LITERATURE SURVEY

ON

Deepfake Detection in Online Social Network (OSN)

By

Name: Tarunikka Suresh Unnikrishnan

Faculty in Charge: Dr Raja Muthalagu

Bits Pilani, Dubai Campus

Dubai International Academic City, Dubai

Abstract

Digital trust and online safety are in jeopardy by the spread of deepfake media, enabled by generative adversarial networks (GAN's) and diffusion models. The majority of deepfake detection systems currently in use solely employ machine learning models, such as transformers or Convolutional Neural Networks, to analyze the audio or video content. This project offers a framework for context-aware detection that integrates two viewpoints: the content and its online dissemination. At the content level, transformer-based models are used to analyze both video frames and audio signals for subtle inconsistencies. At the context level, graph neural networks examine how the media is shared on social networks, including repost patterns, speed of spread, and user credibility. By combining these two types of information, the system is expected to be more reliable and resistant to manipulation in real-world scenarios. To guarantee practical applicability, the framework will be assessed using social media data as well as benchmark datasets.

Keywords : Deepfake detection, Transformer Models ,Graph Neural Networks (GNNs), Multimodal Analysis ,Online Social Networks (OSNs)

Introduction

The rapid advancement of artificial intelligence has revolutionized how digital media is produced, shared, and consumed. Among the most striking developments is the rise of *deepfake* which are synthetic audio, video, or images generated using deep learning techniques such as generative adversarial networks and diffusion models. These technologies can create highly convincing fabricated content that is nearly indistinguishable from authentic media. While deepfakes have legitimate applications in entertainment, education, and accessibility, their malicious use has raised significant concerns for society. From political disinformation and identity fraud to reputational harm and financial scams, manipulated media undermines trust in online communication and challenges the integrity of digital ecosystems.

To counter these threats, researchers have developed a wide range of deepfake detection methods. Early approaches focused on visual artifacts such as abnormal eye blinking, texture irregularities, or inconsistencies in facial landmarks. With progress in generative models, however, these cues have become less reliable, prompting the adoption of convolutional neural networks and transformer-based architectures capable of learning subtle, data-driven patterns. Parallel efforts have extended detection beyond video to include audio deepfakes, exploring spectral and temporal inconsistencies in synthesized speech. More recently, multimodal detection techniques have combined audio and visual streams, showing that inconsistencies between lip movement and speech can serve as powerful forensic signals.

Despite these advancements, online social networks (OSNs) remain fertile ground for the spread of deepfakes. These platforms amplify the risks by enabling rapid, large-scale dissemination while introducing distortions such as compression, re-encoding, and format changes that degrade detection accuracy. At the same time, OSNs provide contextual information such as repost patterns, diffusion structures, and user credibility that has been underexplored in deepfake forensics. Leveraging both multimedia signals and social propagation dynamics opens a promising direction for improving robustness in real-world settings.

This project builds on these insights by proposing a context-aware deepfake detection framework that unites content-based analysis with social-context features. By employing transformer architectures for audio–visual feature extraction and graph neural networks for modeling OSN diffusion, the framework aims to deliver more resilient and adaptable detection. Evaluating the approach on state-of-the-art benchmarks and datasets will ensure not only technical accuracy but also practical applicability for safeguarding online information integrity.

Literature Survey

Deepfakes have become more challenging to differentiate from real media due to the quick development of generative models like GANs, autoencoders, and diffusion techniques. As a result, deepfake detection has become a crucial area of study, attracting attention from the domains of multimedia forensics, computer vision, and natural language processing. Even though unimodal and multimodal detection have advanced significantly in the existing literature, there are still persistent issues, especially with regards to dataset generalization, resistance to hostile attacks, and incorporating contextual data from online social networks (OSNs).

A comprehensive review of detection techniques is given by Heidari et al. [1], who classified them into three groups: transformer-based, recurrent, and convolutional. According to their research, RNNs capture temporal dependencies but are not scalable, whereas CNNs are excellent at identifying low-level pixel inconsistencies but are fragile when compressed. Additionally, the review emphasizes that hybrid architectures are more successful, even though the majority of these models continue to overfit to particular datasets. Gong and Li [2] reinforce this by cataloging datasets and algorithms, showing that detectors trained on curated benchmarks fail dramatically on real-world content. The significance of taking social dynamics into account which is an aspect that is largely missing from content-only detection pipelines is also emphasized by their analysis. With case studies showing how deepfakes have been used in fraud and disinformation campaigns, Patel et al. [3] broaden this viewpoint and call for detection systems to be developed with realistic adversarial use cases in mind. This is further supported by Sunil et al. [4], who assess autonomous detection techniques in image, video, and audio modalities. They conclude that although deep learning has enhanced detection performance, most systems are not resilient to distortions like reposting and re-encoding, which are common on OSNs.

Recent research has placed a strong emphasis on the value of benchmarking. One of the most varied benchmarks available is DF40, a next-generation dataset that was introduced by Yan et al. [5] and includes 40 forgery types. Their tests show that models developed on more recent diffusion-based manipulations, such as FaceForensics++ and DFDC, are unable to identify them. This illustrates the need for cross-dataset resilience and the vulnerability of dataset-specific training. By carefully investigating the shortcomings of the existing benchmarks, Kaur et al. [6] reaffirm these worries by highlighting problems like an uneven distribution of classes, poor annotations, and a narrow range of manipulations. Together, these pieces show that although benchmarks are useful, they frequently inflate model performance in controlled settings, giving the impression of security, while real-world deepfakes continue to be far more complex.

In terms of architectural innovations, Bansal et al. [7] designed DFN, a lightweight hybrid that integrates MobNet and XGBoost to reduce computational complexity while maintaining accuracy. This demonstrates a move toward useful detection that can function on devices with constrained resources. In a comparative analysis of traditional CNN architectures, Ashani et al. [8] discovered that VGG19 performed better than the others in image-based tasks. Although these comparative analyses are useful, they frequently fall short of providing insights into context-aware or multimodal detection. In order to increase accuracy and safeguard privacy in collaborative detection scenarios, Heidari et al. [9] presented a blockchain-based federated learning model that combines SegCaps and CNNs. Although it is still primarily content-focused and does not make use of contextual metadata, this direction demonstrates the growing interest in privacy preserving frameworks.

Multimodal approaches are increasingly favored for their ability to exploit cross-domain inconsistencies. Astrid et al. [10] introduced an audio-visual detection framework that detects minute discrepancies between speech and lip movement. Their results enhance resilience against dataset bias and show that AV inconsistencies are powerful markers of manipulation. This was further developed by Khan et al. [11] with LGDF-Net, which combines global semantic features with local fine-grained features. Despite their effectiveness, their findings demonstrate that when models are exposed to unconstrained datasets, generalization is still restricted. The transition to deep learning-based multimodal fusion is further supported by Kumar et al.'s [12] survey of detection techniques, which highlights the obsolescence of manual methods like blink detection.

Other works investigate ensemble and boosting approaches. Bansal et al. [13] presented Faceswap Finder, a fusion-based method tailored for face-swap manipulations. AdaBoost was applied by Devi and Rajasekaran [14] to improve the classification of challenging samples in DFDC; however, their method was still vulnerable to high-quality manipulations created using more recent methods. In their comparison of CNN architectures such as Xception, ResNet, and VGG, Abbasi et al. [15] discovered that Xception was the most accurate, but it was also the most susceptible to adversarial perturbations. Together, these studies show that although ensemble approaches can slightly increase accuracy, they are unable to address the more fundamental problems of adversarial robustness and dataset dependency.

The rise of Vision Transformers (ViTs) has shifted the landscape. Chepchirchir and Mboli [16] compared CNNs, RNNs, and ViTs, reporting that ViTs generalized better across datasets, achieving accuracies exceeding 84%. According to their findings, transformers might be more resilient to manipulation artifacts because of their capacity to capture global dependencies. Sharma and Rawat [17] explored expression dynamics with their MADDM model, a masked autoencoder that leverages subtle facial expressions as cues. This method proved to be very successful, despite being computationally demanding,

because deepfake generators have trouble simulating real micro-expressions. In addition to these, Tolosana et al. [18] offer a more comprehensive review that emphasizes the transition from single-modality artifact detection to multimodal and context-aware detection.

When comparing across these works, several themes emerge. First, a common flaw is still dataset dependence. When applied to unconstrained datasets, even advanced models like LGDF-Net [11] and ViTs [16] exhibit discernible drops, demonstrating that benchmark-driven advancement is not equivalent to practical robustness. Second, there is not enough emphasis placed on adversarial robustness. Few models are specifically made to fend off adversarial perturbed samples, as Abbasi et al. [15] showed. Third, multimodal approaches like those proposed by Sharma and Rawat [17] and Astrid et al. [10] enhance resilience, but they only examine the manipulated media without considering contextual cues from the way content circulates on OSNs. Lastly, while privacy-preserving frameworks like federated learning [9] and lightweight models like DFN [7] increase deployment feasibility, they do not address the contextual spread of false information.

When combined, these studies show that although research on deepfake detection has advanced significantly, it still primarily focuses on content-level characteristics. The need for novel strategies is highlighted by the persistent gaps across datasets, adversarial robustness, and contextual modeling. To address these, the current project suggests a context-aware framework that combines graph neural networks trained on repost cascades, temporal diffusion, and user credibility with transformer-based models for audio and video. This framework attempts to attain both technical accuracy and practical dependability in online ecosystems by fusing social-contextual dynamics with content-level cues.

Proposed Work / Problem Definition

Although significant progress has been made in developing deepfake detection models, the majority of existing approaches remain narrowly focused on content-level artifacts present in visual or auditory data. While transformers, multimodal fusion systems, and convolutional neural networks have all shown promising performance on benchmark datasets, their performance significantly deteriorates in real-world situations where videos are compressed, re-encoded, or reposted repeatedly. Furthermore, the efficacy of solely media-centered solutions is still being undermined by adversarial attacks and generative refinements. However, the social context in which deepfakes spread has received comparatively little attention, especially in online social networks (OSNs), where manipulation spreads quickly and frequently display structural patterns that diverge from the diffusion of real media.

By combining content-level feature analysis and propagation-level contextual modeling, the proposed work seeks to close these gaps by presenting a context-aware detection framework. To detect subtle inconsistencies that might evade CNN-based detectors, transformer-based architectures will be employed at the media level to extract spectral cues from audio signals and spatiotemporal features from video frames. Graph neural networks (GNNs) will be used to model temporal dynamics, user credibility, and repost cascades in OSNs at the contextual level. The system is anticipated to achieve increased robustness against adversarial manipulations and greater reliability across heterogeneous datasets by combining these two complementary viewpoints.

Formally, the problem can be defined as follows: given an input sample M consisting of a video–audio pair and an associated propagation graph G derived from OSN sharing activity, the objective is to learn a mapping function $f(M,G)\rightarrow\{0,1\}$ where 0 denotes authentic content and 1 denotes manipulated content. Unlike traditional detection models that rely solely on M, this formulation explicitly incorporates G as an additional source of discriminative information. The central hypothesis of this work is that combining media-level inconsistencies with social-contextual diffusion patterns will provide more reliable detection of deepfakes in real-world digital ecosystems

Methodology

To improve deepfake detection in actual online social network settings, the suggested framework integrates social-context and multimedia features. There are six stages to the methodology: data gathering, preprocessing, analysis at the content and context levels, classification and fusion, and assessment.

DATA COLLECTION

To guarantee balanced training and evaluation, the dataset will include both real-world social media samples and well-known benchmarks. A foundation of various manipulation types, such as face-swapping, lip-syncing, and diffusion-based synthesis, will be provided by standard corpora like FaceForensics++, Celeb-DFv2, DFDC, and DF40. These datasets offer baseline comparisons against current techniques and enable controlled experimentation. Additional samples will be collected from online social networks to capture the complexity of real-world situations. To make sure the framework learns to function accurately outside of lab settings, these OSN-based examples include repost chains, platform-induced compression, and user interactions that mimic how manipulated content spreads in actual environments.

PREPROCESSING

Multimedia data will be preprocessed to mimic the distortions created during online sharing before features are extracted. In addition to being compressed, re-encoded, and subjected to other degradations typical of OSNs, video streams will be normalized in terms of resolution and frame rate. Similarly, to replicate the quality losses that occur during uploads, audio streams will be re-encoded at different bitrates and then transformed into spectrograms. In addition to avoiding overfitting to perfect benchmark conditions, this step guarantees consistency across datasets. By introducing controlled distortions during preprocessing, the system becomes better prepared to handle noisy, degraded, and heterogeneous data encountered in practice.

CONTENT-LEVEL ANALYSIS

The content-level module uses advanced transformer-based backbones to analyze both visual and auditory features. Vision Transformers (ViT) or Swin Transformers process video streams by modeling long-range spatial–temporal dependencies. This allows them to identify subtle inconsistencies like frame-level artifacts, mismatched textures, or unnatural facial expressions. To capture pitch irregularities, speech mismatches, or lip movement misalignments, audio transformers or conformer networks process audio streams in parallel to turn them into spectrograms. Multimodal fusion layers are then used

to align and combine the outputs from both modalities, improving the system's capacity to identify subtle audiovisual irregularities that are frequently missed by single-modality detectors.

CONTEXT-LEVEL ANALYSIS

Beyond content, the framework incorporates social-contextual information by modeling how media spreads across online social networks. Repost data, interaction logs, and user credibility metrics are used to build diffusion graphs where nodes represent accounts and edges represent repost or sharing actions. Temporal features such as repost burstiness, depth of diffusion cascades, and propagation speed are encoded alongside user-level credibility indicators. The application of Graph Neural Networks (GNNs) to these structures allows the system to discern between manipulated campaigns and natural sharing behavior by capturing both local and global diffusion patterns. An extra line of defense is offered by this context-aware method, especially against adversarial deepfakes, which look realistic from the outside but spread erratically through social networks.

FUSION AND CLASSIFICATION

The fusion stage integrates the outputs of the multimedia and context-level modules to form a unified detection decision. A late-fusion strategy is employed, where embeddings from the visual, audio, and graph modules are concatenated and passed through fully connected layers for joint optimization. This fused representation captures both media-level cues and contextual irregularities, allowing the classifier to rely on complementary information sources. The classification head then outputs authentic or manipulated which is a binary label while confidence scores provide interpretability and allow threshold tuning for deployment in different application scenarios.

EVALUATION

To guarantee balanced performance, the framework will be carefully evaluated on social network samples as well as benchmark datasets. To gauge efficacy under various circumstances, metrics like accuracy, precision, recall, F1-score, and area under the ROC curve will be presented. Robustness tests will be carried out under distortions such as recompression, re-encoding, and cross-platform reposting, in addition to standard benchmarks. The effectiveness of integrating multimedia and context features will be evaluated through comparative experiments against the most advanced models. This multi-level assessment guarantees that the system is both technically sound and practically feasible for implementation in actual online environments.

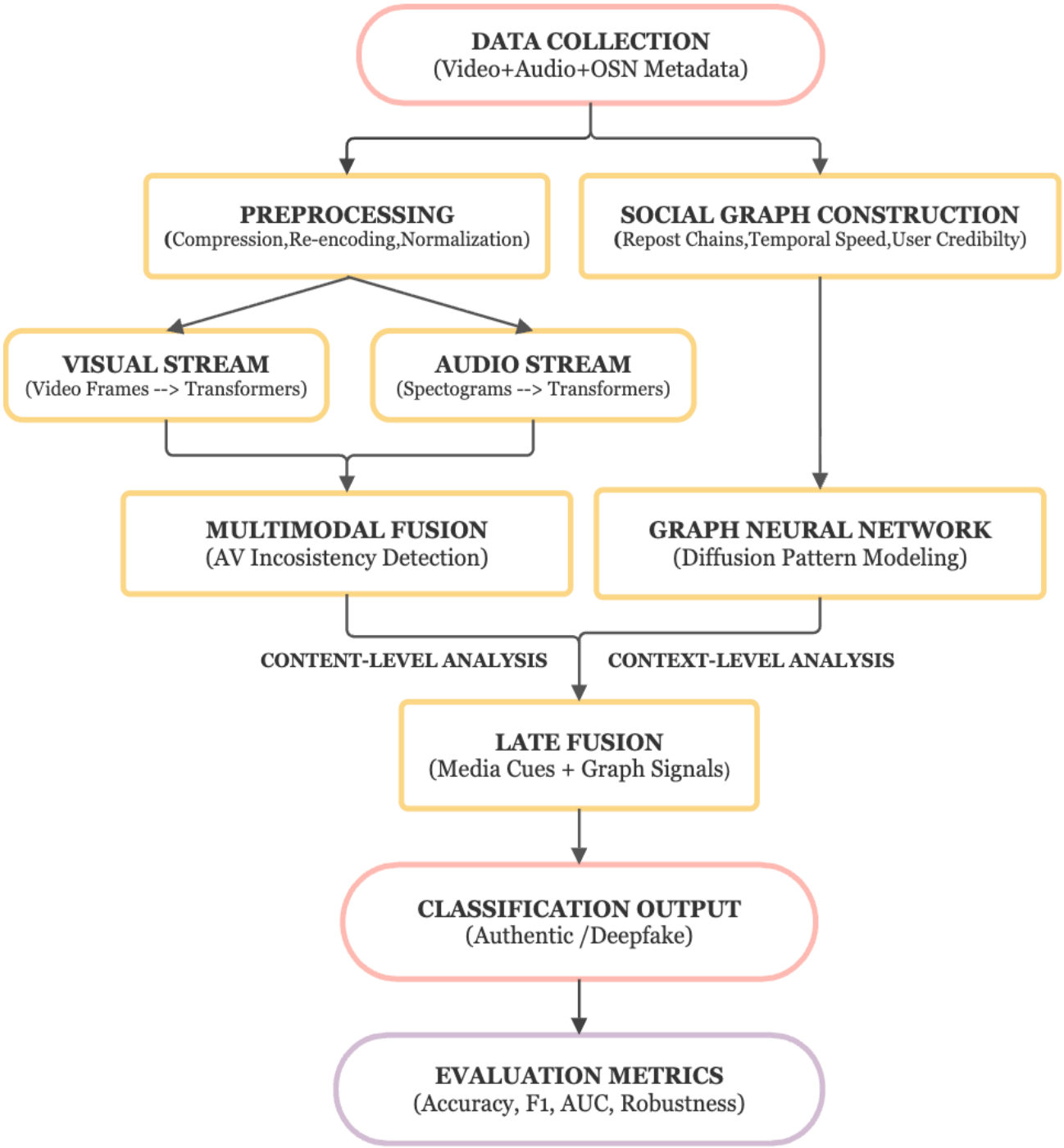

Fig 1: The proposed Methodology Diagram

References

[1] A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, “Deepfake detection using deep learning methods: A systematic and comprehensive review,” *WIREs Data Mining and Knowledge Discovery*, vol. 14, no. 2, p. e1520. [Online]. Available: https://doi.org/10.1002/widm.1520

[2] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, C. Wang, S. Ding, Y. Wu, and L. Yuan, “DF40: Toward next-generation deepfake detection,” in *Proc. NeurIPS Datasets and Benchmarks Track*, 2024. [Online]. Available: https://github.com/YZY-stack/DF40

[3] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, “Deepfake video detection: Challenges and opportunities,” *Artificial Intelligence Review*, vol. 57, pp. 159–184. [Online]. Available: https://doi.org/10.1007/s10462-024-10810-6

[4] Y. Patel, S. Tanwar, R. Gupta, P. Bhattacharya, I. E. Davidson, R. Nyameko, S. Aluvala, and V. Vimal, “Deepfake generation and detection: Case study and challenges,” *IEEE Access*, vol. 11, pp. 143296–143315. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3342107

[5] R. Sunil, P. Mer, A. Diwan, R. Mahadeva, and A. Sharma, “Exploring autonomous methods for deepfake detection: A detailed survey on techniques and evaluation,” *Heliyon*, vol. 11, p. e42273. [Online]. Available: https://doi.org/10.1016/j.heliyon.2025.e42273

[6] N. Bansal, T. Aljrees, D. P. Yadav, K. U. Singh, A. Kumar, G. K. Verma, and T. Singh, “Real-time advanced computational intelligence for deep fake video detection,” *Applied Sciences*, vol. 13, no. 3, p. 3095. [Online]. Available: https://doi.org/10.3390/app13053095

[7] Z. N. Ashani, I. S. C. Ilias, K. Y. Ng, M. R. K. Ariffin, A. D. Jarno, and N. Z. Zamri, “Comparative analysis of deepfake image detection method using VGG16, VGG19 and ResNet50,” *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 47, no. 1, pp. 16–28. [Online]. Available: https://doi.org/10.37934/araset.47.1.1628

[8] A. Heidari, N. J. Navimipour, H. Dag, S. Talebi, and M. Unal, “A novel blockchain-based deepfake detection method using federated and deep learning models,” *Cognitive Computation*, vol. 16, pp. 1073–1091. [Online]. Available: https://doi.org/10.1007/s12559-024-10255-7

[9] M. Astrid, E. Ghorbel, and D. Aouada, “Detecting audio-visual deepfakes with fine-grained inconsistencies,” in *Proc. British Machine Vision Conference (BMVC)*, 2024. [Online]. Available: https://bmva-archive.org.uk/bmvc/2024/papers/Paper_695/paper.pdf

[10] H. Khan, Z. Liu, Y. Li, and L. Song, “LGDF-Net: Local and global feature-based dual-branch fusion networks for deepfake detection,” *Applied Intelligence*, 2024. [Online]. Available: https://doi.org/10.1007/s10489-024-05170-4

[11] S. Kumar, A. Gupta, R. Mishra, and T. Singh, “Deepfake detection techniques: A comparative survey,” *SSRN Preprint*. [Online]. Available: https://doi.org/10.2139/ssrn.5240416

[12] V. Bansal, R. Kumawat, and S. Anand, “Faceswap Finder: A fusion-based deepfake detection technique,” *Procedia Computer Science*, vol. 235, pp. 231–238. [Online]. Available: https://doi.org/10.1016/j.procs.2025.01.123

[13] K. Devi and S. Rajasekaran, “Deepfake video detection using AdaBoost on DFDC dataset,” *Procedia Computer Science*, vol. 235, pp. 210–218. [Online]. Available: https://doi.org/10.1016/j.procs.2025.01.117

[14] H. Abbasi, F. Riaz, and A. A. Shah, “Comparative evaluation of deepfake detection using CNN architectures: Xception, ResNet, and VGG,” *Procedia Computer Science*, vol. 235, pp. 198–209. [Online]. Available: https://doi.org/10.1016/j.procs.2025.01.115

[15] B. Chepchirchir and K. Mboli, “Comparative analysis of CNN, RNN, and Vision Transformers for deepfake detection,” *Procedia Computer Science*, vol. 235, pp. 239–247. [Online]. Available: https://doi.org/10.1016/j.procs.2025.01.124

[16] P. Sharma and R. Rawat, “Enhancing deepfake detection through dynamics of facial expressions using masked autoencoders,” *Applied Sciences*, vol. 15, no. 3, p. 1225. [Online]. Available: https://doi.org/10.3390/app15031225

[17] R. Tolosana, C. Rathgeb, A. Morales, and C. Busch, “Deepfake generation and detection: State of the art, open challenges, and future directions,” *Information Fusion*, vol. 99, p. 101861. [Online]. Available: https://doi.org/10.1016/j.inffus.2023.101861

[18] Y. Zhang, J. Liu, and M. Chen, “Micro-expression dynamics for deepfake detection: A novel approach,” *Neural Computing and Applications*, 2025. [Online]. Available: https://doi.org/10.1007/s00521-025-09876-9