Review on Deepfake Detection

Sreejith Mohan
Department of CSE,
Sree Narayana Gurukulam
College of Engineering
smsreejith29@gmail.com

Jisto Kuriakose
Department of CSE,
Sree Narayana Gurukulam
College of Engineering
jistokuriyakose@gmail.com

Vishnudas T
Department of CSE,
Sree Narayana Gurukulam
College of Engineering
vishnudas82507@gmail.com

Mithun Raju
Department of CSE,
Sree Narayana Gurukulam
College of Engineering
Mithunraju142@gmail.com

Abstract- This review delves into the dynamic landscape of deepfake technology and advancements in detection methods. It addresses concerns about misinformation and trust erosion due to the rapid evolution of synthetic media created by algorithms like GANs. The paper examines various detection methodologies, including deep learning models (Xception, CNN, VGG, ResNet18) and innovative approaches like error-level analysis and attention mechanisms. While each method shows strengths, they also have limitations, such as generalization challenges and computational complexity. The review emphasizes ongoing efforts to improve these techniques, highlighting the necessity for updated datasets and ethical considerations. A crucial resource for researchers and policymakers combating the risks associated with deepfake misuse.

Keywords: -GANs, Xception, CNN, VGG, ResNet18

1. INTRODUCTION

The rapid evolution of deepfake technology, fuelled by advancements in deep learning algorithms, has ushered in a new era of synthetic media creation. Deepfakes, utilizing techniques like generative adversarial networks (GANs), enable the seamless manipulation of facial expressions, speech, and body movements to convincingly superimpose one individual's likeness onto another. However, the escalating sophistication of deepfake technology has raised serious concerns about its potential misuse, particularly in spreading misinformation, influencing public opinion, and eroding trust in authentic media.

This review paper comprehensively examines various research endeavours aimed at tackling the challenges posed by deepfake technology. The selected studies employ diverse methodologies, ranging from deep learning models like Xception, CNN, VGG, and ResNet18 to innovative approaches like error-level analysis, stream descriptors, and attention mechanisms. The focus of these investigations spans the detection and classification of deepfakes, often with an emphasis on diverse applications such as financial fraud prevention, identification of celebrity manipulated videos, and curbing the dissemination of fake news and misinformation.

The strength of these methodologies lies in their ability to analyse and identify patterns indicative of deepfake manipulation, offering promising results in terms of detection accuracy. However, each approach comes with its own set of limitations, whether it be challenges in generalization across diverse datasets, computational complexity, or susceptibility to certain types of deepfake methods. As the research community grapples with these challenges, the importance of developing effective and reliable deepfake detection systems becomes increasingly evident. This review not only provides an overview of existing deepfake detection techniques but also highlights the strengths and drawbacks of each approach. Additionally, it sheds light on the ongoing efforts to refine and enhance these methods, emphasizing the need for continual updates to datasets and consideration of ethical implications. As the threat of deepfake technology looms large, this review aims to serve as a valuable resource for researchers, policymakers, and technologists working towards mitigating the risks associated with the misuse of synthetic media.

2. LITERATURE SURVEY

There are several studies have proposed various methods for identifying deepfakes. This section will discuss some of the most relevant sources in this area.

In paper [1], their pursuit to identify videos generated using deepfake technology, the authors employ a deep learning model, Xception, to detect fake videos

generated by mainstream deepfake methods with high accuracy. However, they acknowledge a potential limitation as the model might not perform as effectively on fake videos generated using different approaches. The study underscores the challenge of adapting video input to deep learning models designed for images.

The paper [2] presents a web-based platform for users to upload videos and classify them as fake or real, addressing various types of deepfakes. The approach involves splitting the video into frames, followed by face cropping. Notably, the model processes cropped frames directly for detection. A notable drawback is the inability to detect audio deepfakes, emphasizing a potential limitation in the overall detection capabilities.

The survey done in the paper [3], the authors propose a novel deep fake detection and classification method combining Error Level Analysis (ELA) and deep learning. This methodology involves resizing images to CNN's input layer and performing ELA to identify digital manipulation at the pixel level. The proposed technique achieves high accuracy, especially with ResNet18 and KNN. However, a challenge highlighted is the need for regular updates to the dataset, pointing to the dynamic nature of deepfake generation techniques.

The authors of the paper [4] introduces a web application for deepfake detection and generation. The detection side involves face extraction, training on a dataset, and employing a machine learning classifier. 14 On the generation side, users upload two videos, and the model extracts frames for merging faces with enhancements. The study demonstrates slightly better accuracy with CNN compared to VGG. The authors intend to enhance dataset generalization and augmentation techniques to improve the system's ability to detect user-inserted data, emphasizing the importance of regular dataset updates.

In paper [5], This study combines Convolutional Neural Networks (CNNs) for facial feature extraction with Recurrent Neural Networks (RNNs) to capture temporal patterns in video frames for deepfake detection. The model integrates facial detection systems to focus exclusively on faces within frames, achieving a promising 72.5% accuracy on the validation set. While demonstrating potential in deepfake detection, the study acknowledges challenges such as overfitting and computational cost associated with temporal information integration and adaptive face detection.

In the study of paper [6], a deep learning-based solution is proposed for the detection and localization of spatiotemporal manipulation in digital videos. The model, utilizing Long Short-Term Memory (LSTM) architecture, is specifically designed to excel in both temporal and spatial localization of forgery within videos. The research employs the REWIND dataset, comprising 10 original and 10 forged videos, with 16 videos allocated for model training and the remaining four for testing. The evaluation of model performance is conducted at three distinct levels: pixel-level for spatial localization, frame-level for temporal localization, and video-level for assessing forgery detection accuracy. The investigation delves into considerations such as data dependency and computational complexity to provide a comprehensive understanding of the proposed spatiotemporal localization approach.

Paper [7] conducts a comprehensive review of existing literature on deepfake creation algorithms and cutting-edge methods for detecting deepfakes. The authors meticulously analyse various deepfake generation techniques, exploring the underlying algorithms responsible for crafting fake images and videos that pose significant threats to privacy, democracy, and national security. Additionally, the study delves into the current landscape of deepfake detection methods, scrutinizing their effectiveness and limitations. Through this extensive literature review, the paper aims to provide a thorough understanding of the challenges posed by deepfake technologies and identifies potential research trends and directions to address these issues effectively. By highlighting challenges, trends, and future directions in the realm of deepfakes, the study serves as a roadmap for the artificial intelligence research community. Furthermore, emphasizing the consideration of potential misuse, societal impacts, and the responsible development and deployment of countermeasures would provide a more holistic perspective on the subject.

An innovative approach in paper [8] employs an adversarial dual-branch data augmentation framework and introduces a modified attention mechanism within ResNet50 to enhance the detection of forged images. In data preprocessing, traditional random sampling augmentation is combined with adversarial samples, enhancing diversity and uniform hardness in the forged image dataset for a more robust model. This approach improves the accuracy of detecting forged images, even in unpredictable data distributions. The modified attention mechanism addresses uneven attention distribution issues in the baseline ResNet50 by assigning increased weight to forged traces in multi-scale feature maps. Jensen–Shannon divergence loss and cosine annealing algorithms further enhance

accuracy and convergence speed during training. While validation on standard and corrupted datasets shows significant improvements compared to mainstream methods, it's acknowledged that the introduced complexity and increased computational requirements during training and inference are noteworthy considerations.

Siwei Lyu [9] categorizes deepfake detection methods into signal feature-based, physical/physiological-based, and data-driven approaches. Signal feature-based methods focus on anomalies in the generation process but are sensitive to disturbances. Physical/physiological-based methods expose deepfakes through violations of physics or human physiology, offering intuitive results but relying on robust Computer Vision algorithms. Data-driven methods, employing Deep Neural Networks (DNNs), achieve high performance with large datasets, but their success hinges on the quality and diversity of training data and model design. Each category exhibits distinct strengths and limitations, influencing the overall effectiveness of deepfake detection. Training on specific datasets is crucial for ensuring model relevance, but this approach may suffer from a lack of generalization when faced with unseen models, underscoring the need for a balanced consideration of these factors in designing comprehensive deepfake detection systems.

In paper [10], the proposed approach demonstrates a robust capacity to generalize across diverse datasets by leveraging unsupervised learning, specifically the EM algorithm, to uncover hidden structures in Deepfake images. By focusing on local pixel correlations within all layers, particularly the Transpose Convolution layers in GANs, the method forms clusters through the estimation of distribution parameters. This spatially capturing model effectively encodes the structural features of Transpose Convolution Layers, enabling it to discern the characteristic patterns indicative of Deepfakes. The operational dependence on the direct connection between local pixel correlation and the actions performed by Transpose Convolution layers enhances the model's adaptability, allowing it to generalize effectively across a wide range of datasets and reliably identify Deepfake images based on their intrinsic structural properties.

The methodology for deepfake detection encompasses a comprehensive approach in the paper [11], starting with the careful curation of diverse datasets, their standardization, and preprocessing. State-of-the-art deep learning models, such as 3D CNNs, are deployed, and a structured training/validation/testing split, along with data augmentation, ensures model robustness. Feature extraction focuses on capturing temporal patterns and unique artifacts specific to Deepfake generation. The process involves fine-tuning based on rigorous performance evaluations using metrics like accuracy, precision, recall, and F1 score, supported by cross-validation to enhance robustness. Post-processing is employed to refine results, and ethical considerations are addressed in the detection process. Despite the efficiency gained through automation in analysing vast video volumes amid the widespread use of deepfakes, a notable drawback is the persistent challenge of striking a balance between false positives and false negatives, which continues to impact the overall accuracy of the deep learning-based deepfake detection system. This ongoing challenge underscores the need for further refinement to achieve a more optimal trade-off between detection sensitivity and specificity.

The authors introduce a method focused on training a video frame classifier for deepfake detection, incorporating key elements such as face extraction, alignment, and a fine-tuned VGG-16 convolutional model in paper [12]. The classifier is enriched with features like batch normalization, dropout, and a custom two-node dense layer designed for real and fake class identification. Adam Optimizer is employed to enhance learning, and transfer learning evaluates simple features, aiding in the identification of anomalies introduced during fake image creation. The post-processing stage involves video analysis, utilizing the trained model's results for all frames. The method achieves high accuracy in deepfake video detection by leveraging the capabilities of CNNs to learn intricate patterns and features in image data. However, it acknowledges the potential limitation of reduced generalization and detection performance when confronted with limited or biased training data. Addressing these challenges is crucial to ensure the robustness and effectiveness of the deepfake detection system across diverse scenarios and datasets.

The authors of the paper [13] This paper presents a novel temporal-aware system for the automatic detection of deepfake videos. The system employs a two-step process, utilizing a convolutional neural network (CNN) to extract frame-level features and subsequently training a recurrent neural network (RNN) to classify whether a video has undergone manipulation. The use of a simple convolutional LSTM structure enables accurate predictions with as little as 2 seconds of video data, leading to very fast detection of fake videos. The method demonstrates highly accurate results and has been tested

and proven to provide fast and reliable deepfake detection. Specifically designed to identify videos manipulated using popular deepfake generating techniques, the system excels in scenarios where such methods are applied. However, it acknowledges a challenge in identifying videos manipulated using unseen and unorthodox methods, indicating a potential limitation in its adaptability to novel manipulation techniques not covered by the training data. Despite this, the proposed method stands out for its efficiency in detecting common deepfake generation methods quickly and accurately.

The paper [14] introduces an end-to-end deep learning model for deepfake video detection, presenting a novel hybrid framework named Inception ResNet-BILSTM. This model is specifically designed to be robust across different ethnicities and varied illumination conditions. The approach involves extracting faces from videos and feeding them into Inception ResNet-BILSTM to capture frame-level learnable details. The resulting features are then utilized to classify videos as either real or fake. The proposed model demonstrates effectiveness in detecting deepfake videos created using various techniques, coping with variations in illumination conditions, and addressing diverse ethnicities. However, it is important to note that this method is not applicable for the detection of deepfake images or audio. The focus solely on video content suggests a limitation in its scope, and users should consider alternative approaches for comprehensive multimedia-based deepfake detection scenarios.

The paper [15] proposes a method titled "Deepfakes Detection with Automatic Face Weighting" that relies on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for the extraction of visual and temporal features from faces in videos, enabling accurate detection of manipulations. The evaluation is conducted using the DFDC dataset, and the method introduces an automatic face weighting mechanism. It is characterized by fast processing, demonstrating good accuracy in detecting manipulated content. However, the approach dismisses any analysis of audio content, limiting its scope to visual and temporal features. Additionally, the method may face challenges in scenarios where multiple faces are present, potentially making it difficult to identify deepfakes in such complex situations. Despite these limitations, the proposed method offers an efficient and accurate solution for deepfake detection, particularly in cases involving single faces in videos.

3. RESULTS / DISCUSSION

In recent years, the proliferation of deepfake technology has underscored the urgent need for robust and effective detection methods to safeguard the authenticity of visual content. This review examines several cutting-edge approaches in the realm of deepfake detection, each leveraging advanced deep learning models to tackle the challenges posed by manipulated videos. From the temporal-aware system utilizing a combination of convolutional and recurrent neural networks for rapid and accurate detection, to the end-to-end deep learning model designed for video analysis with a focus on ethnic and illumination variations, these methodologies showcase significant advancements in the field. Additionally, the exploration of features such as automatic face weighting highlights innovative strategies for enhancing detection precision. While each approach demonstrates commendable strengths, they also exhibit nuanced limitations, such as overlooking audio content analysis or facing challenges in scenarios with multiple faces. This review aims to provide a comprehensive overview of these state-of-the-art deepfake detection methods, shedding light on their capabilities, strengths, and potential areas for improvement.

The following table presents a comprehensive review of various deepfake detection methodologies, highlighting their key features, methods, and evaluation metrics.

Year	Name	Methods used	Result
2020	[1] Deepfake Detection through Deep Learning	Xception (deep learning model)	High accuracy in detecting mainstream deepfake methods, potential limitation for different approaches
2020	[2] Deepfake Video Detection using Neural Networks	Web-based platform, frame splitting, face cropping	Effective for detecting various types of deepfakes, but unable to detect audio deepfakes
2020	[3] Deep fake detection and classification using error-level analysis and deep learning	Combination of Error Level Analysis (ELA) and deep learning	High accuracy with ResNet18 and KNN, challenges include the need for regular dataset updates
2023	[4] DeepFakeDG : A Deep Learning Approach for Deep Fake Detection and	Face extraction, machine learning classifier (CNN vs. VGG)	Slightly better accuracy with CNN, emphasis on dataset generalization and augmentation

	Generation		
2019	[5] We Need No Pixels: Video Manipulation Detection Using Stream Descriptors	CNNs for facial feature extraction, RNNs for temporal patterns	72.5% accuracy on validation set, challenges include overfitting and computational cost
2020	[6] Video Manipulation Detection and Localization Using Deep Learning	Deep learning-based spatiotemporal manipulation detection (LSTM)	Comprehensive evaluation at pixel, frame, and video levels, considerations include data dependency and computational complexity
2020	[7] Deep Learning for Deepfakes Creation and Detection	Comprehensive review of deepfake creation algorithms and detection methods	Analysis of existing literature, identification of challenges, and potential research trends
2023	[8] Deepfake Detection Algorithm Based on Dual-Branch Data Augmentation and Modified Attention Mechanism	Adversarial dual-branch data augmentation, modified attention mechanism	Improved detection accuracy, but increased computational requirements
2022	[9] Deepfake detection	Categorization into signal feature-based, physical/physiological-based, and data-driven approaches	Discussion of strengths and limitations of each category in deepfake detection
2020	[10] DeepFake Detection by Analyzing Convolutional Traces	Unsupervised learning with EM algorithm for Deepfake image detection	Generalizes effectively across diverse datasets, relies on local pixel correlations
2022	[11] Extensive Analysis of Deep Learning-based Deepfake Video Detection	3D CNNs, fine-tuning, feature extraction, post-processing	Efficient deepfake detection but challenges in balancing false positives and false negatives
2020	[12] Deepfake Video Detection Using Convolutional Neural Network	Video frame classifier with VGG-16, face extraction, alignment	High accuracy in deepfake video detection, potential limitation in generalization with biased training data
2021	[13] Deepfake Video Detection Using Recurrent Neural Networks	Two-step process using CNN and RNN for temporal-aware deepfake detection	Fast and reliable detection of common deepfake methods, potential challenge with unseen and unorthodox methods
2022	[14] Deepfakes Examiner: An End-to-End Deep Learning Model for Deepfakes Videos Detection	Inception ResNet-BILSTM hybrid framework	Robust across ethnicities and illumination conditions, limited to video content detection
2023	[15] Deepfakes Detection with Automatic Face Weighting	CNNs and RNNs for visual and temporal features, automatic face weighting	Efficient and accurate detection of manipulated content, limited scope to visual and temporal features

Overall, the studies reviewed demonstrate a growing concern and active research efforts in the field of deepfake detection. Different methods and approaches have been explored to address the challenges posed by the widespread use of deepfake technology, which threatens democracy, justice, and public trust.

4. CONCLUSION

In conclusion, this extensive literature review on deepfake detection methods reveals a complex landscape marked by growing concerns about the potential threats posed by synthetic media. The surveyed studies collectively emphasize the urgency of addressing challenges associated with deepfake technology, spanning privacy, democracy, justice, and national security. Various methodologies, ranging from the use of deep learning models like Xception and Inception ResNet-BILSTM to the integration of techniques like Error Level Analysis (ELA) and adversarial dual-branch data augmentation, are explored. Despite significant progress, persistent challenges include the dynamic nature of deepfake generation techniques requiring regular dataset updates and the need for generalization across diverse scenarios. The importance of efficient and fast detection methods is underscored, with innovative approaches such as unsupervised learning using the EM algorithm and temporal-aware systems leveraging convolutional LSTM structures. Ethical considerations are highlighted, emphasizing responsible development and

deployment of countermeasures to address concerns about privacy, trust, and the democratic process. The interdisciplinary nature of these efforts, spanning computer vision, deep learning, and ethical considerations, underscores the complexity of the deepfake detection landscape. The surveyed literature provides a valuable roadmap for the AI research community, offering insights into current methodologies, limitations, and identifying potential research trends to fortify society against the impacts of deepfake technologies. Continued collaboration and innovation are essential to develop adaptable detection systems capable of staying ahead of the evolving synthetic media landscape.

5. REFERENCE

[1] Deng Pan, Lixian Sun, Rui Wang, Xingjian Zhang, Richard O. Sinnott "Deepfake Detection through Deep Learning" 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), DOI 10.1109/BDCAT50828.2020.00001

[2] Abhijit Jadhav, Abhishek Patange, Jay Patel, Hitendra Patil, Manjushri Mahajan "Deepfake Video Detection using Neural Networks" IJSRD - International Journal for Scientific Research & Development | Vol. 8, Issue 1, 2020 | ISSN (online): 2321-0613

[3] Rimsha Rafque, RahmaGantassi, RashidAmin, Jaroslav Frnda, Aida Mustapha, Asma HassanAlshehri "Deep fake detection and classification using error-level analysis and deep learning" https://doi.org/10.1038/s41598-023-34629-3

[4] Zeina Ayman, Natalie Sherif, Mariam Mohamed, Mohamed Hazem, Diaa Salama " DeepFakeDG: A Deep Learning Approach for Deep Fake Detection and Generation" Journal of Computing and Communication Vol.2 , No.2 , PP. 31-37 , 2023

[5] David Guera, Sriram Baireddy, Paolo Bestagini, Stefano Tubaro, Edward J. Delp "We Need No Pixels: Video Manipulation Detection Using Stream Descriptors" arXiv:1906.08743v1 [cs.LG] 20 Jun 2019 https://doi.org/10.1007/978-3-030-01228-1_8

[6] Hemal Mamtora, Kevin Doshi, Shreya Gokhale, Surekha Dholay, Chandrashekhar Gajbhiye "Video Manipulation Detection and Localization Using Deep Learning" 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) DOI: 10.1109/ICACCCN51052.2020.9362923

[7] Thanh Thi Nguyen, Tien Dung Nguyen, Cuong M. Nguyen, Saeid Nahavandi "Deep Learning for Deepfakes Creation and Detection" https://www.researchgate.net/publication/336058980

[8] Da Wan, Manchun Cai, Shufan Peng, Wenkai Qin and Lanting Li "Deepfake Detection Algorithm Based on Dual-Branch Data Augmentation and Modified Attention Mechanism" Appl. Sci. 2023,13, 8313. https://doi.org/10.3390/app13148313

[9] Siwei lyu "Deepfake detection" 2022 H.T.Sencaretal.(eds.), MultimediaForensics, Advances in Computer Vision and Pattern, https://doi.org/10.1007/978-981-16-7621-5_12

[10] Luca Guarnera,Oliver Giudice,Sebastiano Battiato "DeepFake Detection by Analyzing Convolutional Traces" 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),DOI 10.1109/CVPRW50498.2020.0034126

[11] D. Myvizhi1, J. C. Miraclin Joyce Pamila2 "Extensive Analysis of Deep Learning-based Deepfake Video Detection " 2022 Journal of Ubiquitous Computing and Communication Technologies, March 2022, Volume 4, Issue 1, Pages 1-8 DOI: https://doi.org/10.36548/jucct.2022.1.001

[12] Aarti Karandikar, Vedita Deshpande, Sanjana Singh, Sayali Nagbhidkar, Saurabh Agrawal " Deepfake Video Detection Using Convolutional Neural Network " Volume 9 No.2, March -April 2020 International Journal of Advanced Trends in Computer Science and Engineering, https://doi.org/10.30534/ijatcse/2020/62922020

[13] David Guera, Edward J. Delp " Deepfake Video Detection Using Recurrent Neural Networks" Deep feature interpolation for image content changes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6090–6099, July 2021

[14] Hafsa Ilyas, Aun Irtaza, Ali Javed, Khalid Mahmood Malik " Deepfakes Examiner: An End-to-End Deep Learning Model for Deepfakes Videos Detection" 2022 16th International Conference on Open Source Systems and Technologies (ICOSST) | 978-1-6654-6477-2/22/2022 IEEE | DOI: 10.1109/ICOSST57195.2022.10016871

[15] Daniel Mas Montserrat, Hanxiang Hao, S. K. Yarlagadda, Sriram Baireddy, Ruiting Shao, Janos Horv'ath, Emily Bartusiak, Justin Yang, David G ´ uera, Fengqing Zhu, Edward J. Delp "Deepfakes Detection with Automatic Face Weighting" 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)/DOI10.1109/CVPRW50498.2020.00342