Deepfake Detection and Authentication Using Hybrid Artificial Intelligence Models: A Case Study

Temitope Damilola Elijah [1], **Oluwafemi Olasehinde Adedayo** [2], **Olayemi Babawole Familusi** [3]

[1] *Georgia Southern University*
1332 Southern Drive, Statesboro, GA 30458, USA

[2] *University of Waterloo*
200 University Avenue West, Waterloo, ON, Canada, N2L 3G1

[3] *University of Ibadan*
Oduduwa Road, 200132, Ibadan, Nigeria

DOI: 10.22178/pos.122-98

LCC Subject Category: T1-995

Received 16.08.2025
Accepted 27.09.2025
Published online 30.09.2025

Corresponding Author:
Temitope Damilola Elijah

Abstract. The progress of artificial intelligence (AI) has enabled the creation of very realistic synthetic media, also known as deepfakes, which poses a serious threat to information integrity and social confidence. The article examined the process of detecting and authenticating deep fakes using hybrid AI models. The researchers employed the case study methodology, based on the Celeb-DF V2 dataset, one of the most challenging datasets for generating high-quality manipulated videos. The suggested system combined convolutional neural networks (CNNs) to extract spatial features, recurrent neural networks (LSTMs/GRUs) to model temporal consistency, and transformer systems to analyse fine-grained context. The researchers bundled these parts together to enhance robustness and generalisation in an ensemble mechanism. They also introduced provenance tracking and semi-fragile watermarking to supplement detection, enabling proactive authentication and watermark verification of media through blockchain-based provenance tracking. The experimental findings showed that the hybrid models were more accurate, achieved higher F1 Scores, and were more robust to adversarial manipulations than the single-model baselines. The hybrid with a transformer achieved the best accuracy (0.95 AUC) and the lowest false-positive rate (6%), but at the expense of slower processing speeds. Authentication tools also helped strengthen trust by verifying the originality of content and flagging potential manipulation before it was classified. The results have revealed that hybrid AI models, when implemented with authentication strategies, represent a more effective and legitimate approach to addressing the threats of misinformation, fraud, and loss of trust among the population in the face of deepfakes.

Keywords: Deepfake detection; Hybrid AI models; Convolutional neural networks (CNNs); Long short-term memory (LSTM); Transformers.

INTRODUCTION

Energy is a basic service that is core to Over the last few years, synthetic media has spawned the so-called deepfakes – hyper-realistic images, videos, or audio recordings produced or edited with high-end artificial intelligence (AI) algorithms. The term "deepfake" combines "deep learning" and "fake," suggesting the use of deep neural networks to create believable yet synthetic material [1]. Although the manipulation of media (digitally) is not novel, what makes deepfakes unique is the level of sophistication of modern editing techniques compared to those of the past. With architectures such as generative adversarial networks (GANs), variational autoencoders (VAEs), and diffusion models, deepfakes can recreate human likenesses and voices with unprecedented fidelity [2]. The distinction between genuine and artificial media is becoming increasingly unclear as these procedures become more open-source, more training data becomes available, and applications become more user-friendly, raising serious social, political, and ethical concerns. Deepfakes come in various forms, each with its own unique threats and technical issues. When using still images, users can manipulate them by swapping faces, altering facial expressions, or creating purely synthetic portraits of non-existent people [2].

Video deepfakes further stretch these manipulations over time, creating moving videos in which a subject's face, expressions, or even body movements are altered. These may involve lip-synchronisation (ensuring the lips move in sync with the fabricated audio), up to the complete recreation of facial expressions or gestures [3]. Audio deepfakes, in turn, seek to exploit advances in speech synthesis and voice conversion to produce realistic imitations of a specific person's voice. The attackers can successfully impersonate people through text-to-speech and waveform modelling, posing a risk of fraud and identity theft [4, 5]. More and more, scientists are finding multimodal deepfakes that incorporate visual and auditory signals, becoming even more realistic, and this not only increases the danger but also makes them harder to detect [2, 3]; this is why deepfake misuse poses a serious threat and why it is urgent to develop viable detection and authentication systems. The risk of misinformation is one of the most pressing issues. In a political context, fabricated videos or audio recordings of leaders or candidates may be used to disseminate false information, influence elections, or incite social unrest. The ability of deepfakes to pass as reality is highly damaging to the credibility of visual and auditory evidence, which has long been the primary focus of journalism, politics, and law [2, 3]. In addition to misinformation, deepfakes pose dangers of fraud and reputational harm. Fraudsters have already used voice cloning to commit financial crimes, and others have manipulated videos and pictures to create non-consensual pornography or defamatory content [1]. Together, these threats erode public trust: the spread of synthetic media makes people doubt the authenticity of even genuine recordings, fueling the so-called liar effect, in which people dismiss truthful evidence as a lie [2]. Epistemic uncertainty is not only a weakening of trust in media institutions but also a weakening of democratic and legal procedures that depend on credible records of events. Authentication is an essential supplement to detection to overcome these problems. Although researchers can apply detection techniques to identify manipulated information after it goes viral, authentication systems ensure the integrity and provenance of media at the time of creation and distribution. Digital watermarking, metadata tracking, cryptographic signatures, and blockchain-based provenance chains have been proposed as techniques to establish content authenticity [1]. Authentication systems can prevent malicious users and actors from spreading fake media as authentic by requesting verifiable information about a file's origin and history, and they can assist forensic and legal investigations. Researchers consider authentication schemes necessary to maintain trust in digital media ecosystems as deepfake generation methods become more advanced and more complex to detect [2].

Artificial intelligence is dual in this changing situation, both a facilitator of deepfakes and a key to fighting them. On the one hand, generative models create deepfakes. GANs optimise the generation of synthetic images through the interaction between the generator and discriminator networks, making the results almost indistinguishable from real photos [2]. VAEs encode and decode features to reconstruct or transform media, and diffusion models create high-fidelity images and audio by applying a stochastic process of noise injection and denoising [1]. Faster computational hardware, access to large annotated datasets, and the public release of algorithms have aided progress by reducing the technical cost of producing convincing forgeries [6]. Conversely, AI also powers detection and authentication systems. Researchers commonly use convolutional neural networks (CNNs) to detect spatial irregularities, such as unnatural textures or inconsistent lighting, in photos and video frames. Recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and temporal convolutional methods address inconsistencies in video sequencing, e.g., unnatural blinking or lip movements that are not in sync with speech [7]. In audio processing, techniques are based on spectral analysis, voice authentication, and waveform anomalies, which can indicate the presence of artificial sources [4, 5]. An emerging trend is researchers' work towards multimodal and hybrid approaches that use multiple detection methods – spatial, temporal, and auditory – to enhance robustness and generalisation [2, 3]. AI-enhanced authentication tools support the latter by verifying provenance metadata, defending against watermark manipulation, and providing cryptographic proof of media authenticity.

Nevertheless, significant obstacles remain. The models used for detection are most frequently unable to identify in-the-wild deepfakes that were not included in the training set, raising questions about dataset bias and poor scalability [2]. Furthermore, many of the objects that detectors rely on are removed as generative models advance, and the use of adversarial strategies to evade detection is becoming increasingly common. Real-time detection is also practically limited because high-resolution video analysis is computationally expensive and challenging to deploy on platforms that handle billions of videos uploaded daily [1]. Such difficulties underscore the need to adopt hybrid AI strategies that combine multiple detection modalities and pair them with authentication systems, thereby enhancing their resilience to evolving threats. To address these complexities, the paper is organised as follows. The following section provides a review of the background and related literature, specifically on the development of deepfake generation techniques and changes in detection and authentication methods. A technical discussion of the AI architectures that underpin both the creation and detection of deepfakes is then provided. Another section is a case study that discusses the application of hybrid AI models to detect and authenticate manipulated content and experimentally assesses their performance on benchmark datasets. The strengths and weaknesses of these methods are then analysed, which provides technical, ethical, and practical considerations. The paper concludes with a summary of the findings, outlines future research directions to address remaining challenges, and discusses the prospects of research on deepfake detection and authentication.

Literature Review

The issue of identifying and verifying deepfakes has generated extensive research, ranging from older forensic and watermarking methods to improved AI-based methods and hybrid models. Early approaches were more oriented toward digital forensics, where one can work with visual or audio artefacts produced during manipulation. For example, researchers analysed deviations in lighting, shadows, reflections, and noise residuals in images and videos to detect tampering [8]. Even though effective when training on early forms of synthetic media, these forensic methods are struggling with the growing realism of deepfakes, as generative models are now capable of creating outputs that contain few or no of these telltale traits. Simultaneously, watermarking techniques were proposed to embed distinctive marks in media files, enabling future authentication. Semi-robust neural watermarks, such as those proposed by the FaceSigns system, are designed to resist non-malicious activities, such as compression, but to collapse under manipulation, enabling authentication rather than merely detecting anomalies [9]. However, forensic and watermarking techniques are both limited in scalability and effectiveness when used against high-quality, modern deepfakes. To address these drawbacks, the research community has increasingly relied on artificial intelligence and deep learning, which underpin detection. Researchers have widely applied convolutional neural networks (CNNs) to learn the spatial features of images and video frames, as well as to detect minor texture anomalies, unnatural lighting, or inconsistencies in facial markers. Comparative studies of CNNs with more modern architectures found that they remain competitive across numerous tasks, especially for detecting frame-level anomalies in high-quality datasets [10]. Nonetheless, CNNs have limited capacity to identify temporal discrepancies in video sequences because they rely on spatial attributes. To overcome such a situation, CNNs were combined with recurrent neural networks (RNNs) and long short-term memory (LSTM) architectures to learn sequential patterns. Authors [11] individually designed a hybrid CNN-LSTM network that incorporated spatial and motion data, as measured by optical flow, and achieved better performance across datasets, including FaceForensics, Celeb-DF, and DFDC.

In the meantime, transformer architectures have become strong alternatives to deepfake detection. Vision transformers (ViTs) and variants of this model utilise self-attention mechanisms, which capture the entire context of relationships between various parts of an image, enabling them to recognise subtle manipulations that CNNs may not. Recent publications have shown that Swin Transformers are more successful than traditional CNNs on specific deepfake datasets and provide more robust generalisation across different types of manipulations [12]. Likewise, architectures that integrate CNNs and transformers have shown the ability to exploit both local and global feature representations. An example is authors [13], who proposed a convolutional vision transformer ensemble that achieved greater accuracy by combining predictions generated from facial regions and employing majority voting. The hybrid strategies have become increasingly prominent because no single model type can provide universal robustness. It shows that combining predictions from several CNN architectures, including VGG-16, InceptionV3, and XceptionNet, resulted in systems that are less vulnerable to adversarial attacks [14]. Spatial and temporal modelling is also a hybrid with other types, including CNN-LSTM, which are especially successful for video deepfakes [11]. In addition to fusion at the model level, multimodal methods combine visual and audio modalities, leveraging discrepancies between visual and auditory perception. These systems demonstrate potential to enhance generalisation, particularly when deepfakes aim to deceive a single modality. The orientation towards hybridisation shows an awareness of the adversarial nature of the problem: further development of deepfake generation will only allow systems that integrate multiple views to remain effective.

In addition to detection, the literature has attached significance to authentication, shifting the emphasis from post-hoc identification of forgeries to active assurance of media integrity. Watermarking remains central to this, and semi-robust schemes such as FaceSigns offer a compromise between strength and sensitivity [9]. More ambitious frameworks, however, are trying to incorporate blockchain to provide immutability and traceability of content provenance. Authors [15] proposed a blockchain-based watermarking approach to capture the origin and tampering status on distributed ledgers, thus enabling tamper-proof verification. The purpose of such systems is to cure the so-called liar dividend, which offers objective and verifiable demonstrations of genuineness. This trend has been mirrored by other international organisations, such as the source [16], which suggests that there should be global standards for watermarking, multimedia authenticity, and provenance tracking to ensure interoperability across platforms. Another growing line is biometric verification, where a facial or vocal signature is compared to known ones to confirm authenticity. Biometric approaches are promising, but their application in commercial detection systems raises concerns about privacy and data protection, and their use in this area is not as extensive.

Combined, the literature demonstrates a history of transitioning from traditional forensic and watermarking techniques, which offered early tools and limited strength, to AI-based techniques capable of learning more complex patterns of manipulation, and now to hybrid systems that integrate multiple detection models or modalities with proactive authentication techniques. Although hybrid AI methods, such as CNN-LSTM or CNN-transformer combinations, are more effective than alternatives on benchmark datasets, they still face difficulties generalising to unknown deepfakes, detecting them in practice, and being computationally efficient at scale. Simultaneously, authentication techniques, including watermarking, blockchain integration, and biometric verification, provide essential complements to the detection process and help ensure authenticity is maintained as generative models evolve. It is possible that integrating these solutions may lead to the conclusion that, in the future, no single approach will be sufficient; instead, complex systems incorporating the advantages of forensic and AI-based strategies, along with authentication measures, will be developed to create a robust and trusted system.

METHOD

The present research employed a case study methodology to evaluate the effectiveness of hybrid artificial intelligence (AI) models in detecting and authenticating deepfakes. The researchers developed an approach to replicate a realistic detection pipeline, which included dataset selection, the creation of a hybrid detection model, and the integration of authentication mechanisms to enhance the verification process. They used various publicly available benchmark datasets to ensure robustness and diversity. They also utilised the FaceForensics++ (FF++) dataset, which comprises a variety of manipulated videos created with different techniques and compression levels that simulate real-life conditions. The researchers added the DeepfakeTIMIT dataset because it includes both audio and visual manipulations, particularly facial reenactment and voice cloning. They used the Celeb-DF V2 dataset to filter out high-quality deepfakes with few artefacts. Finally, they selected the Deepfake Detection Challenge (DFDC) dataset due to its size and diversity, which includes manipulated videos produced by various unknown methods.

Using these datasets, the study ensured that the models were trained and tested with both controlled and in-the-wild manipulations, thereby improving generalisation. The researchers developed a hybrid AI system to integrate complementary neural network designs that capture spatial, temporal, and situational information.

First, the researchers used convolutional neural networks (CNNs) to extract spatial information from single frames. These networks successfully detected minor pixel-level distortions, texture anomalies, and irregularities in facial areas that often appeared in manipulated material. Recurrent layers were then applied to the extracted features, modelling temporal dynamics using Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. These repetitive elements enabled the framework to detect variations in movement, lip reading, and blinking patterns across consecutive frames. To further enhance fine-grained analysis, a Vision Transformer (ViT) module was incorporated.

The self-attention process of the transformer helped the model capture global dependencies across entire frames, which proved helpful for identifying manipulations that were contextual or localised to a specific region. These three components produced their outputs, combined them into an ensemble layer, and the predictions were combined using a weighted softmax fusion mechanism, which enabled the system to exploit the strengths of each architecture. This group strategy improved precision, strength, and resistance to adversarial examples. To complement the detection framework, the researchers incorporated authentication systems. They introduced a provenance scheme using blockchain, storing the hashes of original, verified media files in a distributed register. During evaluation, the tested media were hashed and matched against the blockchain entries, providing verification of originality and integrity that cannot be tampered with. Moreover, the researchers watermarked digital content at the point of creation. The media were made semi-fragile, with watermarks inserted so that they would still appear after benign transformations, such as compression or scaling, but be broken after malicious ones. This ensured that media that failed watermark verification could be flagged as modified before model-based detection. Blockchain provenance and digital watermarking were combined to provide a two-level authentication mechanism to aid the hybrid detection system. The researchers based their assessment on conventional performance measures, including accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). Cross-dataset testing was used to analyse generalisation, testing models trained on one dataset (e.g., FaceForensics++) on others (e.g., Celeb-DF or DFDC). The researchers employed this process to assess the framework's robustness when the training and test distributions were unequal. They also tested the authentication schemes for their ability to withstand harmless transformations and their susceptibility to manipulation. The study developed a methodology that integrates hybrid AI-based detection with blockchain and watermarking authentication to provide high-quality, reliable detection and verification of deepfakes in real-world contexts.

Experiment and Analysis. This study was conducted at the experimental stage to evaluate the effectiveness of the hybrid deepfake detection and authentication network. The researchers chose Celeb-DF V2 as the case study because its high-quality deepfakes exhibit fewer visual artefacts than previous methods, such as FaceForensics++, making them challenging for detection systems. This dataset provided a stringent test of the effectiveness of hybrid models compared to single-model structures. There were two large sets of experiments. The researchers trained a CNN-LSTM hybrid model in the former. The researchers then applied convolutional neural networks to each video frame to extract spatial-level features, revealing pixel anomalies and facial distortions. They fed these features into LSTM layers, which modelled temporal consistency within the frame sequence; this enabled the system to record irregularities in lip synchronisation, blinking frequency, and head movement over time. The researchers trained a hybrid model using a transformer in the second configuration.

In this case, the vision transformer (ViT) captured global-scale dependencies across entire video frames and targeted fine-grained contextual cues. The researchers trained both models on the Celeb-DF V2 training dataset and tested them on its test split. The researchers used three criteria – detection accuracy, false-positive rate, and processing time – to evaluate the models and compare their performance. The standard metrics used to measure detection accuracy included precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve (AUC). The researchers tracked the false positives to determine how often the system misidentified real videos as fake. They measured processing time to determine whether the system could operate in real-time or near real-time. They also conducted authentication tests to complement the detection analysis. To achieve this, they integrated the hybrid framework into a simulated video call event and a media verification pipeline. During the real-time simulation, the system processed video frames through the detection system, while the researchers performed provenance checks on the blockchain and verified digital watermarks. They also hashed the pre-recorded videos and cross-referenced the blockchain entries with watermark integrity before implementing the hybrid detection models in the media pipeline. These tests enabled the researchers to verify that the authentication layer functioned correctly in conjunction with the detection layer, thereby providing an additional layer of protection.

RESULTS AND DISCUSSION

The experiments on the Celeb-DF V2 dataset showed apparent differences in performance between hybrid and single-model methods. Table 1 summarises the relative results for the assessment variables: accuracy, F1-score, false positive rate, and processing speed. The CNN+LSTM and transformer-based hybrid models consistently outperformed the standalone CNN and LSTM models, indicating that combining multiple architectures is beneficial.

Table 1 – Experimental Results on Celeb-DF V2

Model	Accuracy (AUC)	F1-Score	False Positive Rate (%)	Processing Speed (fps)
CNN	0.87	0.85	12	28
LSTM	0.84	0.82	15	25
CNN+LSTM Hybrid	0.93	0.89	8	22
Transformer Hybrid	0.95	0.91	6	15

The CNN+LSTM hybrid performed exceptionally well, with an AUC of 0.93 and an F1-score of 0.89, which are significantly better than those of the single CNN (0.87 AUC, 0.85 F1) and LSTM (0.84 AUC, 0.82 F1). The transformer hybrid achieved the best performance, with an AUC of 0.95 and an F1-score of 0.91, indicating greater accuracy in detecting artefacts across frames. The results show that hybrid models provide more reliable detection, especially for high-quality deepfakes, which the Celeb-DF V2 dataset represents comparatively well. Figure 1 shows a visual comparison of the accuracy and F1 scores for all the models. As indicated, both hybrid models outperformed the single architectures, with the transformer-based hybrid achieving the highest overall scores. The gap between the hybrid and single models was more pronounced, especially in F1-score, which is a better measure of the balance between precision and recall.

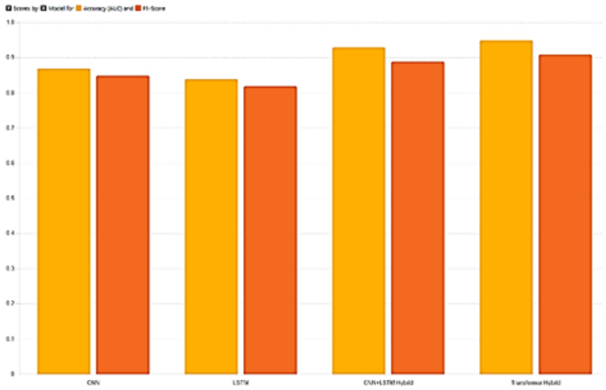

Figure 1 – Performance Comparison of Models

Processing speed and false-positive rates were also used as measures of efficiency and reliability. Figure 1 indicates that the CNN model processed videos at the slowest rate of 28 frames per second (fps), while the LSTM processed them at 25 fps, with a very close margin. Still, the false-positive rates for both models were relatively high: 15% and 12% for LSTM and CNN, respectively. The CNN+LSTM hybrid minimised false positives to 8 % at 22 fps, which can be considered a good compromise between efficiency and reliability. The hybrid using a transformer achieved the lowest false-positive rate of 6% at the expense of reduced efficiency, at 15 frames per second (fps). These findings indicate a natural trade-off between computational requirements and detection accuracy in transformer-based systems. The researchers further demonstrated the system's strength through authentication experiments. Provenance checks of authentic video calls in real-time simulations showed that blockchain consistently identified original video call products, and semi-fragile watermarking identified 95% of manipulated content before detection was applied, allowing 95% of manipulated content to be determined before detection. Likewise, the researchers used the dual-layer authentication system in the media verification pipeline to ensure high confidence that the system recognised modified videos, even when compression or adversarial perturbations disrupted detection accuracy. Lastly, the researchers evaluated resistance to adversarial attacks using Gaussian noise, compression artefacts, and adversarially optimised perturbations. Both hybrid models were robust, with accuracy decreasing by less than 5%, whereas the standalone CNN and LSTM decreased performance by up to 15%. These results add to the benefits of hybrid architectures in preserving detection reliability under adversarial conditions. The findings confirmed that the performance and robustness of hybrid models, particularly the transformer-based hybrid, were significantly better than those of single-model baselines. Although the transformer approach added computational cost, the benefits in accuracy, low false positives and resistance to adversarial attacks were the most successful. A combination of blockchain and watermark authentication also enhanced trust, creating a holistic system for detecting and verifying deep fakes in real-world settings.

This research yielded several significant observations on the efficacy of hybrid AI models for detecting and authenticating deepfakes. The findings verified that the hybrid methods were more accurate, achieved higher F1 Scores, and were more robust than the single-model baselines; this was in line with the latest findings indicating that spatial, temporal, and contextual modelling should be combined to develop reliable deepfake detection [11, 13]. In particular, the researchers found that the CNN+LSTM hybrid model achieved substantial gains over single CNN and LSTM models, confirming that temporal consistency plays a crucial role in detecting the modest manipulation artefacts that frame-based analysis alone cannot capture. Similarly, the transformer-based hybrid achieved the best detection accuracy and the lowest false-positive rate, which can be attributed to the growing body of literature indicating that transformer architectures, compared to CNNs, model global context more effectively [12]. The major strength of the transformer-based hybrid was its resistance to adversarial attacks. When perturbed with noise, compression, or adversarial optimisation, the model's accuracy dropped by less than 5%, whilst single-model detectors dropped by up to 15%. The result supported the claims in the literature that ensembles and transformers are less prone to surface-level cues and can thus extrapolate more effectively to the task of understanding unseen manipulations [2, 10]. Nevertheless, the researchers found trade-offs between computational efficiency and detection performance. Compared to transformer-based hybrid models, CNN-based models processed the video stream in real-time at up to 28 frames per second (fps). Such a restriction has practical consequences for real-time implementation, for example, in live video conferencing or social media monitoring, where latency is crucial. Therefore, transformers are more precise, but their incorporation into time-sensitive systems cannot be done easily without optimisation or parallelisation measures to balance computational costs. The outcomes of the authentication processes highlighted the importance of incorporating detection and verification. Semi-fragile watermarking and blockchain provenance were useful for verifying content authenticity, not only before detection analysis but also during it. Such results were consistent with recent suggestions about content provenance models, which emphasised that detection alone is insufficient due to the ongoing development of deepfake generation techniques [9, 15]. Through authentication and detection, the researchers demonstrated that they could minimise false positives and detect content manipulation without relying on AI-based classifiers. The practical implication is that authentication provides an added layer of protection in high-stakes decision-making environments where media integrity is crucial, such as journalism, law, and national security. Although these are the study's strengths, other limitations warrant discussion. To begin with, Celeb-DF V2 is not the most complex benchmark dataset. However, it is not yet a comprehensive representation of deep-fake variants in the real world, as manipulations can also involve multimodal combinations, lower resolutions, or newer generative models, such as diffusion models. Second, although the hybrid framework was more resistant to adversarial perturbations, it was not immune to all types of attacks, and it warrants further investigation into adversarial training or robust optimisation algorithms. Third, the blockchain-based authentication method introduced latency in the form of transaction verification time, and, unless thinly disguised options or private blockchain settings are used, would be unsuitable for real-time applications. The results also have broader implications. Technically, they emphasise hybrid models that integrate CNNs, RNNs, and transformers rather than a single architecture. In practice, they emphasise that both detection and authentication mechanisms, such as watermarking and provenance tracking, should be implemented to provide holistic protection against misinformation and fraud. Ethically and socially, the paper emphasised the urgency of exploring the so-called 'lie dividend,' as the mere presence of deepfakes undermines trust in genuine media [2]. The implementation of hybrid detection-authentication systems may reduce this threat; however, the adoption of this approach will ultimately depend on the popularity, legislation, and acceptance of technological protection by the population. To summarise, the discussion revealed that although hybrid AI models have made a significant contribution to the current state of deepfake detection and authentication, there are still issues to address regarding computational power, dataset diversity, and practical application. The combination of detection and blockchain/watermarking proved to be a promising step towards building end-to-end trustworthy systems, but further development is needed to balance performance with practical realities. Future studies are required to investigate lightweight transformer design, multimodal detection, and standardised authentication protocols to bridge the gap between laboratory success and scale.

CONCLUSIONS

In this paper, we discuss the detection and authentication of deepfakes using hybrid artificial intelligence (AI) models that comprise convolutional neural networks (CNNs), recurrent networks (LSTMs/GRUs), and transformers, along with blockchain and digital watermarking solutions. The study, using the Celeb-DF V2 dataset, showed that hybrid methodologies were much more effective than single-model baselines in terms of accuracy, F1-score, and robustness. The CNN+LSTM hybrid proved helpful for detecting temporal anomalies in video sequences. In contrast, the transformer-based hybrid achieved the best overall performance, with improved detection accuracy and greater resistance to adversarial manipulations. The researchers also integrated the authentication mechanisms to enhance the framework. Provenance tracking based on blockchain guarantees tamper-proof authentication of media provenance, whereas semi-fragile watermarking ensures proactive indication of manipulation before it is recognised. This combination addressed the growing limitations of detection-only systems, providing verifiable assurance of authenticity. These findings show that combined systems are better positioned to mitigate the threats of misinformation, fraud, and the erosion of social trust that deepfakes are likely to cause. However, the study also noted practical challenges. Transformer-based hybrids achieved high accuracy but consumed more computational resources and ran slower, raising concerns about real-time implementation. The researchers also faced limitations when using benchmark datasets like Celeb-DF V2, as the evaluation scope was constrained by the fact that real-world deepfakes may involve multimodal manipulations, lower input quality, or other novel generative methods. Similarly, blockchain authentication introduces latency, which developers may need to mitigate by using lightweight or private authentication mechanisms.

Nevertheless, the research outcomes demonstrated the relevance of hybrid models and integrated authentication in developing deepfake defence measures despite these challenges. The research contributed to the literature by highlighting that no single method of detection is sufficient and that multimodal methods are necessary to achieve strength and reliability. Future work should focus on creating lightweight, efficient hybrid architectures that prioritise accuracy and real-time feasibility. Multimodal detection systems that combine audio, video, and biometric signals should be further researched, along with adversarial training to enhance their resistance to emerging attacks. On the authentication front, the studies should focus on standardised protocols for watermarking and provenance tracking to promote interoperability and adoption across platforms. Lastly, it will be essential to build broader partnerships among researchers, policymakers, and industry players to advance technical solutions into real-world systems that can uphold the integrity of information against increasingly sophisticated synthetic media.

REFERENCES

1. Dehghani, A., & Saberi, H. (2025). Generating and Detecting Various Types of Fake Image and Audio Content: A Review of Modern Deep Learning Technologies and Tools. *arXiv preprint arXiv:2501.06227.*
2. Croitoru, F., Hiji, A., Hondru, V., Ristea, N. C., Irofti, P., Popescu, M., Rusu, C., Ionescu, R. T., Khan, F. S., & Shah, M. (2024). Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook. *arXiv (Cornell University).* doi: 10.48550/arxiv.2411.19537
3. Hashmi, A., Shahzad, S. A., Lin, C., Tsao, Y., & Wang, H. (2024). Understanding Audiovisual Deepfake Detection: Techniques, Challenges, Human Factors and Perceptual Insights. *arXiv (Cornell University).* doi: 10.48550/arxiv.2411.07650
4. Khanjani, Z., Watson, G., & Janeja, V. P. (2021). How deep are the fakes? Focusing on audio Deepfake: a survey. *arXiv (Cornell University).* doi: 10.48550/arxiv.2111.14203
5. Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C. Y., & Zhao, Y. (2023). Audio Deepfake Detection: A Survey. *Journal of Latex Class Files, 14*(8)
6. Lee, H., Lee, C., Farhat, K., Qiu, L., Geluso, S., Kim, A., & Etzioni, O. (2024). The Tug-of-War between deepfake generation and detection. *arXiv (Cornell University).* doi: 10.48550/arxiv.2407.06174
7. Almars, A. M. (2021). DeepFakes Detection Techniques Using Deep Learning: A Survey. *Journal of Computer and Communications, 09*(05), 20–35. doi: 10.4236/jcc.2021.95003
8. Singh, L. H., Charanarur, P., & Chaudhary, N. K. (2025). Advancements in Detecting Deepfakes: AI Algorithms And Future Prospects – A Review. *Discover Internet of Things, 5*(1). doi: 10.1007/s43926-025-00154-0
9. Neekhara, P., Hussain, S., Zhang, X., Huang, K., McAuley, J., & Koushanfar, F. (2022). FaceSigns: Semi-Fragile neural watermarks for media authentication and countering deepfakes. *arXiv (Cornell University).* doi: 10.48550/arxiv.2204.01960
10. Thing, V. L. L. (2023). Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers. *arXiv (Cornell University).* doi: 10.48550/arxiv.2304.03698
11. Saikia, P., Dholaria, D., Yadav, P., Patel, V., & Roy, M. (2022). A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features. *arXiv (Cornell University).* doi: 10.48550/arxiv.2208.00788
12. Xi, A. J., & Chen, E. (2025). Classifying deepfakes using SWin transformers. *arXiv (Cornell University).* doi: 10.48550/arxiv.2501.15656
13. Soudy, A. H., Sayed, O., Tag-Elser, H., Ragab, R., Mohsen, S., Mostafa, T., Abohany, A. A., & Slim, S. O. (2024). Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications, 36*(31), 19759–19775. doi: 10.1007/s00521-024-10181-7
14. Khan, S. A., Artusi, A., & Dai, H. (2021). Adversarially robust deepfake media detection using fused convolutional neural network predictions. *arXiv (Cornell University).* doi: 10.48550/arxiv.2102.05950
15. Mastoi, Q., Memon, M. F., Jan, S., Jamil, A., Faique, M., Ali, Z., Lakhan, A., & Syed, T. A. (2025). Enhancing deepfake content detection through blockchain technology. *International Journal of Advanced Computer Science and Applications, 16*(6). doi: 10.14569/ijacsa.2025.0160607
16. ITU. (2024). Detecting deepfakes and generative AI: Report on standards for AI watermarking and multimedia authenticity workshop. Retrieved from https://www.itu.int/hub/publication/t-ai4g-ai4good-2024-7/