A Review on Explainable AI for Deepfake Detection Leveraging Hybrid Deep Learning Techniques

Hitarth H. Raval[1], Mehul S. Patel[2], Shweta D. Parmar[3]

M.Tech Student, Dept. of CE, Sankalchand Patel College of Engineering, Sankalchand Patel University,Visnagar, India[1]
Assistant Professor, Dept. of IT, Sankalchand Patel College of Engineering, Sankalchand Patel University, Visnagar, India[2]
Assistant Professor, Dept. of CE, Sankalchand Patel College of Engineering, Sankalchand Patel University,Visnagar, India[3]

hitarthraval.hr@gmail.com[1], mspatelit_spce@spu.ac.in[2], sdparmarce_spce@spu.ac.in[3]

Abstract: The advent of deepfake technology, leveraging advancements in generative artificial intelligence, has catalyzed a substantial threat to the integrity and trustworthiness of digital media. Deepfakes, which include hyper-realistic synthetic images, videos, and audio generated using techniques such as Generative Adversarial Networks (GANs), have been widely exploited to create fake content that is increasingly indistinguishable from reality. This work investigates the intersection of Explainable Artificial Intelligence (XAI) with deepfake detection, emphasizing the importance of transparency and interpretability in this field. We provide a detailed analysis of existing deepfake detection strategies, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid and multimodal approaches. The paper further emphasizes the importance of integrating XAI techniques to enhance model interpretability, reliability, and robustness, thus enabling more transparent and ethical AI systems. In addition, we assess various evaluation metrics and benchmark datasets utilized in deepfake detection research and discuss the limitations of current models. Finally, the paper outlines future research directions, advocating for continuous innovation and interdisciplinary collaboration to mitigate the pervasive threat posed by deepfake technology.

Keywords*:* Deepfake, Generative Adversarial Networks (GAN), Hybrid Models, Deep Learning, Explainable AI (XAI)

I. INTRODUCTION

The rapid proliferation of deepfake technology represents one of the most formidable challenges in the digital information landscape. Deepfake material, developed with deep learning models like GANs, may effectively duplicate an individual's look and speech, typically with malevolent intent. These manipulations pose serious risks, from spreading misinformation and political propaganda to undermining privacy and national security. The need for effective detection mechanisms has never been more critical, especially as the technology to create deepfakes becomes more accessible.

Deepfake content, which is produced with the use of deep learning models like GANs, may mimic people's speech and look quite well, frequently with malevolent intent. However, most detection models are complex, acting as “black boxes” whose decision-making processes are not easily interpretable. This opaqueness raises ethical and trust-related concerns, as the models' decisions cannot be easily justified or scrutinized, particularly in sensitive applications involving legal, political, or media contexts. Explainable AI (XAI) has been offered as a remedy for these issues. XAI seeks to make AI models more visible and intelligible, promoting informed trust and allowing stakeholders to comprehend and confirm the system's judgments.

This paper highlights the crucial role of XAI while offering a thorough analysis of deepfake detection techniques. We examine various techniques, including CNNs for spatial analysis, RNNs for temporal feature extraction, hybrid models that combine both, and advanced multimodal frameworks that integrate audio-visual cues. Furthermore, we explore how XAI methods can improve model interpretability and discuss the metrics and datasets used to benchmark these detection systems. Our analysis concludes with a discussion of the current limitations and recommendations for future research in this ever-changing field.

http://doi.org/10.63766/spujstmr.24.000034

II. LITERATURE REVIEW

The rapid advancement of deepfake technology has led to the development of numerous detection methods, each employing different approaches and architectures. Mas Montserrat et al. (2020) introduced a method combining convolutional and recurrent neural networks for efficient face manipulation detection, highlighting the importance of temporal analysis in deep-fake videos. Pashine et al. (2021) provided a comprehensive survey of facial manipulation detection techniques, comparing various CNN models like VGG-19, ResNet-50, and Xception, and underscoring the trade-offs between model complexity and accuracy. The efficacy of hybrid models, such as CNN-LSTM architectures, was further explored by Shaikh et al. (2023), who demonstrated how combining spatial and temporal feature extraction improves detection robustness. Additionally, Ismail et al. (2021) and Raza and Malik (2023) proposed innovative approaches incorporating multimodal frameworks, such as the YOLO-Face CNN-XGBoost model and Multimodal trace architecture, which jointly analyze audio and visual cues to enhance detection performance. To address challenges related to model interpretability and explainability, Groh et al. (2021) examined the role of Explainable AI (XAI), emphasizing the use of feature attribution and visualization techniques to make deepfake detection models more transparent and trustworthy. Collectively, these studies underscore the ongoing efforts to develop efficient, accurate, and interpretable deepfake detection systems, while also highlighting the necessity of future research that focuses on model generalizability and robustness against adversarial attacks.

III. DEEPFAKE DETECTION TECHNIQUES

Deepfake detection techniques have evolved significantly over time, primarily due to advances in deep learning and computer vision. The following subsections delve into the key methodologies used in this domain.

A. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a cornerstone of deepfake detection, particularly adept at extracting spatial features from images and video frames. CNN-based models like VGG-19, ResNet-50, and Xception have been extensively utilized for their ability to learn intricate patterns and detect artifacts indicative of deepfake manipulations [5]. The Xception network, for instance, employs depth wise separable convolutions, making it both efficient and powerful in identifying subtle inconsistencies.

The performance of CNN models is often benchmarked on datasets like FaceForensics++ and Celeb-DF, where they achieve high detection accuracies [6]. However, these models face challenges when dealing with unseen deepfake variants, highlighting a lack of generalizability. Moreover, CNNs are susceptible to adversarial attacks and often require extensive computational resources for training and deployment. Despite these limitations, CNNs remain a fundamental component of deepfake detection systems, and ongoing research focuses on enhancing their robustness and efficiency.

B. Recurrent Neural Networks (RNNs) and Temporal Analysis

Temporal analysis is crucial in deepfake detection, especially for video content where temporal inconsistencies can be a tell-tale sign of manipulation. Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, are used to capture and analyze temporal dependencies between video frames [9]. For example, the YOLO-Face convolution recurrent approach combines a CNN for spatial feature extraction and a Bi-LSTM for temporal analysis, resulting in commendable performance on large-scale datasets [6]. The strength of RNN-based models lies in their ability to detect subtle temporal anomalies, such as unnatural head movements or frame-to-frame inconsistencies. These models are often trained on datasets like the DeepFake Detection Challenge (DFDC) and evaluated based on metrics like recall, precision, and the F1-score [4]. Nevertheless, RNNs are computationally intensive and may suffer from vanishing gradient problems, prompting researchers to explore more efficient architectures or hybrid approaches that mitigate these issues.

C. Hybrid Models

Hybrid models that combine CNNs and RNNs have gained traction in the deepfake detection landscape. These models make use of the advantages of both architectures: RNNs for temporal feature modeling and CNNs for spatial feature extraction. A prominent example is the CNN-LSTM architecture, which has demonstrated superior performance in detecting complex deepfake manipulations by capturing both spatial and temporal features [9].

Despite their effectiveness, hybrid models are not without limitations. They require substantial computational power and may exhibit reduced performance when faced with adversarially crafted deepfake content [8]. Moreover, their interpretability remains a concern, as the combination of multiple architectures adds to the complexity of the model.

Integrating XAI techniques into hybrid models could enhance their transparency and facilitate a better understanding of the decision-making process.

D. Multimodal Approaches

The use of multimodal data, such as audio and visual cues, has emerged as a promising strategy for deepfake detection. The Multimodal trace framework, for example, uses Intramodality and InterModality Mixer Layers to process audio and visual features together, achieving state-of-the-art accuracy on the FakeAVCeleb dataset [16]. These models can detect inconsistencies that are not visible when each modality is considered separately by analyzing the synchronization of the audio and visual modalities.

Multimodal approaches have demonstrated robustness in cross-dataset evaluations, making them suitable for real-world applications. However, they also present challenges, including increased model complexity and the need for large, diverse datasets that encompass various types of deepfake manipulations. Future research in this area is expected to explore alternative integration techniques, such as attention mechanisms and graph-based models, to improve both performance and interpretability [17].

TABLE: I
COMPARATIVE ANALYSIS OF EXISTING DEEPFAKE DETECTION TECHNIQUES

Technique	Model Type	Accuracy (%)	Computational Efficiency	Interpretability	Robustness
CNN-Based	Deep Learning	90-95	Moderate	Low	Moderate
RNN-Based	Deep Learning	85-92	High	Low	Moderate
Hybrid CNN-RNN	Hybrid Model	92-97	Moderate	Moderate	High
Transformer-Based	Deep Learning	93-98	Low	High	High
Handcrafted Features	Traditional ML	80-88	High	High	Low
GAN-Based Detection	Deep Learning	88-94	Moderate	Low	Moderate

IV. EXPLAINABLE AI IN DEEPFAKE DETECTION

The increasing complexity of deepfake detection models necessitates a focus on Explainable Artificial Intelligence (XAI) to make these systems transparent and understandable. As detection algorithms become more sophisticated, it becomes crucial for stakeholders—including researchers, policymakers, and the public—to comprehend how these models arrive at their decisions. This section delves into the role of XAI in deepfake detection, emphasizing its methods and the benefits of incorporating interpretability into AI models.

A. Feature Attribution and Visualization

Feature attribution techniques are fundamental in XAI, helping to highlight which features most influence a model’s output. In the context of deepfake detection, feature attribution can identify specific facial regions or temporal segments of a video that contribute significantly to the decision of the model [12]. Methods like integrated gradients, Grad-CAM (Gradient-weighted Class Activation Mapping), and saliency maps are commonly used to visualize these attributions.

- Grad-CAM is particularly effective for CNN-based models, as it generates heatmaps that show which areas of an image are most important for a given classification [13]. For example, a Grad-CAM heatmap may reveal that a deepfake detector focuses on unnatural artifacts around the eyes or mouth, areas that are often difficult for generative models to replicate perfectly.

- Saliency maps can similarly be used to visualize which pixels or regions in an image influence the model's predictions. These maps provide insights into how the model perceives discrepancies in facial features, such as asymmetrical lighting or unnatural skin textures, which may be indicative of deepfake content [14].

Feature attribution not only aids researchers in understanding model behavior but also helps in debugging and refining detection algorithms. By identifying which features are most informative, developers can improve model robustness and efficiency [15]. Furthermore, these visualizations can be used to educate non-expert audiences about the strengths and limitations of deepfake detection technologies, fostering a more informed and skeptical public.

B. Model Transparency and Interpretability

Beyond feature attribution, model transparency encompasses the broader goal of making AI systems comprehensible at a structural and operational level. In deepfake detection, interpretability is crucial for several reasons:

Trust and Accountability: In applications where deepfake detection models are used for forensic purposes or media verification, understanding the model's reasoning is essential [16]. For instance, journalists or legal experts need to know why a model flagged a video as a deepfake, especially if the content has legal or societal implications

Bias Detection: XAI can help reveal biases in detection models. If a model consistently performs better or worse on certain demographic groups, interpretability tools can uncover these discrepancies, prompting the development of fairer algorithms [17]. For example, a model might exhibit bias against individuals with darker skin tones if it was trained on a dataset lacking sufficient diversity.

Techniques for enhancing model transparency include rule-based explanations, decision trees, and surrogate models that approximate the behavior of more complex neural networks [18]. LIME (Local Interpretable Model-agnostic Explanations) is one such tool that can explain individual predictions by approximating the deepfake detection model with an interpretable one [19].

Another emerging approach is the use of attention mechanisms in neural networks. Attention layers highlight which parts of an input the model is focusing on, making the decision-making process more interpretable [20]. For example, an attention-based model may allocate more focus to areas where facial artifacts are most pronounced, providing a clear rationale for its predictions.

V. EVALUATION METRICS AND DATASETS

The efficacy of deepfake detection models is often assessed using a combination of standard evaluation metrics and benchmark datasets. These metrics and datasets play a crucial role in establishing the reliability and robustness of different detection methods.

A. Evaluation Metrics

In deepfake detection, evaluation metrics such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUROC) are frequently employed. Different viewpoints on model performance are offered by each of these metrics:

The percentage of accurate predictions among all predictions is known as accuracy. Accuracy is helpful, but if the dataset is unbalanced, it can be deceptive because it ignores false positives and false negatives.

Precision and Recall offer a more nuanced view. Precision indicates how many of the detected deepfakes are truly deepfakes, while recall measures how many actual deepfakes were correctly identified. These metrics are crucial in scenarios where false positives or false negatives have severe consequences.

A balanced metric that is especially helpful when working with imbalanced datasets is the F1-score, which is the harmonic mean of precision and recall.

AUROC assesses how well the model can differentiate between classes at various threshold values. A model that performs well in distinguishing between authentic and fraudulent content under various circumstances is indicated by a high AUROC value.

In addition to these metrics, latency and computational efficiency are important considerations for real-time applications, such as social media platforms that need to screen uploaded content. The robustness of a model, often tested through adversarial attacks, is another critical metric, as it measures the model's resilience to intentional manipulations designed to evade detection.

B. Benchmark Datasets

The development and evaluation of deepfake detection models rely heavily on publicly available datasets. These datasets vary in size, complexity, and the types of manipulations they include. Some of the most popular datasets are:

- **FaceForensics++:** This dataset includes over 1.8 million manipulated images and videos created using four different deepfake generation methods. It provides a benchmark for testing the effectiveness of detection models, particularly in the presence of compression artifacts and low-quality videos.

- **Celeb-DF (V2):** Known for its high-quality deepfake videos, Celeb-DF (V2) presents a significant challenge for detection models. The dataset addresses limitations found in earlier datasets, such as unnatural head poses and low visual quality, by offering more realistic manipulations.

- **DeepFake Detection Challenge (DFDC):** Created by Facebook in collaboration with other institutions, the DFDC dataset contains over 100,000 videos with a diverse set of actors and deepfake techniques. It serves as a comprehensive resource for evaluating model performance on a large and varied dataset.

- **WildDeepfake:** This real-world dataset includes deepfake videos collected from the internet, making it a more challenging testbed for detection models. Unlike controlled datasets, WildDeepfake features diverse scenes and lighting conditions, reflecting the complexities of real-world scenarios.

The choice of dataset significantly influences a model's performance and generalizability. While some models excel on specific datasets, they may struggle with others, highlighting the need for cross-dataset evaluations. Moreover, dataset diversity is crucial for training robust models that perform well across different demographics and environmental conditions.

VI. LIMITATIONS AND FUTURE DIRECTIONS

Despite significant advancements, deepfake detection models face several limitations. These include issues related to generalizability, computational efficiency, and the ever-evolving nature of deepfake generation techniques. Most detection methods are optimized for specific types of manipulations and may falter when confronted with novel or highly sophisticated deepfakes. Additionally, the reliance on large and diverse datasets poses challenges, as obtaining and annotating such data is resource-intensive.

Future research should focus on developing models that can generalize across different domains and withstand adversarial attacks. Hybrid and multimodal approaches that combine visual, audio, and contextual data hold promise for improving detection accuracy and robustness. Moreover, the integration of XAI techniques will be critical for making these models more transparent and trustworthy. Interdisciplinary collaboration among AI researchers, ethicists, and policymakers will also be essential in addressing the ethical and societal implications of deepfake technology.

VII. CONCLUSION

In conclusion, the integration of Explainable AI in deepfake detection offers a path toward more transparent, interpretable, and reliable AI systems. As deepfake technology continues to advance, the need for robust and explainable detection methods becomes increasingly urgent. By combining state-of-the-art detection techniques with XAI principles, researchers can build systems that are not only effective but also ethically sound and trustworthy. This paper provides a comprehensive overview of the current landscape in deepfake detection and highlights the critical role of XAI in shaping the future of this rapidly evolving field.

REFERENCES

[1] Mas Montserrat, D., Hao, H., Yarlagadda, S. K., Baireddy, S., Shao, R., Horváth, J., & Delp, E. J. (2020). Deepfakes Detection with Automatic Face Weighting. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2851-2859.

[2] Pashine, S., Mandiya, S., Gupta, P. K., & Sheikh, R. (2021). *Deep Fake Detection: Survey of Facial Manipulation Detection Solutions. ArXiv*, abs/2106.12605.

[3] Groh, M., Epstein, Z., Firestone, C., & Picard, R. W. (2021). *Deepfake detection by human crowds, machines, and machine-informed crowds.* Proceedings of the National Academy of Sciences of the United States of America, 119.

[4] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Canton-Ferrer, C. (2020). *The DeepFake Detection Challenge Dataset*. *ArXiv*, abs/2006.07397.

[5] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1-11.

[6] Ismail, A., Elpeltagy, M. S., Zaki, M. S., & Eldahshan, K. A. (2021). *Deepfake video detection: YOLO-Face convolution recurrent approach*. PeerJ Computer Science, 7.

[7] Vidyavathi, B. M., Ahmed, A. F., Ayain, L., & Fatima, S. M. (2024). *Deepfake Detection using Deep Learning*. International Journal of Advanced Research in Science, Communication and Technology.

[8] Heidari, A., Navimipour, N. J., Dag, H., & Unal, M. (2023). *Deepfake detection using deep learning methods: A systematic and comprehensive review.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14.

[9] Shaikh, M. S., Nirankari, L., Pardeshi, V., Sharma, R., & Kale, S. (2023). *Deepfake Detection Using Deep Learning (Cnn+Lstm).* International Journal of Scientific Research in Engineering and Management.

[10] Naskar, G., Mohiuddin, S. M., Malakar, S., Cuevas, E., & Sarkar, R. (2024). *Deepfake Detection using Deep Feature Stacking and Meta-learning.* Heliyon.

[11] Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y. G. (2020). WildDeepfake: *A Challenging Real-World Dataset for Deepfake Detection.* Proceedings of the 28th ACM International Conference on Multimedia.

[12] Chen, L., Zhang, Y., Song, Y., Liu, L., & Wang, J. (2022). *Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection*. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18689-18698.

[13] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). *Deepfake Detection: A Systematic Literature Review. IEEE Access*, 10, 25494-25513.

[14] Ismail, A., Elpeltagy, M. S., Zaki, M. S., & Eldahshan, K. A. (2021). *A New Deep Learning-Based Methodology for Video Deepfake Detection Using XGBoost*. *Sensors*, 21.

[15] Ramachandran, S., Nadimpalli, A. V., & Rattani, A. (2021). An Experimental Evaluation on Deepfake Detection using Deep Face Recognition. *2021 International Carnahan Conference on Security Technology (ICCST)*, 1-6.

[16] Raza, M. A., & Malik, K. M. (2023). *Multimodal trace: Deepfake Detection using Audio visual Representation Learning*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 993-1000.

[17] Rekha, G., & Shashi, P. (2023). *Deepfake: Creation and Detection using Deep Learning*. International Journal for Research in Applied Science and Engineering Technology.

[18] Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2019). *Deepfake video detection through optical flow based CNN*. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.