A Comprehensive Review of Media Forensics and Deepfake Detection Technique

Megha Kandari
Dept. of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, India
megakandari134@gmail.com

Vikas Tripathi
Dept. of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, India
vikastripathi.be@gmail.com

Bhaskar Pant
Dept. of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, India
pantbhaskar2@gmail.com

***Abstract*—In recent years, as the use of the internet and social media has advanced, so the number of fake media content also grows rapidly. Media forensic techniques are becoming essential to analyze and extract information from various types of media, such as images, videos, and audio recordings, to be used as evidence in legal proceedings. As of recent trends, social media is being flooded with sophisticated deepfakes. So evolving deepfake detection technique is required to curb the flow of misinformation via deepfake videos. This paper intends to present a review of evolving media forensics and deepfake detection techniques.**

Keywords—Deepfake, Deep learning, internet, Media forensic, Misinformation, and social media

I. INTRODUCTION

The quick advancements in the field of multimedia content manipulations are due to the easily available mobile applications provided by big tech companies such as Facebook, Snapchat, and FaceApp [1, 2]. It becomes difficult to differentiate between real and deepfakes videos. Sometimes other manipulated media are falsely considered Deepfakes, whereas these manipulated media is not Deepfake and are known as shallow fakes [3]. Deepfakes are those which are generated using deep learning techniques.

Deepfake technology is named after an anonymous Reddit user who went by the name "deepfakes" (deep learning + fake), who first shared videos of celebrities in adult clips created using deep learning techniques in 2017 [1]. Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's similarities. Deepfake algorithms can produce fake photos and videos that are so convincing that no one can tell them apart from real ones, as shown in Figure 1, the popular deepfake content present on the internet.

Although producing false content is not new, "the first known attempt at trying to swap someone's face, circa 1865, can be found in one of the iconic portraits of U.S. President Abraham Lincoln" [2]. Deepfake uses potent machine learning and artificial intelligence techniques (as shown in Figure 2) to edit or create audio and visual information that can be more easily deceiving.

Media forensics is a rapidly growing field that focuses on the detection, analysis, and manipulation of digital media [6, 7]. The use of deepfake technology has brought new challenges to the field, as it allows for the creation of highly realistic and convincing fake videos. Deepfake technology can proliferate different malicious activities like high school bullying, blackmailing people by creating fake adult content, changing public opinions in the political campaign, and spreading fake information [4].

Fig. 1. Example of the deepfake from YouTube video. Top: deepfake videos, bottom: original videos [4].

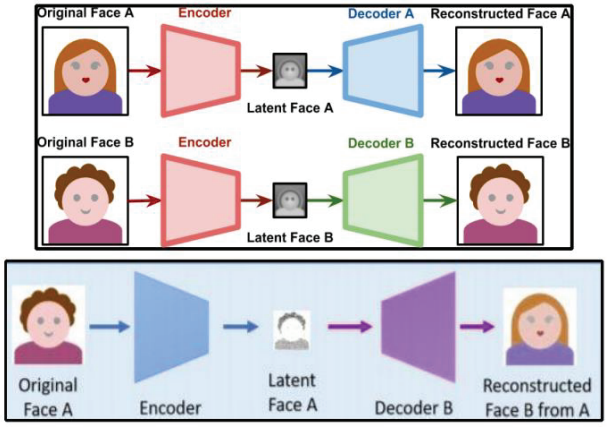

Fig. 2. A model of deepfake creation, combining two pairs of encoders and decoders. For training, two networks adopt the same encoder but a distinct decoder (top). An image of face A is encoded and decoded by decoder B to create a deep fake (bottom) [5].

This has led to the development of new techniques and methods for detecting deepfakes and other manipulated media [8].

II. MEDIA FORENSICS TECHNIQUES

Media forensics technology refers to the various techniques and tools used to detect and analyze digital media for authenticity and integrity. According to Bhagtani et al. [8], "The goals of media forensics are to answer the following questions: Is the media element manipulated

(detection)?; Where is it modified (localization)?; What tools and/or who modified it (attribution)?; and Why did they modify it (characterization)?".

This includes methods for detecting and mitigating deepfake content, which is created using artificial intelligence and machine learning techniques to manipulate or generate realistic-looking media. According to Gardiner [9], multimedia forensics plays a vital role in detecting and removing deepfake content and some other kinds of digital manipulations.

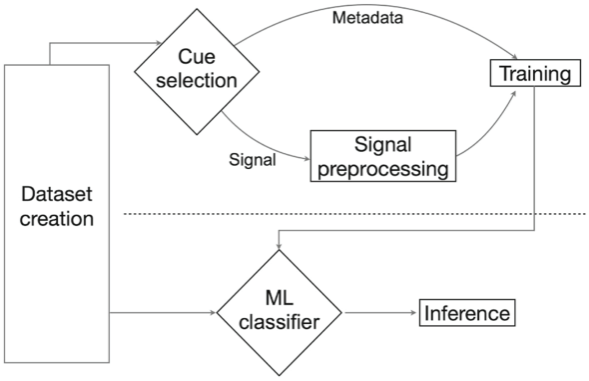

Fig. 3. A pipeline of the approaches of platform provenance analysis using signal processing and metadata [7].

Media forensics technology can be used in a variety of fields, including law enforcement, digital forensics, and news verification. The acquisition method, any post-processing that would have left a distinctive data trail, like a fingerprint, and any phase of image history that involves storage in a compressed or alternative format are all considered by multimedia forensics [10]. Multimedia signal processing techniques (shown in figure 3) to overcome the spread of deepfake by detecting traces of manipulation operations, are often underutilized [11]. Some common Media forensics techniques include:

- Image enhancement: techniques used to improve the visibility of details in an image, such as adjusting brightness and contrast [7].
- Video stabilization: techniques used to steady shaky or unstable video footage [7].
- Audio enhancement: techniques used to improve the clarity of audio recordings, such as filtering out background noise [4].
- Metadata analysis: techniques used to analyze and extract information from the metadata of a media file, such as the location, the date and time the file was created, the camera or device used to create the file by matching the sensor's fingerprints, and other information that may be useful in an investigation [7].
- Digital watermarking and steganography: Digital techniques used to detect or extract hidden information or watermarks embedded in a media file [12, 13].
- Hash value analysis: This technique is used to create a unique numerical value for a media file, which can be used to verify the authenticity of the file and detect any changes made to it [14, 15].
- Error Level Analysis: Technique to detect any manipulation in the image file. Error Level Analysis can be used for checking the compression ratio of images or video frames, for detecting manipulations in images, because the compression levels of both original and deepfake images are always different [16].

These are just a few examples of the many techniques used in media forensics, as the field is constantly evolving, and new techniques are being developed.

III. Deepfake Detection Techniques

Deepfake detection techniques are methods used to identify and flag manipulated media, such as videos or images that have been created using deepfake technology. There are several forms of deepfake and deepfake detection techniques [17], as given in Figure 4, but in this paper, we mainly focus on deepfake videos.

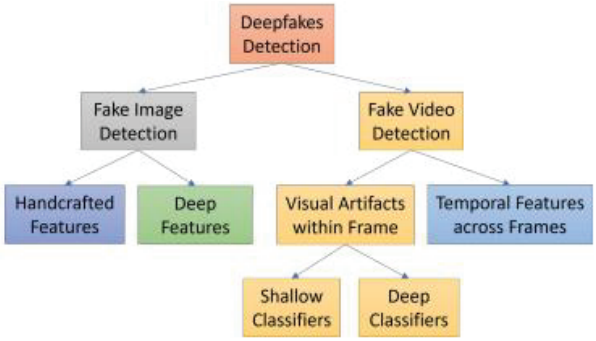

Fig. 4. Categories of deepfake image and deepfake video detection techniques [17].

Koopman, et al. [18] have tested the Photo Response non-uniformity (PRNU) technique for detecting deepfake videos. The authors created a dataset of 10 original videos with 20 to 40 seconds long length and 16 deepfake videos. The proposed model gave the resultant false positive and false negative rates as 3.8 % and 0% respectively. But the likelihood ratio was not that good because of the small dataset. PRNU procedure for forgery localization is shown in Figure 5.

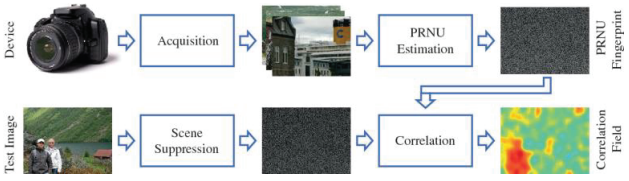

Fig. 5. Figure 5: Manipulation localization process based on PRNU [24].

Vamsi, et al. [19] have demonstrated a deepfake detection approach using residual neural network (ResNet), Convolutional Neural Network (CNN) algorithm, and Long short-term memory (LSTM). The authors used the Celeb-DF dataset for the training and testing of the model. In pre-processing stage video gets divided into frames with cropped face parts by using face recognition methods [20, 21]. The model obtains an accuracy of 91%. Detection allows

appropriate action to be taken to either delete the content or to flag the content tampered. For large datasets, the data augmentation technique on a trained CNN architecture is proven best [10, 22]. Siegel et al. [23], have recommended a method alternative to the deep learning method for deepfake detection. The authors described three sets of handcrafted features and some fusion techniques. The proposed approach analyses the mouth region only and does not consider if the lip sync with the audio is present or not. Hence more research needs to be done on this. Bhagtani et al. [8], implemented the data augmentation technique to binary cross entropy (trained CNN architecture). Future work needed to be done in combining neural network features and statistical features with each other to improve the attribution and confirmation of any media document.

Siegel et al. [25], introduced a domain-adapted forensic data model. The limitation of the paper is training conditions should be more focused on forensic workflow. The performance of detection tools that give better results can be decreased in real-time scenarios, so forensic tools need to be more robust to counter such issues [4]. There are some prominent challenges in deepfake detection, such as, if a fake video is just for a very small duration in a video, then it is difficult to detect [2]. CNN, Recurrent Neural Network (RNN), and LSTM can be effective to detect Deepfake by saving the information of short-term sequences [15]. If a Deepfake video consists of fake audio it can't be detected yet [23]. To overcome these types of challenges the fusion of audio deepfake detection methods [26] and video deepfake methods can be effective.

For coping with the Deepfake impacts in spreading false information on social media, integrated detection techniques should provide checking and deletion of such Deepfakes [17].

IV. Analysis

As deepfake technology continues to advance, it becomes more challenging to detect deepfakes accurately, requiring constant improvement and adaptation of detection algorithms.

Media forensics faces several issues and challenges, including:

- Tampering and manipulation: Digital media is often subject to manipulation, either intentionally or unintentionally, making it difficult to verify its authenticity [3, 4].
- The complexity of media: The increasing complexity and improvement of digital media, such as high-resolution images and videos, make it difficult to perform accurate and efficient analysis [27].
- Data degradation: Digital media can suffer from data degradation over time, which can affect the accuracy of forensics analysis.
- Privacy concerns: Media forensics often require access to sensitive data, which can raise privacy concerns and limit the availability of data for analysis [4].

Sometimes the sensor traces and PRNU fingerprints of sensors through which the photo was clicked can be used for detection [7, 4]. But sometimes in counter forensics, there are challenges, the PRNU fingerprint can also be tempered by skilled attackers [4]. Therefore, deep learning techniques can be effective in these scenarios.

There are numerous ways to create deepfakes, each with its own unique set of challenges for detection. This requires continuous research and development of new methods to detect each new type of deepfake.The effectiveness of the deepfake detection model is highly dependent on the quality and diversity of the training data. However, it is challenging to get large and diverse datasets [28].

TABLE I. Performance Summary of Deepfake Detection Techniques

Study	Approach	Result	Limitations
[1]	Photo Response Non Uniformity (PRNU)	p-value = $5.21*10-5$	To formulate guidelines for likelihood ratios, the dataset used was small.
[2]	Conv-LSTM, 20 frames, 40 frames, 80 frames	Test accuracy = 96.7 (20 frames), 97.1 (40 frames), 97.1 (80 frames)	The model is not able to detect videos that have been altered using undiscovered methods while being trained.
[29]	Metric learning approach.	AUC score = 99.2% (Celeb-DF), Accuracy = 90.71% (Neural Texture)	Methods do not have the ability to generalize to various datasets.
[30]	Data augmentation (CNN)	ROC AUC:.7% increase in AUC by using HF augmentation	The approach used is not so effective in the case of intra-dataset.
[16]	Alex Net and Shuffle Net, Error Level Analysis	Shuffle Net via KNN accuracy = 88.2% and Alex Net's vector accuracy = 86.8%.	Research needs to be done on the bases of pixel-level noise generation and motion.
[31]	DFDT framework.	AUC: 99.41%, 99.31%, and 81.35% for different datasets	Not effective in real-world datasets such as wilddeepfake.
[32]	VGGnet, CNN And DenseNet Data augmentation.	VGG19 accuracy = 95%	For increasing effectiveness, further research could be performed on unsupervised clustering methods.
[34]	modeling media forensic investigation pipelines	$\kappa < 0.4$ (DFeye, DFEAR and DFprob)	Quality of detectors should be enhanced.
[35]	Fusion techniques - Xception, DSP-FWA, and Capsule Network	AUC = 99%	To strengthen the efficacy of the false detectors against attacks not discovered during learning, further work needs to be done on inter-database scenarios.
[36]	CNN	Accuracy=98.2%	To protect privacy and integrity more accurately, the identification of bogus videos can be improved.

There are several deepfake detection techniques developed by researchers to date, Table I shows the advancements of deepfake detection techniques throughout recent years. Different approaches and limitations of the reviewed papers are mentioned in Table I. From the given comparison in Table I, we find out that Visual Geometry Group very deep CNN (VGGnet), CNN, and Densely connected convolutional Networks (DenseNet) implemented

with data augmentation give the best result among all other techniques with an accuracy of 95%.

V. Conclusion

Media forensics and deepfake detection are emerging areas of research that are attracting significant attention due to the increasing use of deepfake technology and its potential impact on society. There have been significant advances in the development of new methods for detecting deepfakes, including both model-based and signal-based approaches. Despite these advances, there remain numerous challenges and limitations in the field, particularly in terms of ensuring the robustness and scalability of the detection methods. As manipulators create better manipulation techniques, there's always a need for improving deepfake detection technologies.

References

[1] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?" Business Horizons, vol. 63, no. 2, pp. 135–146, 2020.

[2] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," In Proc. of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6.

[3] A. Kennedy, "What is a shallowfake video? - understanding the impact of deepfake videos video tutorial: Linkedin learning, formerly Lynda.com," LinkedIn, 13-Aug-2020.

[4] L. Verdoliva, "Media forensics and deepfakes: an overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.

[5] H. Vyas, "Deep fake creation by deep learning," International Research Journal of Engineering and Technology, vol. 7, no. 7, pp. 960-963, 2020.

[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. 5, p. 3095, 2023.

[7] C. Pasquini, I. Amerini, and G. Boato, "Media Forensics on Social Media Platforms: A survey," EURASIP Journal on Information Security, vol. 2021, no. 1, 2021.

[8] K. Bhagtani, A. K. S. Yadav, E. R. Bartusiak, Z. Xiang, R. Shao, S. Baireddy, and E. J. Delp, "An overview of recent work in media forensics: Methods and threats," arXiv preprint arXiv:2204.12067, 2022.

[9] N. Gardiner, Facial re-enactment, speech synthesis and the rise of the Deepfake. Edith Cowan University, Theses 2019.

[10] A. Piva, "An overview on image forensics," International Scholarly Research Notices, 2013, doi: 10.1155/2013/496701.

[11] E. Izquierdo, K. Chandramouli, A. T. S. Ho, and H. J. Kim, "Large-scale multimedia signal processing for security and digital forensics," Multimedia Tools and Applications, vol. 80, pp. 23313–23317, 2021.

[12] K. J. Giri and R. Bashir, "A block based watermarking approach for color images using discrete wavelet transformation," International Journal of Information Technology, vol. 10, pp. 139–146, 2018.

[13] S. Malik and R. K. Reddlapalli, "Histogram and entropy based digital image watermarking scheme," International Journal of Information Technology, vol. 11, no. 2, pp. 373–379, Nov. 2018.

[14] A. M Nagm, "A New Approach For Image Authentication Framework For Media Forensics Purpose," Journal of Computer Engineering & Information Technology, vol. 06, no. 06, 2017.

[15] C. Peersman, C. Schulze, A. Rashid, M. Brennan, and C. Fischer, "iCOP: Live forensics to reveal previously unknown criminal media on P2P networks," Digital Investigation, vol. 18, pp. 50–64, Sep. 2016.

[16] R. Rafique, M. Nawaz, H. Kibriya, and M. Masood, "Deepfake detection using error level analysis and deep learning," In Proc. of the 4th International Conference on Computing & Information Sciences (ICCIS), pp. 1–4, IEEE, 2021.

[17] T. T. Nguyen, Q. V. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," Computer Vision and Image Understanding, vol. 223.

[18] Koopman, M., Rodriguez, A.M. and Geradts, Z., "Detection of deepfake video manipulation," In Proc. of the 20th Irish machine vision and image processing conference (IMVIP), 2018, pp. 133-136.

[19] V. V. V. V. N. S. Vamsi, S. S. Shet, S. S. M. Reddy, et al., "Deepfake Detection in Digital Media Forensics," Global Transitions Proceedings, vol. 3, no. 1, pp. 74-79, 2022.

[20] S. S. Gangonda, P. P. Patavardhan, and K. J. Karande, "VGHN: variations aware geometric moments and histogram features normalization for robust uncontrolled face recognition," International Journal of Information Technology, vol. 14, pp. 1823–1834, 2022.

[21] M. K. Rusia and D. K. Singh, "An efficient CNN approach for facial expression recognition with some measures of overfitting," International Journal of Information Technology, vol. 13, no. 6, pp. 2419–2430, Sep. 2021.

[22] V. Kumar, V. Tripathi, and B. Pant, "Learning compact spatio-temporal features for fast content based video retrieval," International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 2, pp. 2402-2409, 2019.

[23] D. Siegel, C. Kraetzer, S. Seidlitz, and J. Dittmann, "Media forensics considerations on deepfake detection with hand-crafted features," Journal of Imaging, vol. 7, no. 7, 2021.

[24] D. Cozzolino and L. Verdoliva, "Multimedia Forensics Before the Deep Learning Era,". In Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks, 2022, (pp. 45-67). Cham: Springer International Publishing.

[25] D. Siegel, C. Krätzer, S. Seidlitz, and J. Dittmann, "Forensic data model for artificial intelligence based media forensics-Illustrated on the example of DeepFake detection," Electronic Imaging, vol. 34, pp. 1-6, 2022.

[26] Z. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," Algorithms, vol. 15, no. 5, 2022.

[27] J. Hendrix and D. Morozoff, "Media Forensics in the Age of Disinformation," Multimedia Forensics, p. 7-40, 2022.

[28] J. Parsola, D. Gangodkar, and A. Mittal, "Post event investigation of Multi-stream Video Data utilizing Hadoop Cluster," International Journal of Electrical and Computer Engineering, vol. 8, no. 6, p. 5089, 2018.

[29] A. Kumar, A. Bhavsar, and R. Verma, "Detecting deepfakes with metric learning," In Proc. of the 8th International Workshop on Biometrics and Forensics (IWBF), 2020, pp. 1–6.

[30] L. Bondi, E. D. Cannas, P. Bestagini, and S. Tubaro, "Training strategies and data augmentations in cnn-based deepfake video detection," In Proc. of the IEEE international workshop on information forensics and security (WIFS), 2020, pp. 1–6.

[31] A. Khormali and J.-S. Yuan, "Dfdt: An end-to-end deepfake detection framework using vision transformer," Applied Sciences, vol. 12, no. 6, 2022.

[32] M. Taeb and H. Chi, "Comparison of deepfake detection techniques through deep learning," Journal of Cybersecurity and Privacy, vol. 2, no. 1, pp. 89– 106, 2022.

[33] S. H. Silva, M. Bethany, A. M. Votto, I. H. Scarff, N. Beebe, and P. Najafirad, "Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models," Forensic Science International: Synergy, vol. 4, 2022.

[34] C. Kraetzer, D. Siegel, S. Seidlitz, and J. Dittmann, "Process-driven modelling of media forensic investigations-considerations on the example of deepfake detection," Sensors, vol. 22, no. 9, p. 3137, 2022.

[35] R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, and J. Fierrez, "Deepfakes detection across generations: Analysis of facial regions, fusion, and performance evaluation," Engineering Applications of Artificial Intelligence, vol. 110, 2022.

[36] P. Kale, "Forensic verification and detection of fake video using Deep Fake algorithm," International Journal for Research in Applied Science and Engineering Technology, vol. 9, no. 6, pp. 2789–2794, 2021.