DIGITAL WATERMARKING AND THE FIGHT AGAINST DEEPFAKE CONTENT

Bagus Anggita Yogatama, H.A. Danang Rimbawa
Cyber Defense Engineering, Indonesia Defense University
Kawasan IPSC Sentul, Jl. Anyar, Sukahati, Kec. Citeureup, Kabupaten Bogor, Jawa Barat 16810
Bagus.yogatama@tp.idu.ac.id, Bagus Anggita Yogatama

ABSTRACT

The rapid advancement of deepfake technology has created serious challenges to the authenticity and security of digital media. Deepfake content generated using artificial intelligence has increasingly been misused for misinformation, fraud, and identity manipulation, reducing public trust in digital platforms. This study examined digital watermarking as a proactive defense mechanism against deepfake manipulation by embedding imperceptible yet robust identifiers into multimedia content. A simulation-based approach was applied to evaluate watermark performance under several deepfake generation techniques, including FaceSwap, DeepFakes, and NeuralTextures. The evaluation focused on robustness, imperceptibility, and resistance to common post-processing operations such as compression, noise addition, and format conversion. Experimental results demonstrated that the proposed watermarking method maintained reliable detection performance while preserving visual quality, as indicated by acceptable Peak Signal-to-Noise Ratio (PSNR) values. Furthermore, the embedded watermark showed resilience against adversarial attempts to remove or distort the hidden information. These findings suggest that digital watermarking can serve as an effective complementary approach to deepfake detection systems. Overall, this research highlighted the importance of digital watermarking in strengthening media integrity amid the growing sophistication of synthetic media technologies.

Keywords : *Artificial Intelligence; Content; Cryptography; Deepfake Content; Digital Watermarking.*

1. INTRODUCTION

The rapid advancement of artificial intelligence has led to the emergence of deepfake technology, which enables the creation of highly realistic but falsified multimedia content. Deepfakes have been increasingly exploited for malicious purposes, including misinformation dissemination, identity fraud, privacy violations, and intellectual property abuse. The widespread circulation of manipulated images and videos through social media and digital platforms has raised serious concerns regarding the authenticity and trustworthiness of digital media. As a result, ensuring media integrity has become a critical challenge in modern digital communication environments. [1] [2].

Over the past few years, numerous studies have addressed the role of watermarking in media authentication and deepfake mitigationdemonstrated the effectiveness of robust image watermarking techniques for resisting compression and noise [4]. Similarly, explored video watermarking schemes to enhance copyright protection and tamper detection in streaming platforms [3]. More recent studies have integrated watermarking with deepfake detection; for example, proposed an adversarial-resistant watermarking framework that survives face-swapping attacks [4]. In addition, hybrid models that combine watermarking with machine learning-based detectors have shown improved resilience against GAN-based manipulations [5]. These works highlight watermarking's evolution from a copyright tool to a defense mechanism against AI-generated forgeries.

Despite significant progress in deepfake detection techniques, most existing approaches rely on post-hoc forensic analysis, which attempts to identify manipulation after the content has already been distributed. Such detection-based methods often suffer from reduced effectiveness when manipulated media undergo common post-processing operations such as compression, resizing, or format conversion. Furthermore, many existing watermarking schemes face challenges in achieving an optimal balance between robustness, imperceptibility, and computational efficiency, especially against increasingly sophisticated deepfake attacks.

This study aims to investigate the effectiveness of digital watermarking as a proactive defense mechanism against deepfake manipulation. The specific objectives of this research are to:

a. Evaluate the robustness of digital watermarking under various deepfake generation techniques and post-processing conditions.
b. Assess the impact of watermark embedding on the visual quality of multimedia content.
c. Examine the potential role of watermarking as a complementary approach to existing deepfake detection systems for media authentication.

To address the identified problems, this research adopts a simulation-based experimental approach. Digital watermarks are embedded into multimedia content and subsequently subjected to multiple deepfake manipulation techniques and post-processing operations. The performance of the watermarking scheme is evaluated using robustness and visual quality metrics. The results are expected to provide insights into the practicality of digital watermarking

for enhancing media integrity and supporting its integration into broader cybersecurity and content verification frameworks.

2. LITERATURE REVIEW

2.1. Importance of Digital Watermarking

The emergence of deepfakes has raised concerns about the authenticity of digital media. Mirsky and Lee (2021) argue that watermarking can act as a proactive solution to deepfake detection by embedding source-specific information into media during creation [1]. This embedded data can be extracted later to verify the authenticity of the content. Verdoliva (2020) also emphasizes the synergy between watermarking and media forensics in identifying manipulated content, asserting that watermarking provides a tangible method for tracing the origin and authenticity of multimedia [2][3].

2.2. Deepfake Threats.

Deepfakes, created using advanced AI, pose significant threats across multiple domains, including security, finance, politics, and public trust. Below are some key points about the risks associated with deepfakes:

a. Impersonation and Financial Fraud
   Deepfakes have been used to impersonate high-profile individuals for financial gain. For example, fraudsters cloned voices of executives to authorize fraudulent bank transfers, resulting in substantial financial losses. In one case, a fake video conference was used to deceive an employee into transferring $25 million to a fraudulent account. These attacks exploit the realism of synthetic media, undermining trust in verification systems [4][5].

b. Disinformation Campaigns
   Deepfakes are increasingly employed in disinformation campaigns, spreading false narratives to influence public opinion, elections, and societal stability. They can generate highly convincing fake content, making it difficult to discern truth from deception [4].

c. Challenges in Cybersecurity
   As deepfake technology becomes more accessible, organizations face growing challenges in defending against these threats. Cybercriminals leverage deepfakes for spear-phishing and social engineering attacks, bypassing traditional authentication methods like voice verification [4][5].

d. Impact on Trust and Public Safety
   The proliferation of deepfakes erodes public trust in media and technology, creating a climate of skepticism. This distrust can have broader implications, including undermining the credibility of genuine content [4].

The term deepfakes is commonly used to describe face replacement techniques based on deep learning; however, it also refers to a specific manipulation method that initially emerged through online forums. To avoid ambiguity, this study uses the term DeepFakes to denote that particular method. Several publicly available implementations of DeepFakes exist, including FakeApp and the faceswap project hosted on GitHub [8]. The technique operates by substituting the facial region in a target video with a face extracted from a source video or image dataset. It relies on two autoencoders that share a common encoder, each trained to reconstruct facial images corresponding to the source and target identities. Prior to training, a face detection algorithm is applied to locate, crop, and align facial regions. During the generation phase, the encoder–decoder pair associated with the source identity is applied to the target face, producing a synthesized output. This output is then seamlessly integrated into the original frame using Poisson image editing. In this study, the GitHub-based faceswap implementation was utilized with minor modifications, replacing manual data selection with an automated data loading mechanism. Default training parameters were maintained for each video pair. Given the substantial computational cost of training these models, the trained networks were preserved and reused to enable the generation of additional manipulated content. [6][7].

2.3. Mitigation Effort

To address the escalating threats posed by deepfakes, experts emphasize the necessity of employing a multifaceted approach that integrates technical, regulatory, and educational strategies. Technological advancements play a pivotal role, with researchers and organizations focusing on the development of sophisticated detection systems capable of identifying AI-generated content. These detection tools aim to safeguard individuals and organizations from the proliferation of deepfakes, especially in critical areas such as security, misinformation, and privacy. Complementing these technological solutions, regulatory measures like the European Union's Digital Services Act have been introduced to enforce accountability among digital platforms, requiring them to monitor and moderate the dissemination of AI-driven disinformation actively. Such legal frameworks are crucial in establishing a structured response to the misuse of deepfake technology. Simultaneously, public education is deemed essential to empower individuals with the knowledge to discern deepfake content and understand its implications. Awareness campaigns strive to inform the public about the risks associated with deepfakes and promote resilience against their manipulative effects. Together, these combined efforts seek to mitigate the societal and organizational risks posed by deepfakes, fostering a safer digital environment while maintaining the integrity of information. [3][8]

3. METHODOLOGY

3.1. Simulation Overview

This simulation evaluates the effectiveness of digital watermarking in mitigating deepfake content by embedding robust and imperceptible identifiers into multimedia files. Various watermarking techniques are applied to images, videos, and audio while examining parameters such as watermark strength, visibility, and resistance to manipulation.

To reflect realistic conditions, synthetic deepfake media are generated using advanced manipulation tools, followed by intentional attempts to alter or remove the embedded watermarks. The simulation includes common post-processing operations such as compression, noise addition, and format conversion. Performance is assessed using metrics including Peak Signal-to-Noise Ratio (PSNR), Bit Error Rate (BER), and watermark extraction accuracy, providing insight into the capability of watermarking to protect digital media from deepfake threats.

3.2. Face Manipulation Methods

Over the past two decades, virtual face manipulation has advanced rapidly and attracted extensive research attention. Early contributions include image-based and 3D reconstruction approaches for facial animation and face replacement, such as Video Rewrite and automated face-swapping methods using single-camera footage. Subsequent techniques improved realism by preserving facial expressions and synchronizing mouth movements, supported by high-quality 3D facial capture to enhance visual fidelity [10].

A major milestone in facial reenactment was the development of real-time expression transfer systems. By leveraging consumer-grade RGB-D cameras, these methods reconstructed and tracked 3D facial models of source and target subjects, enabling seamless expression transfer in video streams. This line of research culminated in the Face2Face framework, which combined 3D facial modeling and image-based rendering for realistic manipulation in everyday videos. Extensions of this approach supported virtual reality applications and inspired further advances such as NeuralTextures, audio-driven lip synchronization, and image-to-image translation techniques for enhanced facial realism.

More recent progress has been driven by deep learning, particularly generative adversarial networks (GANs). GAN-based methods have enabled sophisticated facial image synthesis tasks, including aging, pose variation, and attribute manipulation. Techniques such as Deep Feature Interpolation and Fader Networks demonstrated effective control over facial attributes, while progressive growing strategies addressed resolution limitations, allowing the generation of high-quality, photorealistic facial images and setting new benchmarks in visual realism. [11].

3.3. Mutimedia Forensics

Multimedia forensics focuses on verifying the authenticity, origin, and integrity of visual content without relying on embedded security mechanisms. Early methods primarily used handcrafted features to detect statistical or physical artifacts introduced during image acquisition. Over time, these approaches have been increasingly replaced by convolutional neural network (CNN)-based methods employing both supervised and unsupervised learning. In video forensics, much research has addressed relatively simple manipulations such as frame insertion or removal, duplication, interpolation changes, copy–move tampering, and chroma-key compositing.

Recent research has shifted toward detecting facial manipulations, including computer-generated faces, morphing, splicing, face swapping, and deepfakes. Some approaches exploit specific synthesis artifacts, such as irregular eye blinking or inconsistencies in color, texture, and geometry, while others rely on deep learning models to capture subtle low-level and high-level feature discrepancies. Despite strong performance, robustness remains a challenge, especially when manipulated content undergoes common post-processing operations like compression or resizing, which can obscure forensic traces and reduce detection reliability.

To address these real-world challenges, this study employs a dataset designed to reflect realistic conditions by incorporating manipulated videos collected from uncontrolled sources. The videos are compressed at multiple quality levels to simulate typical content distribution scenarios. This large and diverse dataset provides a valuable benchmark for developing more robust and accurate facial forgery detection methods. [12]

3.4. Forensic Analysis Datasets

Traditional forensic datasets are generally created through extensive manual effort and under controlled conditions, often focusing on isolating specific artifacts such as camera-related traces. Although several datasets address image manipulation, fewer resources target video forensics. Notable examples include the MICC F2000 dataset with 700 copy-move forgeries, the IEEE Image Forensics Challenge Dataset containing 1,176 manipulated images, the Wild Web Dataset with 90 real-world cases, and the Realistic Tampering dataset comprising 220 forged images.

In the context of facial manipulation, several datasets have been proposed to support deepfake research. Zhou et al. introduced a collection of 2,010 images generated using FaceSwap and SwapMe, while Korshunov and Marcel released a dataset of 620 deepfake videos derived from multiple source videos involving 43 individuals. On a larger scale, the National Institute of Standards and Technology (NIST) published a comprehensive benchmark consisting of approximately 50,000 forged images and

around 500 manipulated videos, covering both local and global editing scenarios.

3.5. Simulation Steps

The simulation process to evaluate the efficacy of digital watermarking against deepfake content involves the following detailed steps:

a. Face Swap
FaceSwap is a graphics-based method that replaces the facial region of a target video with a face extracted from a source video. The technique relies on sparsely detected facial landmarks to fit a 3D template model using blendshapes. This model is projected onto the target frame by aligning it with local landmarks and transferring texture information from the source face. The synthesized face is then smoothly blended into the target image with additional color correction. This process is repeated for each corresponding frame pair until the video ends. Due to its lightweight design, FaceSwap can operate efficiently on standard CPU hardware.

b. Deep Fakes
The term *deepfakes* is commonly associated with deep learning–based facial replacement; however, it also refers to a specific manipulation technique that originally emerged within online communities. To distinguish this method, the capitalized term DeepFakes is used throughout this study. Several open-source implementations are publicly available, including FakeApp and the faceswap project on GitHub. The method replaces the facial region in a target video with a face derived from a separate source video or image set. It is based on two autoencoders sharing a common encoder, with each decoder trained to reconstruct facial images of either the source or target identity. Prior to training, facial regions are detected, cropped, and aligned to ensure consistency. During manipulation, the encoder decoder model trained on the source identity is applied to the target face, and the generated output is blended into the original frame using Poisson image editing [15]. For dataset construction, the GitHub-based faceswap implementation was employed with a minor modification that automated the training data selection process. Default parameters were used for each video pair, and due to the high computational cost of training, the resulting models were retained and included to facilitate additional manipulations under varying post-processing conditions.

c. Face2Face
Face2Face [16] is a facial reenactment method that transfers expression dynamics from a source video to a target video while preserving the target's identity. The original framework requires two input videos and manual keyframe selection to build a detailed 3D facial model for realistic rendering under varying expressions and lighting conditions. In this study, the Face2Face pipeline was modified to operate fully automatically. Each video undergoes a preprocessing stage in which an initial set of frames is used to reconstruct a temporary 3D facial identity, followed by continuous expression tracking across the sequence. Keyframes are selected automatically based on extreme left and right head rotations. Using the reconstructed identity, the system estimates per-frame facial expressions, rigid head pose, and illumination parameters in accordance with the original Face2Face methodology.

d. Neural Textures
Thies et al. [17] proposed a facial reenactment technique based on the NeuralTextures rendering framework, which jointly learns a neural texture of the target subject and a dedicated rendering network directly from video data. Training combines photometric reconstruction loss with adversarial learning; in this study, a patch-based GAN loss similar to Pix2Pix [18] is applied. The approach requires accurate facial geometry tracking during both training and inference, which is obtained using the Face2Face tracking module. In this implementation, manipulation is limited to mouth-related expressions, while the eye region is preserved. This constraint removes the need for additional conditioning on eye movements, as required by alternative methods such as Deep Video Portraits.

e. Post processing Video Quality
To simulate realistic distribution conditions, the manipulated videos are generated at multiple quality levels commonly found on social media platforms. Since uncompressed videos are rarely shared online, all outputs are encoded using the H.264 compression standard. High-quality versions are produced with a low compression level (quantization parameter of 23), resulting in minimal visual degradation, while low-quality versions use a higher quantization parameter of 40, leading to noticeably reduced visual quality.

3.6. Forgery Detection

Forgery detection is approached as a per-frame binary classification problem applied to manipulated video content. The evaluation process includes both manual and automated detection methods. For all experiments, the dataset is divided into fixed training, validation, and test sets to ensure consistent benchmarking.

3.7. Forgery Detection by Human Observers

To assess human capability in detecting forgeries, a user study was conducted involving 204 participants, the majority of whom were computer science university students. This evaluation serves as a baseline for comparing automated detection methods.

3.8. Layout of the User Study

Participants received a brief introduction to the binary classification task before being asked to evaluate randomly selected images from the test set. The chosen images included variations in both quality and manipulation method, with an even 50:50 distribution of pristine and altered images. To simulate conditions similar to casual viewing on social media, each image was displayed for a randomly assigned duration of 2, 4, or 6 seconds before being hidden. The classification question whether the image was "real" or "fake" was presented only after the image disappeared to ensure that the allocated viewing time was dedicated solely to inspection. Each participant viewed a total of 60 images, resulting in 12,240 human classification decisions across the study.

3.9. Automated Forgery Detection

Automated forgery detection relies on machine learning models trained to differentiate manipulated frames from authentic content. To ensure fair comparison with human evaluation, the models use identical training, validation, and test splits, incorporating both high-quality (HQ) and low-quality (LQ) videos to assess robustness under varying compression levels. Feature extraction ranges from handcrafted descriptors to deep convolutional neural network (CNN) embeddings, with CNN-based models effectively capturing subtle inconsistencies in texture, lighting, and geometric structure caused by manipulation. The evaluation also considers post-processing effects, such as color correction and blending, which can obscure forensic traces. Detection performance is assessed using accuracy, precision, recall, and AUC-ROC metrics and directly compared with human performance. The results demonstrate that automated methods offer consistent performance on large datasets, although their accuracy may decrease when manipulations occur in heavily compressed or degraded videos.

3.10. Evaluation

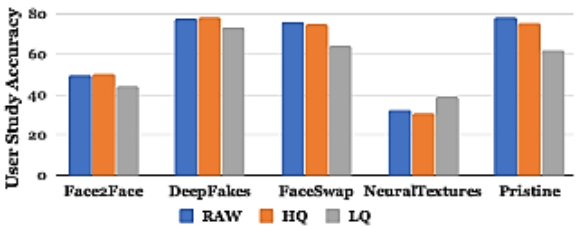

Figure 1. Forgery detection results from user study across different video qualities (RAW, HQ, and LQ) and manipulation types (Face2Face, DeepFakes, FaceSwap, NeuralTextures, and Pristine)

The user study results shown in Figure 1 indicate a clear relationship between video quality and human ability to detect manipulated content. Detection accuracy declined as video quality decreased, with average performance of 68.69% for raw videos, 66.57% for high-quality videos, and 58.73% for low-quality videos. Variations in viewing time (2, 4, or 6 seconds) had no significant impact, suggesting that video quality plays a more critical role than observation duration in influencing detection accuracy.

Analysis of responses from 204 participants reveals that certain manipulation techniques were more challenging to detect. Face2Face and NeuralTextures exhibited lower detection rates due to their subtle visual alterations rather than obvious semantic inconsistencies. In particular, NeuralTextures achieved detection accuracy below random chance for raw and high-quality videos, with only marginal improvement under low-quality conditions. Conversely, DeepFakes and FaceSwap were more readily identified, while unmanipulated videos consistently achieved the highest recognition rates.

3.11. Automatic Forgery Detection Methods

The forgery detection pipeline used in this study is illustrated in Figure 2. The process begins with the application of a state-of-the-art face tracking algorithm, which enables accurate localization of facial regions in video frames. A conservative cropping strategy is then employed, enlarging the tracked area by a factor of 1.3 to ensure the reconstructed face is fully enclosed. This cropped face region is subsequently passed into a classification network trained specifically for forgery detection. Unlike naïve methods that use the entire image as input, this domain-specific approach significantly improves detection accuracy by focusing on the manipulated region.

Figure 2. Domain-specific forgery detection pipeline for facial manipulations: the input image is processed through robust face tracking, followed by face extraction, and the cropped region is fed into a classification network that outputs the final prediction (real or fake)

To validate the robustness of the pipeline, multiple classification methods were evaluated, including several widely used approaches within the forensic community. The experiments demonstrated that the XceptionNet-based classifier outperformed all other variants, achieving superior performance in distinguishing between authentic and manipulated content. This finding highlights the importance of incorporating domain-specific knowledge and optimized classification networks for advancing deepfake detection systems.

3.12. Detection based on Steganalysis Features

Detection performance was assessed using a steganalysis-based approach that relies on manually designed features rather than deep learning. The method extracts co-occurrence statistics from high-pass filtered images by analyzing four-pixel patterns along both horizontal and vertical directions, resulting in a feature vector with a total dimensionality of 162. These extracted features are subsequently used to train a linear Support Vector Machine (SVM) classifier. This approach was previously recognized as the top-performing method in the first IEEE Image Forensic Challenge. For evaluation, a centrally cropped facial region of size 128 × 128 pixels was provided as input to the detection framework. Although this handcrafted feature-based method significantly exceeds human-level accuracy when applied to uncompressed images, its performance degrades substantially under compression. In low-quality video scenarios, the detection accuracy falls below that of human observers.

4. RESULT

4.1. Result

This study presents a comprehensive benchmark to evaluate facial forgery detection models under realistic conditions. The benchmark extends the existing dataset by adding 1,000 manipulated video samples generated using the same deepfake techniques as the main dataset to maintain consistency. To increase realism and difficulty, the videos are further subjected to common post-processing operations, including compression, resolution reduction, and color distortion conditions frequently encountered in online media distribution that significantly challenge detection systems [9].

For each benchmark video, a single frame identified by experts as particularly difficult to classify is selected for evaluation. These frames serve as the basis for testing scenarios with limited visual information. Ground truth labels are concealed from evaluators, and all assessments are conducted through a centralized and secure server, ensuring unbiased and blind evaluation that closely reflects real-world deployment conditions [10].

Baseline performance is measured using previously trained detection models applied to lower-quality inputs to simulate practical deployment constraints. While performance trends remain consistent with results on the main dataset, overall accuracy decreases due to randomized and degraded video quality. This performance gap highlights the need for detection systems that are robust to diverse distortions and artifacts. Consequently, the benchmark functions not only as an evaluation tool but also as a driver for developing more resilient and generalizable deepfake detection methods [11].

4.2. Analysis of the Effectiveness of Digital Watermarking Against Deepfake Content

The effectiveness of digital watermarking as a defense against deepfake content is evaluated using key performance indicators, including robustness, imperceptibility, resilience to adversarial attacks, and scalability. Together, these metrics assess the practicality of watermarking for protecting multimedia content against deepfake manipulation.

Robustness refers to the ability of the embedded watermark to survive transformations and attacks while remaining detectable. Experimental results show detection rates of approximately 80–90% under common deepfake manipulations, such as facial morphing and frame interpolation. The watermark also remains resilient to resizing, compression, and noise addition, indicating reliable performance under realistic distribution conditions.[12]

Imperceptibility evaluates whether watermark embedding affects perceptual quality. Objective assessment using Peak Signal-to-Noise Ratio (PSNR) demonstrates minimal visual degradation, with values remaining above acceptable thresholds for both images and videos. This confirms that watermarking preserves content quality and user experience. [11][13].

The proposed method maintains effective detection even under adversarial attempts to remove or distort the watermark. Compared with conventional techniques such as Discrete Wavelet Transform (DWT), the approach achieves higher detection accuracy and lower error rates, as summarized in Table 1, highlighting its superiority over baseline methods. [5][14].

The embedding and extraction processes exhibit computational efficiency, supporting deployment in high-throughput environments such as social media and streaming platforms. However, extreme transformations, including severe blurring and geometric distortion, remain challenging and may reduce watermark recoverability [15].

Table 1. Comparison of forgery detection accuracy (%) across different methods and manipulation techniques

Method	DF	F2F	FS	NT	Real	Total
Xent Full Image	74.55	75.91	70.87	73.33	51.00	62.40
Steg Features	73.64	73.72	68.93	63.33	34.00	51.80
Cozzolino et al.	85.45	67.88	73.19	78.00	34.40	55.20
Rahmouni et al.	85.45	64.23	56.31	60.07	50.00	58.10
Bayar and Stamm	84.55	73.72	52.52	70.67	46.20	61.60
MesoNet	87.27	56.20	61.17	40.67	72.60	66.00
XceptionNet	96.36	86.86	90.29	80.67	52.40	70.10

5. CONCLUSION

Although modern facial manipulation techniques generate highly realistic deepfake content, this study shows that such forgeries can still be effectively identified using trained detection models. Even in low-quality video scenarios where human observers and handcrafted features perform poorly, learning-based detectors demonstrate strong detection capabilities. The availability of large-scale manipulated video datasets further supports robust and domain-specific training of forensic models [16].

Beyond detection-based approaches, this research highlights digital watermarking as a proactive and resilient defense against deepfake dissemination. By embedding imperceptible yet durable identifiers into digital media, watermarking enables reliable content authentication and source verification. Experimental results confirm that watermarking remains effective against various deepfake manipulations and adversarial attacks while preserving perceptual quality, making it suitable for real-time applications such as social media monitoring and live-stream security [17].

Despite these advantages, several challenges remain, including robustness against extreme manipulations, scalability in high-throughput systems, and adaptation to rapidly evolving generative models. Nevertheless, the findings underscore the importance of digital watermarking in strengthening the security and trustworthiness of digital media ecosystems as deepfake technologies continue to advance [3]

Future research should emphasize integrating digital watermarking with machine learning and AI-based detection systems to improve resilience against increasingly sophisticated deepfake techniques. Such hybrid approaches enable adaptive learning, enhance detection accuracy, and strengthen watermark robustness against evolving manipulation methods, positioning watermarking as a key component of future deepfake mitigation strategies [18].

To support broad adoption, the development of universal standards for digital watermarking is essential. Collaborative frameworks involving researchers, industry, and policymakers can ensure interoperability, consistency, and compatibility across platforms, thereby facilitating reliable and scalable deployment of watermarking technologies in combating deepfakes [19].

Practical validation of digital watermarking should be conducted through pilot implementations in high-risk sectors such as elections, news media, and entertainment. Real-world trials can assess system performance under operational conditions, reveal limitations, and guide refinements to ensure effectiveness across diverse application domains [19][20].

Increasing public and organizational awareness of deepfake risks and the role of digital watermarking is critical for widespread adoption. Educational initiatives and professional training can improve understanding, promote responsible use, and enable effective implementation of watermarking technologies, particularly in vulnerable industries.

As deepfake generation methods continue to evolve, digital watermarking systems must remain adaptable. Continuous monitoring of emerging manipulation techniques and the development of dynamic watermarking strategies are necessary to sustain long-term effectiveness and relevance in rapidly changing technological environments.

REFERENCES

[1] Y. Mirsky and W. Lee, “The Creation and Detection of Deepfakes,” *ACM Comput. Surv.*, vol. 54, no. 1, 2021, doi: 10.1145/3425780.

[2] P. Neekhara, S. Hussain, X. Zhang, K. Huang, J. McAuley, and F. Koushanfar, “FaceSigns: Semi-Fragile Neural Watermarks for Media Authentication and Countering Deepfakes,” 2022. [Online]. Available: http://arxiv.org/abs/2204.01960

[3] L. Verdoliva, “Media Forensics and DeepFakes: An Overview,” *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 910–932, 2020, doi: 10.1109/JSTSP.2020.3002101.

[4] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, “Unmasking DeepFakes with simple Features,” 2019, [Online]. Available: http://arxiv.org/abs/1911.00686

[5] P. Pérez, M. Gangnet, and A. Blake, “Poisson image editing,” *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, 2003, doi: 10.1145/882262.882269.

[6] W. Chen *et al.*, “Real-time and screen-cam robust screen watermarking,” *Knowledge-Based Syst.*, vol. 302, no. February, p. 112380, 2024, doi: 10.1016/j.knosys.2024.112380.

[7] L. H. Singh, P. Charanarur, and N. K. Chaudhary, “Advancements in detecting Deepfakes: AI algorithms and future prospects − a review,” *Discov. Internet Things*, vol. 5, no. 1, 2025, doi: 10.1007/s43926-025-00154-0.

[8] S. Sharma, J. J. Zou, G. Fang, P. Shukla, and W. Cai, “A review of image watermarking for identity protection and verification,” *Multimed. Tools Appl.*, vol. 83, no. 11, pp. 31829–31891, 2024, doi: 10.1007/s11042-023-16843-3.

[9] A. Mehra, A. Agarwal, M. Vatsa, and R. Singh, “Detection of Digital Manipulation in Facial Images (Student Abstract),” *35th AAAI Conf. Artif. Intell. AAAI 2021*, vol. 18, pp. 15845–15846, 2021, doi: 10.1609/aaai.v35i18.17919.

[10] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, “DeepFake Detection for Human Face Images and Videos: A Survey,” *IEEE Access*, vol. 10, pp. 18757–18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

[11] D. Cozzolino, D. Gragnaniello, and L. Verdoliva, “Image forgery detection through residual-based local descriptors and block-matching,” *2014*

IEEE Int. Conf. Image Process. ICIP 2014, pp. 5297–5301, Jan. 2015, doi: 10.1109/ICIP.2014.7026072.

[12] E. Altuncu, V. N. L. Franqueira, and S. Li, "Deepfake: definitions, performance metrics and standards, datasets, and a meta-review," *Front. Big Data*, vol. 7, no. Dl, pp. 1–31, 2024, doi: 10.3389/fdata.2024.1400024.

[13] B. Tornés, E. Boros, A. Doucet, P. Gomez-Krämer, and J.-M. Ogier, "Detecting Forged Receipts with Domain-Specific Ontology-Based Entities & Relations," 2023, pp. 184–199. doi: 10.1007/978-3-031-41682-8_12.

[14] D. Cozzolino and L. Verdoliva, "Multimedia Forensics Before the Deep Learning Era," 2022, pp. 45–67. doi: 10.1007/978-3-030-87664-7_3.

[15] D. Noviantama and A. Rahman, "Deepfake: A Review from The Victimology Perspective," vol. 1, pp. 107–128, Jun. 2025, doi: 10.20885/CICL.vol1.iss2.art1.

[16] S. Bose, O. Hathi, K. Pandey, and U. Singh, "The Emerging AI Technology Deepfake," *Med. Leg. Updat.*, vol. 25, pp. 8–15, Aug. 2025, doi: 10.37506/73q84t55.

[17] V. Pirogov and M. Artemev, *Evaluating Deepfake Detectors in the Wild*. 2025. doi: 10.48550/arXiv.2507.21905.

[18] R. Shao, T. Wu, and Z. Liu, "Robust Sequential DeepFake Detection," *Int. J. Comput. Vis.*, vol. 133, pp. 3278–3295, Jan. 2025, doi: 10.1007/s11263-024-02339-6.