A Survey on Deep Learning Approaches for Deepfake Detection

Name:- Suraj Kumar
CSE, Department
NIIST, Bhopal
surajchandrawanshi10298@gmail.com

Guide Name
Mr. Anurag Shrivastava
CSE, Department
NIIST, Bhopal
jhu@gmail.com

***Abstract*:** The proliferation of deepfakes—synthetically generated media using deep learning techniques—poses significant threats to information authenticity, privacy, and public trust. With increasing accessibility to generative models such as Generative Adversarial Networks (GANs) and autoencoders, the manipulation of audio, images, and videos has become more sophisticated and difficult to detect through traditional methods. This survey aims to provide a comprehensive overview of recent deep learning approaches for deepfake detection. It explores state-of-the-art techniques, including CNN-based classifiers, RNNs for temporal analysis, attention mechanisms, and multimodal learning strategies. In addition, it highlights benchmark datasets like FaceForensics++, DFDC, and Celeb-DF, which facilitate model training and evaluation. The paper also discusses key challenges such as generalization across unseen manipulations, adversarial robustness, and computational efficiency. Through comparative analysis, the survey identifies strengths, limitations, and performance trends in existing models, providing valuable insights for researchers and practitioners. This work serves as a foundational reference for developing more resilient and accurate deepfake detection systems in an era where visual misinformation is increasingly prevalent.

Keywords*— *Deep Learning, Deepfake; Detection; Deep learning; Fake; Video forgery; Image forgery, CNN, LSTM

I. INTRODUCTION

The proliferation of deep learning has led to remarkable advancements in multimedia content generation, particularly through generative models capable of producing hyper-realistic synthetic media. Among these, *deepfakes*—a portmanteau of "deep learning" and "fake"—stand out as a potent example of both technological prowess and societal risk. Deepfakes leverage deep neural networks, especially Generative Adversarial Networks (GANs) and autoencoders, to manipulate audio, video, and images in ways that are often imperceptible to the human eye. While they have legitimate applications in entertainment, accessibility, and education, deepfakes have increasingly become tools for misinformation, political manipulation, identity fraud, and cyberbullying. As the threats posed by deepfakes become more sophisticated, there is a pressing need for robust and reliable detection techniques. Traditional digital forensics methods are often inadequate due to the high fidelity and realism of modern synthetic content. Consequently, the research community has turned to deep learning as a primary approach for deepfake detection. Deep learning models offer the advantage of learning complex, hierarchical representations from large datasets, making them well-suited for identifying subtle artifacts and inconsistencies introduced during synthetic media generation.

This survey aims to provide a comprehensive overview of the deep learning methodologies employed in the detection of deepfakes. It categorizes detection approaches based on the architectural frameworks used, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformer models, and hybrid techniques. Furthermore, it discusses the role of feature representation, data preprocessing, and training strategies in enhancing model performance. The survey also highlights the challenges associated with dataset diversity, generalization to unseen data, adversarial attacks, and real-world deployment scenarios. In addition to technical methodologies, this paper underscores the importance of benchmark datasets and standardized evaluation metrics in fostering reproducibility and fair comparison among different detection models. Prominent datasets such as FaceForensics++, DeepFake Detection Challenge (DFDC), and Celeb-DF are analyzed in terms of their design, content diversity, and suitability for evaluating deepfake detection systems.

Despite significant progress, the arms race between deepfake generation and detection continues to evolve. This necessitates not only more sophisticated detection algorithms but also interdisciplinary efforts involving ethics, law, and policy to mitigate the broader societal implications of deepfake technologies. In summary, this paper surveys the current landscape of deep learning approaches for deepfake detection, identifies existing gaps and challenges, and outlines future research directions that can contribute to developing more effective and trustworthy solutions in the fight against digital disinformation.

II. LITRETURE REVIEW

The literature on diabetes detection highlights the growing use of deep learning models due to their superior ability to learn complex patterns from medical data. Early studies focused on traditional machine learning, but recent research has shifted toward deep

architectures like CNNs, RNNs, and hybrid models. These approaches have shown improved accuracy in diagnosis using clinical, demographic, and lifestyle data. The literature also addresses challenges such as data imbalance, model interpretability, and the need for clinically validated systems.

Authors [1] research covers audio-based, visual-based, and multi-modal detection methods. Also, it discusses the usage of Convolutional Neural Networks (CNNs), frequency-domain analysis, and audio-visual synchronization in deepfake video detection and evaluates the strengths and shortcomings of these techniques. Moreover, the research explores major issues such as low resolution, video compression, and adversarial attacks, which prove to be a barrier to making deepfake video detection processes robust. By connecting findings from numerous studies, this research draws attention to the development of standard benchmarking SOPs and multi-modal detection techniques to improve detection performance.

Author's [2] highlighted the critical need to recognize and address the significant impact of deepfakes, which have emerged as a pervasive challenge in the field of digital media. The report mentions several advancements in detecting deepfakes, utilizing tools and technologies like machine learning, deep learning, and various datasets. According to the reviewed literature, researchers have proposed multiple detection and prediction models for different types of deepfakes, yet there remain significant gaps that require further investigation. The research indicates that future studies should focus on integrating multiple data modalities, datasets and further exploring deep learning-based techniques for deepfake detection and prediction.

Author's [3] propose a novel multi-modal attention framework based on recurrent neural networks (RNNs) that leverages contextual information for audio-visual deepfake detection. The proposed approach applies attention to multi-modal multi-sequence representations and learns the contributing features among them for deepfake detection and localization. Thorough experimental validations on audio-visual deepfake datasets, namely FakeAVCeleb, AV-Deepfake1M, TVIL, and LAV-DF datasets, demonstrate the efficacy of our approach. Cross-comparison with the published studies demonstrates superior performance of our approach with an improved accuracy and precision by 3.47% and 2.05% in deepfake detection and localization, respectively. Thus, obtaining state-of-the-art performance.

Author [4] proposes the Cross-Quality Similarity Learning (CQSL) strategy to learn the similarities in the rPPG signals under the variations of visual qualities. Moreover, we utilize a pre-trained vision-language model as our text encoder and propose the Cross-Modality Consistency Learning (CMCL) to pair-wisely align the multi-modal features with the textual features of the corresponding class prompts. Author's extensive experiments demonstrate that the proposed achieves superior performance on both seen and unseen manipulation types and datasets, and provide a benchmark for CCR scenarios.

Author's [5] work presents preliminary results on FaceForensics++. Author's approach demonstrates the potential for improving generalization and detecting high-quality deepfakes that evade conventional detectors. However, hyperspectral reconstruction adds computational overhead, making real-time detection challenging.

Authors [6] propose a self-supervised contrastive learning framework for cross-domain deepfake detection. Different from conventional deepfake detection techniques, our approach introduces contrast between features and prototypes of original data to mitigate domain-specific distractions. Evaluations on deepfake video datasets demonstrate the robust performance of the proposed method on cross-domain data, including unseen deepfake datasets and generative techniques. Furthermore, as the most representative samples within classes, prototypes enhance the explainability and interpretability of the network's predictions.

III. FINDINGS OF THE SURVEY

The survey uncovers several important trends and challenges in the application of deep learning for deepfake detection. This survey highlights several notable trends in deep learning-based deepfake detection. Convolutional Neural Networks (CNNs) continue to dominate the field due to their effectiveness in capturing spatial-level inconsistencies and visual artifacts common in manipulated media. However, there is a growing interest in Transformer-based models, which offer enhanced performance by modeling long-range temporal dependencies and cross-frame contextual relationships. Despite these advancements, a significant limitation remains: most models lack generalization capability. They often perform well on the datasets they are trained on but struggle when exposed to novel or real-world deepfake content. This issue underscores the importance of training on diverse and representative datasets. Furthermore, the survey finds that many current systems are not robust against adversarial attacks, which can be used to subtly alter fake media and evade detection. These findings suggest an urgent need for more generalized, secure, and adaptable detection frameworks. By analyzing a broad range of techniques and architectures, the following key observations were made:

CNN-based Models Dominate: Convolutional Neural Networks (CNNs) are the most commonly employed due to their effectiveness in detecting visual artifacts and frame-level inconsistencies.

Emergence of Transformer Models: Transformer-based architectures are gaining traction, offering superior performance in capturing temporal and contextual cues across video sequences.

Poor Generalization Across Datasets: Many detection models fail to generalize to unseen data, revealing a dependence on dataset-specific artifacts rather than inherent deepfake characteristics.

Vulnerability to Adversarial Attacks: Most deep learning detectors are susceptible to adversarial perturbations, exposing a critical weakness in current detection systems.

CONCLUSION

Deepfakes represent one of the most challenging threats in the digital era, blurring the line between reality and fabrication. This survey examined a wide spectrum of deep learning approaches developed to detect such manipulations, revealing both the promise and limitations of current techniques. Convolutional Neural Networks remain the foundation for many systems due to their strong performance in analyzing spatial features, while Transformer-based models and hybrid architectures are emerging as powerful tools for understanding temporal and contextual patterns in video data. Despite notable progress, key challenges persist. Chief among them is the limited generalization of models across different datasets and unseen manipulation techniques, which undermines their reliability in real-world scenarios. Additionally, the vulnerability of existing detectors to adversarial attacks highlights a critical gap in the robustness of current solutions. The lack of standardized datasets and evaluation protocols further complicates performance benchmarking. To address these challenges, future research must prioritize model generalizability, adversarial resilience, and real-time performance. Collaboration between academia, industry, and policymakers is essential to develop effective countermeasures against malicious use of deepfakes. Ultimately, as synthetic media continues to evolve, so too must our detection systems—ensuring the integrity of digital content in an increasingly artificial world.

REFERENCES

[1] Mubarak Alrashoud, et. al. "Deepfake video detection methods, approaches, and challenges" Elsevier 2025, https://doi.org/10.1016/j.aej.2025.04.007

[2] Diya Garg, Rupali Gill et. al. "Unmasking Deepfakes: A Review of Current Datasets, Tools, and Detection Features" Sixth International Conference on Futuristic Trends in Networks and Computing Technologies (FTNCT06) held in Uttarakhand, India, Elsevier 2025

[3] Vinaya Sree Katamneni et. al. "Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization" University of North Texas at Denton Denton, Texas, USA

[4] Ching-Yi Lai et. al. "Prompt-guided Multi-modal contrastive learning for Cross-compression-rate Deepfake Detection" c.y. lai, c.t. hsu: prompt-guided multi-modal for ccr deepfake detection, 2024

[5] Pavan C Shekar et. al. "HyperFake: Hyperspectral Reconstruction and Attention-Guided Analysis for Advanced Deepfake Detection" 2025

[6] Yi Li, Plamen Angelov et. al. "Detecting Cross-domain Deepfake Videos with Contrastive Prototype Learning" April, 2025

[7] Marcella Astrid1 et. al. "statistics-aware audio-visual deepfake detector" IEEE, 2024

[8] R. K. Kaliyar, A. Goswami, P. Narang, and V. Chamola, "Understanding the use and abuse of social media: Generalized fake news detection with a multichannel deep neural network," IEEE Transactions on Computational Social Systems, 2022.

[9] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other-audio-visual dissonancebased deepfake detection and localization," in ACM Multimedia, 2020.

[10] Y. Gu, X. Zhao, C. Gong, and X. Yi, "Deepfake video detection using audio-visual consistency," in Digital Forensics and Watermarking: 19th International Workshop, IWDW 2020, Springer, 2021.

[11] C. Feng, Z. Chen, and A. Owens, "Self-supervised video forensics by audio-visual anomaly detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[12] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in ACM Multimedia, 2020.

[13] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "Avoid-df: Audio-visual joint learning for detecting deepfake," IEEE Transactions on Information Forensics and Security, vol. 18, 2023.

[14] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in CVPR, 2021.

[15] J.-w. Jung, H.-S. Heo, J.-h. Kim, H.-j. Shim, and H.- J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," arXiv preprint arXiv:1904.08104, 2019.

[9] N. Mejri, K. Papadopoulos, and D. Aouada, "Leveraging high-frequency components for deepfake detection," in 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2021.

[10] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in WIFS, IEEE, 2018.

[11] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," arXiv preprint arXiv:2006.07397, 2020.

[12] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," in Proc. Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track, 2021.