A Comprehensive Review of Deepfake Detection Techniques

Ahmad Raza *, Abdul Basit *, Abdullah Khalid *, Asjad Amin *
*Department of Information and Communication Engineering, Faculty of Engineering and Technology,
The Islamia University of Bahawalpur, Punjab, Pakistan

***Abstract*—Deepfake technology has evolved significantly since 2017, challenging the authenticity of media and digital security. This review analyzes deepfake detection methods in three categories: deep learning, machine learning, and traditional image processing. Deep learning approaches (Transformers, Attention Mechanisms, Capsule Networks, CNNs, GANs) show potential but struggle with second-generation fakes. Machine learning methods (SVM, Random Forest, KNN, Logistic Regression) offer efficient alternatives with high accuracy in controlled settings. Traditional techniques (DCT, PFA, DFT, Edge Detection, and FTWA) identify artifacts in frequency and spatial domains. Recent innovations like HashShield (99.2%) accuracy, Dynamic Difference Learning, and FAMM framework show promise for addressing temporal inconsistencies. Head Pose Estimation effectively identifies alignment issues in manipulated expressions. Challenges include cross-dataset generalization, compression artifacts, and rapid technique evolution. Hybrid approaches and better datasets are needed to combat this growing threat.**

***Index Terms*—Deepfake, Machine Learning, Deep Learning, Forensics**

I. Introduction

Deepfake technology has significantly strengthened since its introduction in 2017. According to a well-known UK newspaper, these AI-generated manipulations have increasingly become more common over time [1].

Most current deepfake detection models face challenges when applied to real-world data. To enhance detection performance, it is recommended to utilize higher-quality, diverse datasets and to employ hybrid approaches by combining multiple deep learning models [2].

CNNs are effective at detecting first-generation fake images and videos but face significant challenges in identifying more sophisticated, second-generation fakes [3].

Deepfake Face Recognition (DFREC) is a method that can be used to detect the real source and target faces from fake images, which helps to trace who was involved in the fake. It performs better than old methods [4]. CNNs perform well in detecting fake images, audio, and videos with higher accuracy. Although better data sets and protection against attacks are needed [5].

A deepfake detection method uses frequency domain analysis with DFT and classifies data using SVM to detect deepfakes, achieving 99.76 percent precision with minimal training data [6].

Deepfakes can be detected using simple, well-designed features instead of complicated neural networks. These features perform well, are easier to understand, and also fit better into forensic processes [7].

Fake Faces In-the-Wild 10K (FFIW 10 K)is a large deepfake video dataset featuring multiple people per frame to represent real-world scenarios better. It also introduces a deep learning model that effectively detects and localizes fake faces using only video-level labels [8].

Analyzing remaining noise texture in the images can detect the deepfakes without depending on complex deep learning models [9].

The method to detect fake audio using Mel-Frequency Cepstral Coefficients (MFCCs) features merges with machine and deep learning models. This approach attains up to 88 percent accuracy and is useful in any case of language or accent by making it useful for cybersecurity and social media safety [10]. A hybrid transformer network that employs early feature fusion using XceptionNet and EfficientNet-B4 as feature extractors, trained in an end-to-end manner. The model is evaluated on the FaceForensics++ and DFDC datasets, showing comparable results to state-of-the-art techniques. Additionally, novel face cut-out augmentations and random cut-out augmentations are introduced to improve detection performance and reduce overfitting [11].

Head Pose Estimation (HPE) is the process of determining a person's head orientation in 3D space using angles yaw, pitch, and roll, often applied in facial analysis and computer vision tasks and patterns show promise for deepfake detection. This study evaluates Fine-Grained Spatial-Angular Net (FSA-Net), SynergyNet, and Web-Shaped Model (WSM) on the FaceForensics++ dataset, finding that FSA-Net, combined with K-Nearest Neighbors (KNN) and Dynamic Time Warping (DTW) classification, achieves the highest accuracy. Results indicate that HPE patterns effectively distinguish between real and fake videos, particularly by identifying alignment inconsistencies in manipulated facial expressions [12].

Dynamic Difference Learning (DDL) addresses spatio-temporal inconsistencies in deepfake videos, which often evade single-frame detection due to highly realistic frames. DDL employs a Dynamic Fine-Grained Difference Capture (DFGDC) module to mitigate normal facial movement disturbances and improve manipulation detection. Furthermore, a Multi-Scale Spatio-Temporal Aggregation (MSA) module efficiently correlates temporal differences between frames. Experiments show DDL surpasses current methods in deepfake

© 2025 TechShield Publications
DOI: https://doi.org/10.67535/tsp.000001.027
Published under the ICACNC 2025 Proceedings

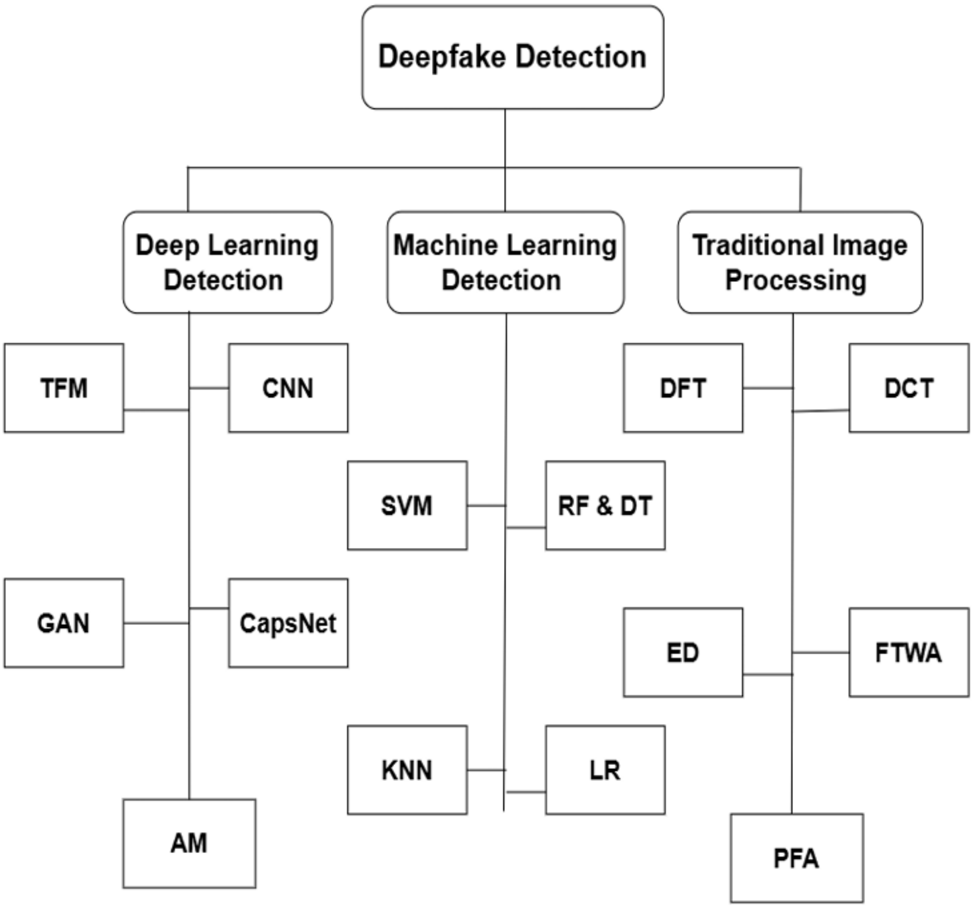

Fig. 1. **Deepfake Detection Techniques Diagram**

detection across diverse datasets [13].

The FAMM framework detects compressed deepfake videos by analyzing temporal inconsistencies in facial muscle motion. It uses geometric relationships between facial landmarks to extract motion features and employs a GRU-SVM classifier, fused via Dempster-Shafer theory, to distinguish real from fake videos. Evaluations on FaceForensics++ and real-world data from social networks demonstrate that the Facial Attribute and Motion Model (FAMM), which detects deepfakes by analyzing facial features and movements, has robust performance and superiority over existing methods, even with compression artifacts [14].

An attention-based multi-task approach integrates spatial and frequency features for detecting and localizing deepfake manipulations. Dual attention mechanisms refine spatial and channel-level features, enhancing classification and localization. Using the Fourier Transform and a U-Net-based architecture, the model achieves robust, competitive performance in both within-dataset and cross-dataset scenarios like FaceForensics++ and Celeb-DF [15].

II. Deepfake Detection Techniques

Deepfake detection has three main techniques as follows:

A. *Deep Learning Detection*

Transformers(TFM):

The paper suggests a deep convolutional Transformer (TFM) that combines CNNs with attention mechanisms using convolutional pooling and re-attention. This hybrid architecture increases generalization and attains state-of-the-art results on multiple Deepfake datasets [16]. The re-attention mechanism renews attention maps by setting up cross-head communications to increase the distinction of attention maps at different layers [17], [18], [19].

Attention Mechanisms (AM):

The network uses the Anchor-Mesh Motion Algorithm (AMM), which tracks facial movements to identify deepfake patterns, to extract temporal features and the Fusion Attention (FA) module, which combines visual and motion data, to fuse spatial and temporal features [20]. By combining attention mechanisms and multi-forensics modules, it controls the limitations of traditional CNN methods and brings out state-of-the-art performance in cross-dataset analysis [21].

Capsule Networks (CapsNet):

This paper investigates the development of DeepFakes and evaluates the effectiveness of fake detection systems using two methods: analyzing the whole face versus specific facial regions. It points out that current detection systems struggle with second-generation DeepFake databases, highlighting the need for more advanced techniques, including Capsule Networks [3]. The challenges faced by current detection methods are particularly in generalizing to new data. Emphasizes the importance of utilizing sophisticated architectures, such as Capsule Networks, to improve detection accuracy [2].

Convolutional Neural Networks(CNN):

The paper discusses DeepFakes, which are AI-generated multimedia used to mislead, and emphasizes the necessity for countermeasures. It also explores the technical developments in mitigating DeepFakes' harmful impacts, including the application of CNNs for DeepFake detection [22]. The paper proposes a method to detect face swapping by comparing the inner face region with its outer context. It utilizes two separate face recognition networks to identify discrepancies between the face and its surrounding context [23], [24], [25], [26].

Generative Adversarial Networks(GANs):

The paper surveys deepfake creation and detection methods, focusing on deep learning and forensic tools.GANs are highlighted as a key technology in generating deepfakes. The paper also discusses how deep learning and forensic tools can be combined to improve deepfake detection [27], [28].

B. Machine Learning Detection

Support Vector Machine(SVM):

The paper explores machine and deep learning techniques for identifying deepfake audio. It employs Mel-frequency cepstral coefficients (MFCCs) to extract essential information from audio samples. The study reveals that SVM outperforms other machine learning models in accuracy for certain datasets [29]. CNNs and SVMs are used for deepfake detection in a 140k face dataset, where CNNs slightly outperform SVMs [30].

Random Forest & Decision Trees (RF & DT):

The paper explores the use of classical machine learning methods, including RF and DT, for Deepfake detection. It demonstrates that these methods can achieve high accuracy in identifying Deepfakes, often outperforming deep learning approaches [31]. Challenges and Future Directions reviews existing Audio Deepfake (AD), AI-generated fake audio, detection methods, including Machine Learning (ML) and Deep Learning (DL) techniques [32].

K-Nearest Neighbors(KNN):

Range of machine learning models, including KNN, to evaluate their approach for deepfake voice detection. They extracted Mel-Frequency Cepstral Coefficients (MFCC) from audio samples and used these features to train the machine learning models [33]. Combination of CountVectorizer and machine learning models, including KNN, to classify tweets as human-generated or bot-generated. It focuses on addressing the challenge of detecting deepfake content on social media, particularly in short-form text like tweets [34].

Logistic Regression:

This study evaluates the performance of classifiers, including Logistic Regression, for identifying machine-generated social media content. Random Forest Classifier achieved the highest accuracy [34].

C. Traditional Image Processing Techniques

Discrete Cosine Transform(DCT):

The HashShield framework, an active forensic system, detects deepfakes and protects privacy by embedding robust, separable perceptual hash codes into the low-frequency domain of images using DCT. This method preserves visual quality, offers computational efficiency, and achieves over 99. 2% deep-fake detection accuracy [35]. This study investigates proactive Deepfake detection using digital watermarking, particularly frequency domain techniques like DCT. This approach enhances robustness and traceability, improving detection accuracy in combating Deepfake content [36].

Pixel And Frequency Analysis(PFA):

The HashShield framework, a forensic system for deepfake detection and privacy, embeds robust, separable perceptual hash codes into images' low-frequency domain using PFA. This method achieves high visual fidelity and computational efficiency while demonstrating over 99.2% deepfake detection accuracy [35], [37]. This paper presents a Local Frequency Analysis (LFA) method, leveraging Krawtchouk moments and DCT-based frequency analysis, to effectively detect diffusion-generated deepfakes. LFA outperforms existing methods on GenImage and DiffusionForensics datasets, exhibiting robustness to JPEG compression and Gaussian blur, and generalizing well across various generative models [38].

Discrete Fourier Transform(DFT):

This study detects deepfakes by analyzing spectral anomalies in color channels within deepfake frames using frequency domain analysis. It proposes statistical features for detection and demonstrates their high accuracy across datasets using supervised and unsupervised learning [39]. Deepfake detection achieved 99.76% accuracy by using frequency domain analysis (DFT), azimuthal averaging for feature extraction, and Support Vector Machines (SVM) for classification [6].

Edge Detection (Sobel,Canny,Laplacian)(ED):

DeepFake detection leverages edge detection methods, such as the Sobel operator, to analyze facial texture variations and identify blending artifacts, especially around the mouth [7]. Image forensics uses image feature extraction, such as edge detection, to identify tampering. This involves preprocessing steps like normalization and noise reduction, followed by analyzing features like edges, colors, and textures to detect image manipulations [40].

Fourier Transform & Wavelet Analysis(FTWA):

FTWA analyzes Deepfake images for frequency domain artifacts using wavelet decomposition and Fourier analysis. This approach extracts discriminative features that expose subtle generative traces, enabling effective identification of the GAN architectures used in their creation [41]. This paper introduces a novel deepfake detection approach using wavelet-packet analysis for multi-scale image representation. By exploiting significant wavelet coefficient differences between real and synthetic images, the method outperforms traditional pixel-space and Fourier-based techniques in deepfake identification [42].

III. CONCLUSION

This review examined deepfake detection methodologies across deep learning, machine learning, and traditional image processing techniques. Deep learning approaches show remarkable potential, with Transformers and attention mechanisms achieving state-of-the-art results. Attention Mechanisms effectively address CNN limitations through innovative approaches like Anchor-Mesh Motion algorithms. Capsule Networks emerge as promising solutions for second-generation deepfakes, while traditional CNNs require enhancement against sophisticated manipulations. GANs serve both in creation and detection when combined with forensic approaches. Machine learning methods offer efficient alternatives. SVMs excel in audio deepfake detection with MFCCs. Random Forest and Decision Trees sometimes outperform complex deep learning approaches. KNN shows promise in audio detection and social media analysis. Traditional techniques provide foundational detection mechanisms. DCT-based approaches achieve 99.2 % accuracy. PFA methods show robustness to compression and blur. DFT approaches reach 99.76 % accuracy. Edge detection effectively identifies blending artifacts. Fourier Transform and Wavelet Analysis outperform pixel-space techniques. Dynamic Difference Learning, Facial Attribute and Motion Model (FAMM) framework, and Head Pose Estimation address temporal inconsistencies effectively. Critical challenges include poor generalization, rapid technology evolution, and compression artifacts. Future research should develop hybrid approaches and better datasets.

REFERENCES

[1] I. Sample, "What are deepfakes–and how can you spot them," *The Guardian*, vol. 13, p. 2020, 2020.

[2] P. Edwards, J.-C. Nebel, D. Greenhill, and X. Liang, "A review of deepfake techniques: Architecture, detection and datasets," *IEEE Access*, 2024.

[3] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez, "Deepfakes evolution: Analysis of facial regions and fake detection performance," in *international conference on pattern recognition*. Springer, 2021, pp. 442–456.

[4] P. Yu, H. Gao, J. Fei, Z. Huang, Z. Xia, and C.-H. Chang, "Dfrec: Deepfake identity recovery based on identity-aware masked autoencoder," *arXiv preprint arXiv:2412.07260*, 2024.

[5] P. Anvekar, A. Gudnavar, K. Naregal, and S. Nagarmunoli, "Detection of manipulated multimedia in digital forensics using machine learning," in *2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*. IEEE, 2024, pp. 33–38.

[6] H. Agarwal, A. Singh *et al.*, "Deepfake detection using svm," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2021, pp. 1245–1249.

[7] 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, p. 108, 2021.

[8] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face forensics in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5778–5788.

[9] A. Méreur, A. Mallet, R. Cogranne, and M. Kuribayashi, "Forensics analysis of residual noise texture in digital images for detection of deepfake," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2025.

[10] H. H. Kilinc and F. Kaledibi, "Audio deepfake detection by using machine and deep learning," in *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. IEEE, 2023, pp. 1–5.

[11] S. A. Khan and D.-T. Dang-Nguyen, "Hybrid transformer network for deepfake detection," in *Proceedings of the 19th international conference on content-based multimedia indexing*, 2022, pp. 8–14.

[12] F. Becattini, C. Bisogni, V. Loia, C. Pero, and F. Hao, "Head pose estimation patterns as deepfake detectors," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 11, pp. 1–24, 2024.

[13] Q. Yin, W. Lu, B. Li, and J. Huang, "Dynamic difference learning with spatio–temporal correlation for deepfake video detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 4046–4058, 2023.

[14] X. Liao, Y. Wang, T. Wang, J. Hu, and X. Wu, "Famm: Facial muscle motions for detecting compressed deepfake videos over social networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7236–7251, 2023.

[15] S. Waseem, S. A. R. S. Abu-Bakar, Z. Omar, B. A. Ahmed, S. Baloch, and A. Hafeezallah, "Multi-attention-based approach for deepfake face and expression swap detection and localization," *EURASIP Journal on Image and Video Processing*, vol. 2023, no. 1, p. 14, 2023.

[16] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep convolutional pooling transformer for deepfake detection," *ACM transactions on multimedia computing, communications and applications*, vol. 19, no. 6, pp. 1–20, 2023.

[17] A. Khormali and J.-S. Yuan, "Dfdt: an end-to-end deepfake detection framework using vision transformer," *Applied Sciences*, vol. 12, no. 6, p. 2953, 2022.

[18] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Protecting celebrities from deepfake with identity consistency transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9468–9478.

[19] C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, and J. Tang, "Istvt: interpretable spatial-temporal video transformer for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1335–1348, 2023.

[20] W. Haiwei, Z. Jiantao, Z. Shile, and T. Jinyu, "Exploring spatial-temporal features for deepfake detection and localization," *arXiv preprint arXiv:2210.15872*, 2022.

[21] P. Wang, K. Liu, W. Zhou, H. Zhou, H. Liu, W. Zhang, and N. Yu, "Adt: Anti-deepfake transformer," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2899–2903.

[22] S. Lyu, "Deepfake the menace: mitigating the negative impacts of ai-generated content," *Organizational Cybersecurity Journal: Practice, Process and People*, no. ahead-of-print, 2024.

[23] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on discrepancies between faces and their context," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 10, pp. 6111–6121, 2021.

[24] 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.

[25] S. A. Shahzad, A. Hashmi, Y.-T. Peng, Y. Tsao, and H.-M. Wang, "How good is chatgpt at audiovisual deepfake detection: A comparative

study of chatgpt, ai models and human perception," *arXiv preprint arXiv:2411.09266*, 2024.
[26] Z. Lai, Y. Zhang, D. Li, and J. Ni, "Leveraging high-frequency diversified augmentation for general deepfake detection," *Journal of Information Security and Applications*, vol. 89, p. 103994, 2025.
[27] W. J. Hadi, S. M. Kadhem, and A. R. Abbas, "A survey of deepfakes in terms of deep learning and multimedia forensics," *International Journal of Electrical and Computer Engineering (Ijece)*, vol. 12, 2022.
[28] C.-C. Chang, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Cyber vaccine for deepfake immunity," *IEEE Access*, vol. 11, pp. 105 027–105 039, 2023.
[29] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via mfcc features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.
[30] J. Mallet, L. Pryor, R. Dave, and M. Vanamala, "Deepfake detection analyzing hybrid dataset utilizing cnn and svm," in *Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, 2023, pp. 7–11.
[31] M. S. Rana and A. H. Sung, "Advanced deepfake detection using machine learning algorithms: A statistical analysis and performance comparison," in *2024 7th International Conference on Information and Computer Technologies (ICICT)*. IEEE, 2024, pp. 75–81.
[32] Z. Almutairi and H. Elgibreen, "A review of modern audio deepfake detection methods: challenges and future directions," *Algorithms*, vol. 15, no. 5, p. 155, 2022.
[33] M. U. Tanveer, K. Munir, M. Amjad, A. U. Rehman, and A. Bermak, "Unmasking the fake: Machine learning approach for deepfake voice detection," *IEEE Access*, 2024.
[34] C. R. Mohan, P. J. Prabhakar, N. Karthik, P. M. Kumar, K. Rithwik, and M. M. S. Prabhas, "Enhanced deepfake detection on social media: Applying countvectorizer and random forest for optimal tweet classification."
[35] M. Yang, B. Qi, R. Ma, Y. Xian, and B. Ma, "Hashshield: A robust deepfake forensic framework with separable perceptual hashing," *IEEE Signal Processing Letters*, 2025.
[36] Z. Lai, S. Arif, C. Feng, G. Liao, and C. Wang, "Enhancing deepfake detection: Proactive forensics techniques using digital watermarking." *Computers, Materials & Continua*, vol. 82, no. 1, 2025.
[37] H. Guo, X. Wang, and S. Lyu, "Detection of real-time deepfakes in video conferencing with active probing and corneal reflection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
[38] J. Weng, "Local frequency analysis for diffusion-generated image detection," in *International Conference on Image Processing and Artificial Intelligence (ICIPAl 2024)*, vol. 13213. SPIE, 2024, pp. 66–73.
[39] M. A. Amin, Y. Hu, H. She, J. Li, Y. Guan, and M. Z. Amin, "Exposing deepfake frames through spectral analysis of color channels in frequency domain," in *2023 11th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2023, pp. 1–6.
[40] D. Kumar, R. C. Pandey, and A. K. Mishra, "A review of image features extraction techniques and their applications in image forensic," *Multimedia Tools and Applications*, pp. 1–102, 2024.
[41] L. Guarnera, O. Giudice, M. Niesner, and S. Battiato, "On the exploitation of deepfake model recognition. 2022 ieee," in *CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.
[42] M. Wolter, F. Blanke, R. Heese, and J. Garcke, "Wavelet-packets for deepfake image analysis and detection," *Machine Learning*, vol. 111, no. 11, pp. 4295–4327, 2022.