Community Guidelines, and may also remove content that constitutes harassment or defamation.
TikTok	The use of deepfake technology to alter images of private individuals or to create composite content containing the likeness of public figures that implies commercial endorsements or violates platform regulations is prohibited. In 2023, TikTok updated its community guidelines, requiring all realistic AI-generated content to be clearly marked as such, typically by including a label in the video captions or within the video itself [6]. For deepfake videos of public figures, such as politicians and government officials, TikTok allows satire or artistic representation of the situation, but it cannot be used to impersonate them, promote products, or spread hate speech. The platform will remove any unlabeled, misleading deepfake content found.
Others	Platforms like Instagram and Reddit generally adhere to similar principles: limiting the spread of misleading deepfake content. LinkedIn explicitly prohibits scams using AI-generated fake profile pictures. Chat platforms like Discord include reminders in their community guidelines against using fabricated media to defraud others. Major platforms are aware of the dangers of deepfakes and have implemented measures such as labeling, traffic reduction, and removal, but enforcement and detection capabilities vary.

Media literacy has emerged as a key defense against misinformation. Educational programs worldwide now include deepfake awareness, and fact-checking organizations are essential in rapidly debunking viral fakes. Still, literacy alone cannot solve the problem and must be complemented by technological detection and platform responsibility.

Challenges for National Security and Criminal Justice

Deepfakes pose risks to national security and criminal justice systems. They can be exploited for propaganda, terrorism, or election interference, while courts struggle to verify digital evidence. In 2022, a deepfake of President Zelensky urging surrender spread online, highlighting the potential of deepfakes in information warfare.

Economic crimes are also affected: AI voice scams have tricked victims into ransom payments, and CEO fraud has cost companies hundreds of thousands of euros. Synthetic identity fraud, combining AI-generated faces with stolen personal data, has surged globally.

In judicial contexts, fabricated evidence undermines the credibility of video/audio records. Defendants may claim authentic evidence is fake, complicating trials. Law enforcement agencies are responding by adopting forensic tools, training officers, and promoting cross-border cooperation. Organizations such as Europol and Interpol have prioritized deepfake threats, while Taiwan and other jurisdictions have incorporated deepfake detection into digital forensics.

Ultimately, combating deepfakes in security and justice requires continuous adaptation, upgraded evidentiary standards, and international collaboration.

Legal and Policies

The improper use of deepfake technology has raised pressing concerns about privacy violations and information manipulation, prompting governments worldwide to re-examine existing legal frameworks and develop new legislation.

In Taiwan, several high-profile deepfake-related crimes—such as the 2021 case where an influencer, "Xiao Yu," used deepfake technology to create and

sell large volumes of non-consensual pornographic videos—accelerated legal reforms. In February 2023, Taiwan amended its Criminal Code, adding Chapter 28-1: Offenses Against Sexual Privacy and Non-Consensual Sexual Images (Articles 319-1 to 319-6). Among them, Article 319-4, known as the "Xiao Yu Clause," criminalizes the creation or distribution of non-consensual deepfake pornography. Offenders face up to five years of imprisonment; if committed for profit, the maximum penalty increases to seven years. This provision closed a legal loophole that previously allowed such cases to be prosecuted only under indirect offenses, such as obscenity or defamation. In May 2023, Taiwan also amended Article 339-4 of the Criminal Code (Fraud) to impose stricter penalties for fraud committed using AI-generated content (images, audio, or video). Offenders may face up to seven years of imprisonment and fines of up to NT$1 million [7]. The rationale is that AI-generated media spreads rapidly and is difficult to detect, posing greater risks to society. Beyond criminal law, the use of deepfakes in elections has also attracted attention. While Taiwan's Election and Recall Act and related laws currently lack specific provisions on deepfakes, the dissemination of disinformation to disrupt elections may still be prosecuted under existing laws. Overall, Taiwan's legal amendments reflect a firm stance against deepfake-enabled crimes in both sexual exploitation and fraud.

As of 2025, the United States has no federal law specifically addressing deepfakes; however, many states have enacted relevant legislation since 2019. State laws mainly target two areas: (1) non-consensual deepfake pornography, often extending existing revenge-porn statutes, and (2) deepfake disinformation in elections, particularly within 60 days of voting. By 2024, approximately 30 states had passed laws covering deepfake-related offenses [8]. For example, California banned non-consensual deepfake pornography and prohibited publishing fabricated candidate videos within 60 days of an election. Texas similarly criminalized the distribution of election-related deepfakes. Where specific laws are lacking, prosecutors often rely on statutes related to identity theft, cyber harassment, or defamation. Federally, bills such as the Deepfakes Accountability Act have been introduced to mandate the labeling of deepfake content, but progress has been limited due to concerns about the First Amendment. Meanwhile, the FBI has repeatedly warned of the risks associated with fraud in AI-generated media. In 2023, the Biden administration issued an executive order on AI governance, which requires the implementation of content authentication mechanisms.

The European Union has moved toward a unified regulatory framework. The EU Artificial Intelligence Act (AI Act) [9], finalized in 2023, requires transparency for AI-generated or manipulated media (Article 50). Any deepfake content provided or used within the EU must be clearly labeled or watermarked as artificially generated, with exceptions for law enforcement, public interest, or artistic expression [10]. Violations may result in significant fines when the Act takes full effect in 2026. Additionally, the Digital Services Act (DSA) requires large platforms to mitigate risks of disinformation, including deepfakes. Platforms signing the updated Code of Practice on Disinformation (2023) committed to developing detection tools and applying warning labels. Beyond the EU, the UK's Online Safety Act (2023) explicitly criminalized the non-consensual sharing of sexual deepfakes. In mainland China, the Deep Synthesis Internet Information Service Regulation (2022, effective 2023) requires that all AI-generated media be prominently labeled as synthetic and prohibits its use in extortion, rumor spreading, or illegal activities [11].

Globally, a consensus is forming that legal frameworks must require transparency and accountability in AI-generated content. By criminalizing malicious uses (e.g., fraud, election interference, sexual exploitation) and mandating content labeling, governments aim to mitigate the growing threats posed by deepfake technologies.

Detection and Forensics

Technical Detection

To mitigate the threats posed by deepfakes, it is crucial to understand both how they are generated and how they can be detected. This subsection introduces three mainstream deepfake generation techniques: Autoencoders, Generative Adversarial Networks (GANs), and Diffusion Models, before outlining current detection methods.

Autoencoders

An autoencoder is an unsupervised neural network architecture designed to learn low-dimensional representations of data and reconstruct them [12]. A typical autoencoder consists of two parts: an **encoder** and a **decoder**. The encoder compresses the input data into a lower-dimensional latent vector, while the decoder attempts to reconstruct the original data from this vector. Since the network is trained to produce an output as close as possible to the input, the autoencoder is compelled to learn how to extract the most salient features of the data, ensuring that the compressed representation can still approximate the original input upon reconstruction. When the latent space has significantly lower dimensionality than the input, the autoencoder effectively performs **dimensionality reduction** and **feature extraction**.

Mathematically, an autoencoder minimizes the difference between the input and the reconstructed output (e.g., mean squared error). Through gradient descent, the encoder and decoder adjust their parameters jointly. When properly trained, the encoder produces an efficient representation of the original data, while the decoder learns to reconstruct the data from that representation.

Autoencoders play a crucial role in deepfake technology, particularly in early **face-swapping** techniques. Some extensions are introduced as follows:

Face Replacement: [13]

Many of the earliest viral deepfake videos (e.g., swapping a celebrity's face onto another actor's body) used a specialized autoencoder architecture. This technique involved two decoders sharing the same encoder (see Figure 1): one decoder was trained to reconstruct person A's face, while the other was trained to reconstruct person B's face. The shared encoder learns to extract general facial features, encoding expressions and pose into a latent vector. At inference time, an image of person A is passed through the encoder, and decoder B reconstructs a version of person B's face that preserves A's expression and pose. By processing every frame of a video sequence, a realistic face-swapped video can be generated.

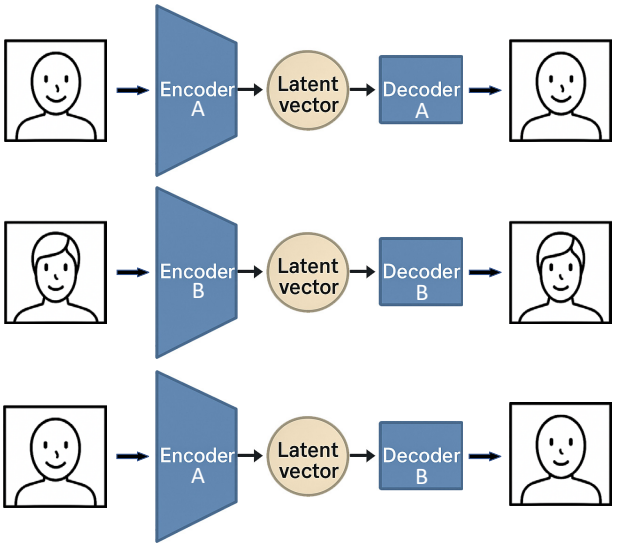

Fig. 1. The illustration of Autoencoder

Variational Autoencoders (VAE):

A notable variant is the **Variational Autoencoder (VAE)** [14], which introduces a probabilistic interpretation to the latent space by assuming a prior distribution. The encoder outputs parameters of this distribution, from which latent vectors are sampled. The decoder reconstructs samples, enabling both the generation of new data and the interpretation of existing data. While VAE-generated images may appear blurrier than those from GANs, VAEs offer stable training and better control over the latent space. Hybrid approaches such as **VAE-GAN** [15] combine the strengths of both methods, benefiting from VAE's stability and GAN's image sharpness.

Voice Cloning and Other Modalities:

The autoencoder concept generalizes beyond images. For example, in **voice conversion**, an autoencoder can map a source speaker's voice to that of a target speaker. The encoder extracts linguistic and prosodic features, while the decoder learns the timbre and identity of the target speaker, enabling the generation of speech in the target's voice. This technology has been used in v**oice deepfakes**, including the impersonation of public figures for fraudulent calls.

Overall, autoencoders and their variants provide efficient representation learning for deepfake tasks. Compared to GANs and diffusion models, which focus

on directly generating realistic data, autoencoders emphasize the reconstruction and transformation of existing samples, making them well-suited for applications such as face swapping and voice mimicry. However, the realism of autoencoder outputs typically falls short of GAN-generated results, often requiring additional techniques for quality enhancement.

Generative Adversarial Networks (GANs)

Ian Goodfellow *et al.* proposed GANs in 2014 [16], a deep learning architecture consisting of two competing neural networks: a **generator**, which produces synthetic data, and a **discriminator**, which distinguishes between real and generated data. The generator aims to produce outputs that can deceive the discriminator, while the discriminator strives to enhance its ability to differentiate between real and fake. This adversarial training continues until a Nash equilibrium is reached, ideally resulting in generator outputs that are indistinguishable from real data.

Mathematically, the training process is expressed as a minimax game: the discriminator maximizes classification accuracy (real vs. fake), while the generator minimizes the probability that its outputs are classified as fake. This process is illustrated in Figure 2.

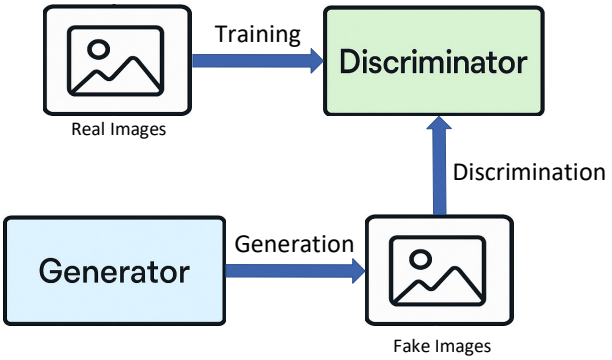

Fig. 2. The illustration of Generative Adversarial Networks

Some deepfake applications and extensions are introduced as follows:

Face Generation:

GANs can generate highly realistic faces from random latent vectors. A prime example is NVIDIA's **StyleGAN** [17] series, capable of synthesizing photorealistic, non-existent human faces. Such techniques are used in deepfake apps to create synthetic portraits or replace faces in videos.

Video Forgery:

Extending GANs to video data enables frame-by-frame generation of facial movements, creating fake talking-head videos. Early deepfake implementations on internet forums employed GAN-based approaches (e.g., Face2Face [18]), mapping source video facial expressions to target faces with remarkable realism.

Image-to-Image Translation:

GANs are widely applied to **style transfer** and **image translation** tasks, such as turning daytime images into nighttime scenes or converting sketches into color images. These techniques can also be misused to fabricate events or alter visual evidence.

While GAN-generated images are highly realistic, they may still exhibit subtle artifacts, such as unnatural textures. GAN training is notoriously challenging, with issues such as mode collapse (a lack of output diversity). These challenges have motivated researchers to explore alternative generative models, resulting in the emergence of diffusion-based approaches.

Diffusion Models

Diffusion models [19]-[20] are a class of generative models that have rapidly gained popularity in recent years. They are inspired by the physical process of diffusion and operate in two stages: a **forward process** that gradually corrupts data with noise until it resembles pure noise, and a **reverse process** that learns to denoise step by step, reconstructing the original data. Conceptually, this is akin to "sculpting an image from noise."

During training, a real data sample (e.g., a clear image) is progressively corrupted with noise over multiple steps until it becomes indistinguishable from random noise. The model is trained to reverse this process, predicting and removing noise at each step, thereby learning the data distribution. Figure 3 illustrates this process.

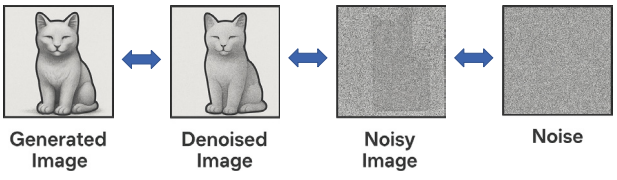

Fig. 3. The illustration of the Diffusion Model

During inference, generation begins with random noise, and the model iteratively applies its learned denoising steps to produce a coherent, high-quality output. When conditioned on text prompts—known as **conditional diffusion models** [21]—the process is guided toward generating images that match the textual description. Popular text-to-image systems such as Stable **Diffusion** [22] and **Midjourney** [23] are based on this principle.

Diffusion models are praised for their training stability and ability to produce fine-grained detail. Because each step introduces only minor refinements, the model can carefully control image features. Compared to GANs, which generate a complete image in a single pass, diffusion models achieve higher quality and diversity but require more computation during inference.

Some deepfake applications and extensions are introduced as follows:

High-Fidelity Portrait and Scene Generation:

Diffusion models can generate photorealistic images, specify styles, and produce high-resolution portraits. This makes them powerful tools for fabricating convincing fake photos and scenes.

Deepfake Video Generation:

Although initially focused on still images, research is extending diffusion models to video, using frame-by-frame or spatiotemporal diffusion. Future deepfake videos may utilize diffusion models to achieve improved temporal coherence and realism, surpassing the capabilities of GAN-based methods.

Image Editing and Restoration:

Diffusion models can perform inpainting (filling in missing regions) and super-resolution tasks. For instance, they can modify facial expressions or reconstruct damaged image areas. Such techniques could be misused for subtle tampering, making detection harder.

Diffusion models provide a powerful and flexible framework for generating deepfakes, enabling the creation of highly realistic and complex content. However, their reliance on stochastic processes introduces variability in outputs, posing new challenges for detection, as traditional deepfake detectors may struggle to identify their characteristic patterns.

Deepfake generation has evolved from autoencoders to GANs and diffusion models, producing increasingly realistic content. Some potential detection strategies include:

- Photo-Response Non-Uniformity (PRNU) analysis [24], which examines sensor fingerprints to identify inconsistencies.
- Deep learning classifiers [25], which utilize CNNs trained on both real and fake datasets, are used to detect subtle artifacts.
- Explainable AI (XAI) methods [26], which employ heatmaps or network dissection to highlight decision-making regions.
- Physical and physiological consistency checks, such as blink frequency [27], rPPG heartbeat signals [28], lighting and shadow coherence, and lip-sync accuracy [29].

Deepfake video forensics

Recently, the Colombian Constitutional Court's ruling T-323/24 established strict requirements for transparency and explainability in AI models used within judicial decision-making [30]. These measures are designed to safeguard impartiality, fairness, and the protection of fundamental rights. Therefore, pending the completion of reliable AI-driven detection and forensic tools, establishing transparent and explainable deepfake forensic procedures remains an urgent necessity. Based on the preceding analysis and discussion, we propose the following recommendations for a deepfake forensic process to ensure that the examination results are reliable and reproducible:

1. ***Evidence Reception and Preservation:***

   The foundational step of forensic examination is to establish a transparent *chain of custody* by extracting the digital video or image evidence from its source medium. Compute a digital hash value (e.g., SHA-1, SHA-256, etc.) to confirm that the file remains unaltered throughout subsequent analyses. Following standard digital forensic practices, the original file must be stored as *read-only*, and all operations should be performed on a verified duplicate to maintain evidentiary integrity.

2. ***Preliminary Assessment:***

   Conduct an initial viewing or listening of the duplicate file in light of the case background to identify the focal points requiring authentication (e.g., whether a particular video segment or voice recording is fabricated). At this stage, investigators should also inquire about the provenance and transfer history of the evidence and record any factors that might influence the interpretation.

3. ***Metadata and Device Verification:***

   Extract and review EXIF and encoding information from the media to assess their plausibility. When possible, obtain a sample of the alleged recording device (e.g., if the evidence is said to come from a suspect's smartphone, acquire other authentic videos recorded on the same device). Comparative analysis of file structures, visual patterns, or even *camera fingerprinting* can substantially strengthen conclusions. In the absence of the physical device, cross-reference the media's attributes with large-scale public databases of camera or smartphone samples to determine whether its properties deviate from expected norms.

4. ***Content Analysis:***

   Combine manual inspection with software-assisted tools to detect visual or auditory anomalies. Frame-by-frame examination should be conducted using video playback software, marking any suspicious details—such as irregular facial edges, shadows, reflections, or mismatched audio-visual synchronization. In accordance with image-forensic standards, the examiner may adjust brightness, contrast, or magnification to reveal distinctive details or artifacts better.

5. ***Technical Analysis:***

   Apply frequency-domain transformations and noise analysis to detect unnatural patterns in the visual content. For audio, analyze waveforms and spectrograms to identify inconsistencies. Compute statistical features such as luminance histograms or inter-frame differences, comparing them with those from authentic videos. Additionally, employ AI-based deepfake detection tools to obtain a *probability score* (commonly from 0 to 1) indicating the likelihood of manipulation, along with identified artifacts of forgery. All analytical findings should be consolidated to determine whether there is consistent evidence pointing toward fabrication.

6. ***Conclusion and Reporting:***

   Integrate findings from all stages to reach a forensic conclusion. For example: *"According to the analysis, the examined video is classified by the XAI model as a deepfake with a probability exceeding 99%. At timestamp XX seconds, a segment exhibits a high likelihood of being a deepfake, and frame YY reveals significant irregularities in both visual content and digital characteristics, inconsistent with the statistical properties of authentic recordings. Collectively, the evidence indicates that this video is highly likely to have been synthesized using deepfake technology."* The forensic report should clearly document the employed methods (including software and hardware tools, versions, and parameter configurations), the specific findings (supplemented with annotated key frames or charts), and the expert's professional opinion. The report must be systematic, objective, complete, and reproducible—allowing any expert of comparable competence to replicate the procedures and obtain consistent results, a crucial factor for admissibility in court.

7. ***Cross-Verification:***

   To further enhance reliability, incorporate verification through alternative methods—such as re-analyzing the evidence using a different tool or an independent analytical principle. Many international forensic laboratories maintain ISO/IEC 17025 *Accreditation systems* to ensure consistent quality and have formed expert communities to share emerging deepfake cases and detection techniques, keeping abreast of evolving threats and countermeasures.

An example of Deepfake video detection and forensics

An example video was downloaded from TikTok (https://www.tiktok.com/@parishiltonsays/video/7181871717126606122). Before we illustrate the proposed forensic process, we utilize the renowned deepfake video detection tool, DeepWare [31], to detect the video. The detection result is shown in Figure 4.

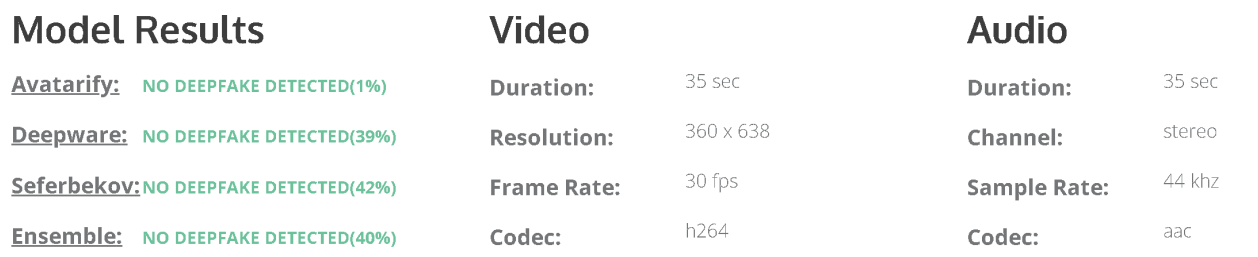

Fig. 4. The detection result of the example video in DeepWare

In our forensic processes, we first calculate the SHA-1 hash value of the downloaded video as "c4cdb a8929a80614b29ab58f9e74021d26a5ab1b." Then, this examination requires determining whether the video is a deepfake. Third, the FFmpeg tools are used to extract the video's metadata. The resolution and frame rate of the video are 360 x 638 and 30 FPS, respectively. After we split the frames and audio, we detect visual or auditory anomalies within them in the fourth process. The right facial image in the 19th frame shows the flaw of the face swap (see Figure 5). Then, we follow the fifth to the seventh processes to complete the examination report.

Fig. 5. The 19th frame of the question video

These recommendations provide a **comprehensive, multi-layered approach**—from data format verification to content analysis, integrating human expertise with machine intelligence. Given that no two forensic cases are identical, practitioners must apply professional judgment and adapt the workflow according to the specific context of each deepfake investigation.

In practice, hybrid approaches that combine rapid screening with forensic validation are preferred. Nevertheless, researchers stress that deepfake detection is a 'cat-and-mouse game,' requiring continuous innovation, hybrid strategies, and possibly watermarking and provenance certification systems.

Conclusion and Recommendations

Deepfake technology, as a double-edged sword of AI, offers innovative applications but also creates new risks. This study proposes five key recommendations:

1. Strengthen legal frameworks and law enforcement, including criminalizing non-consensual deepfake pornography, election manipulation, and AI-enabled fraud; coordinate international standards for content labeling; and create victim-centered remedies.
2. Develop XAI detection tools and anti-forgery mechanisms, including advanced multi-modal detection, anomaly detection, and provenance authentication systems, such as cryptographic signatures.
3. Enhance public media literacy by incorporating deepfake recognition into curricula, launching awareness campaigns, and strengthening fact-checking institutions.
4. Strengthen platform accountability: Require platforms to deploy detection systems, enforce labeling and takedowns, and maintain transparency.
5. Promote international cooperation by establishing new global media standards, sharing intelligence, and

forming multinational task forces for high-impact cases.

While threats are fundamental, integrated approaches across law, technology, education, governance, and cooperation can mitigate risks. At the same time, society must explore positive applications of deepfake technologies in film, medicine, and accessibility.

Acknowledgments

This work on this paper was supported by the Ministry of the Interior, Republic of China (Taiwan). (Project No. 115-0805-02-28-01).

References

1. Geng Y. Comparing deepfake regulatory regimes in the United States, the European Union, and China. Georgetown Law Technol Rev 2023; 7: 157-78.
2. Women's Rescue Foundation. Deepfake pornography severely harms victims: Call for legislation against non-consensual sexual image crimes. Women's Viewpoint Articles. Apr. 2021.
3. Europol Innovation Lab. Facing reality? Law enforcement and the challenge of deepfakes. Publications Office of the European Union, 2022. ISBN 978-92-95236-23-3, DOI: 10.2813/158794.
4. Schiff KJ, Schiff DS, Bueno NS. The liar's dividend: can politicians claim misinformation to evade accountability? American Political Science Review. Published online 2024:1-20. DOI:10.1017/S0003055423001454.
5. Roth Y, Achuthan A. Building rules in public: Our approach to synthetic & manipulated media. Twitter Company Blog, Feb. 4, 2020. [Full text freely available at: https://blog.x.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media].
6. Vincent J. TikTok bans deepfakes of nonpublic figures and fake endorsements in rule refresh. The Verge. Mar. 21, 2023. [Full text freely available at: https://www.theverge.com/2023/3/21/23648099/tiktok-content-moderation-rules-deepfakes-ai].
7. Shan S. Legislature passes stiffer jail, fine for deepfake fraud. Taipei Times. May 17, 2023. [Full text freely available at: https://www.taipeitimes.com/News/front/archives/2023/05/17/2003799936].
8. National Conference of State Legislatures (NCSL). Deceptive audio or visual media ('deepfakes') 2024 legislation. 2024. [Full text freely available at: https://www.ncsl.org/technology-and-communication/deceptive-audio-or-visual-media-deepfakes-2024-legislation].
9. The European AI Office. The EU Artificial Intelligence Act. [Full text freely available at: https://artificialintelligenceact.eu/].
10. Hickman T, Lorenz S, Teetzmann C, Jha A. Long awaited EU AI Act becomes law after publication in the EU's Official Journal. White & Case Insights. Jul. 16, 2024. [Full text freely available at: https://www.whitecase.com/insight-alert/long-awaited-eu-ai-act-becomes-law-after-publication-eus-official-journal].
11. Sumsub. Global deepfake incidents surge tenfold from 2022 to 2023. Identity Fraud Report. Nov. 2023.
12. Vincent P, Larochelle H, Lajoie I, Bengio Y, Manzagol PA. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J Mach Learn Res 2010; 11: 3371-408.
13. Github. Faceswap: Deepfakes software for all, 2018. [Full text freely available at: https://github.com/deepfakes/faceswap]
14. Kingma DP, Welling M. An introduction to variational autoencoders. Now Foundations and Trends, 2019. ISBN:9781680836226.
15. Razghandi M, Zhou H, Erol-Kantarci M, Turgut D. Variational autoencoder generative adversarial network for synthetic data generation in smart home. Proc. ICC 2022 - IEEE International Conference on Communications, Seoul, Republic of Korea, 2022; 4781-786, DOI: 10.1109/ICC45855.2022.9839249.
16. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS), 2014; 2672-80.
17. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Proc. IEEE/CVF Conf Computer Vision and Pattern Recognition (CVPR), 2019: 4401-10.
18. Thies J, Zollhöfer M, Stamminger M, Theobalt C, Nießner M. Face2Face: Real-time face capture and

reenactment of RGB videos. Proc. IEEE/CVF Conf Computer Vision and Pattern Recognition (CVPR), 2016: 2387-95.
19. Sohl-Dickstein J, Weiss EA, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. Proc. Int. conf. machine learning (ICML), 2015: 2256-65.
20. Yang L, Zhang Z, Song Y, Hong S, Xu R, Zhao Y, Zhang W, Cui B, Yang MH. Diffusion models: A comprehensive survey of methods and applications. ACM Comput Surv 2023; 54(4): 1-39.
21. Sun K, Chen Z, Lin X, Sun X, Liu H, Ji R. Conditional diffusion models for camouflaged and salient object detection. IEEE Trans Pattern Anal Mach Intell 2025; 47(4): 2833-48.
22. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022:10674-85. DOI: 10.1109/CVPR52688.2022.01042.
23. Midjourney. [Full text freely available at: https://www.midjourney.com/home].
24. Yang WC, Jiang J, Chen CH. A fast source camera identification and verification method based on PRNU analysis for use in video forensic investigations. Multimed Tools Appl 2021; 80: 6617-38.
25. Kawabe A, Haga R, Tomioka Y, Shin J, Okuyama Y. A dynamic ensemble selection of deepfake detectors specialized for individual face parts. Electronics 2023; 12(18): 3932.
26. Mansoor N, Iliev A. Explainable AI for deepfake detection. Appl Sci 2023; 15(2): 725.
27. Jung T, Kim S, Kim K. DeepVision: Deepfakes detection using human eye blinking pattern. IEEE Access 2020; 8: 83144-54. DOI: 10.1109/ACCESS.2020.2988660.
28. Seibold C, Wisotzky EL, Beckmann A, Kossack B, Hilsmann A, Eisert P. High-quality deepfakes have a heart! Frontiers in Imaging, 4, 2025.
29. Shahzad SA, Hashmi A, Khan S, Peng YT, Tsao Y, Wang HM. Lip sync matters: A novel multimodal forgery detector. Proc. 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, 2022; 1885-92. DOI: 10.23919/APSIPAASC55919.2022.9980296.
30. Herrera-Tapias BA, Guzmán DH, Zambam NJ, Turatti L, Rodríguez FA, Fröhlich S, Staffen MR, Muñoz PC, Reyes AG, de la Torre GS, Duarte NR, Ramos EP, Algorithmic discrimination and explainable artificial intelligence in the judiciary: a case study of the Constitutional Court of Colombia. Procedia Comput Sci 2025; 257: 1227-32.