A Hybrid Approach for Medical Deepfake Detection Using Depth-Wise Convolutions in Vision Transformer and Frequency Domain Analysis

Dhanyalakshmi R[1], Alexander Zakharov[2], Natalia Romanchuk[2], Anitha J[1] and Jude hemanth[1]

[1] Department of ECE Karunya Institute of technology and science, Coimbatore, India
[2]Neurosciences Research Institute, Samara State Medical University, Samara, Russia

* corresponding author: judehementh@karunya.edu

Abstract.

These days, identifying medical deepfakes is crucial for preventing fraudulent activity, to avoid inaccurate diagnoses, as well as to uphold patient confidence. The massive increase in the production of realistic synthetic medical images presents significant challenges for clinical decision-making, highlighting the need for effective detection techniques. This proposed method offers a hybrid deepfake detection model which incorporates a lightweight Depth-Wise Convolution module in a Vision Transformer (DWConv-ViT) and a Fast Fourier Transform (FFT) module to improve feature extraction in the deepfake detection process. In contrast to conventional models which primarily use either frequency-based analysis or spatial analysis, our method integrates both feature types to increase resilience against malicious attacks. The proposed model was trained and tested using two datasets consisting of real knee X-ray images and GAN-generated osteoarthritis X-ray images. By utilizing both spatial and frequency-based details, our approach improves generalization and robustness against sophisticated deepfake approaches. Therefore, this work helps to ensure the reliability and validity of medical diagnoses.

Keywords: Medical Deepfake Detection, Deepfake Detection, Hybrid CNN-Transformer Model, Vision Transformer, Medical AI, Fast Fourier Transform, Depthwise Convolution.

1 Introduction

Deepfake is a cutting-edge technology that makes it possible to create realistic medical images, including CT scans, X-rays, MRIs, and more, in addition to creating artificial human photos, videos and so on [1]. These artificially generated images have become crucial for data augmentation [2], which protects patient privacy while allowing AI models to train on a variety of datasets. Furthermore, by producing realistic simulations of complicated procedures, medical deepfakes strengthen training in surgical procedures [3]. Deepfake-based avatars are used in telemedicine to improve

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

personalized and interactive doctor-patient communication, which ultimately makes remote consultations more accessible [4]. Additionally, the primary issue in medical research is the difficulty of collecting or obtaining data on rare diseases. Whereas, deepfake has become a lifesaver in this case [5], allowing for the creation of more trustworthy diagnostic models. Medical deepfakes have transformed the medical and healthcare sector by strengthening training methodology, diversifying data, ensuring privacy protection, and enhancing data privacy[6].

Despite its importance in the medical sector, medical deepfake carries serious concerns that could endanger patient safety and treatment [7]. Since deepfakes induce abnormalities are indistinguishable from real ones, it could mislead doctors and healthcare professionals, leading to inappropriate therapies. Therefore, Concerns regarding various possible misdiagnoses are raised by the creation of extremely realistic synthetic medical images. The threat was illustrated in a real-world scenario by a team of researchers in 2019 [8]. They used 3D conditional GAN to successfully modify CT scans of patients by adding or deleting lung cancer indications. This demonstrated that deepfake technology can bring false positives or negatives in medical diagnosis, which can lead to a life-threatening situation.

Further, these deepfakes generated medical images and scans can be exploited maliciously to create fake medical records to manipulate the diagnoses or to commit insurance fraud, which can result in monetary losses and violations of ethics. In light of these risks, it is imperative to create sophisticated techniques for identification of medical deepfakes to prevent artificial images from jeopardizing patient confidence or clinical accuracy. Research in this area must be accelerated in order to preserve the integrity of medical diagnostics and decision-making, as medical deepfakes carry similarly serious consequences to those of media deepfakes in terms of misinformation and fraud against identities. Key Focus Areas for Medical Deepfake detection are

- Maximizing Diagnostic Accuracy
- Detecting Medical Irregularities
- Safeguarding Patient Privacy.
- Improving Performance with Limited Data

The problem of identifying medical deepfakes is addressed through our study by introducing a hybrid deep learning architecture that incorporates ViT-Small [9], Depth-Wise Convolution (DWConv) [10], and Fast Fourier Transform (FFT) [11]. We demonstrated this study by training and testing the model with manipulated knee osteoarthritis X-rays images. Although all the components of a detection system play an important role, starting from preprocessing to detection, the feature extraction part majorly influences the decision-making process. So, we have exclusively designed a feature extraction module that performs well on a limited dataset by capturing global, local and spectral features.

Here, DWConv along with ViT facilitates in capturing both local texture features and global anatomical structure with minimal computational load. Additionally, the FFT branch is incorporated to investigate the spectral anomalies that are typically introduced by generative models, enhancing robustness against adversarial attacks. Unlike conventional methods that solely depend on either spectral analysis or frequency

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

analysis, our approach utilizes both frequency and spectral information to classify real and deepfake images. Crucially, the majority of medical deepfake detection models that have been developed to date require a substantial quantity of training data in order to be trained. Our approach, on the other hand, can do better with a smaller dataset.

2 Related work

2.1 General Deepfake generation

The creation of realistic deepfake data is significantly impacted by the advancements in Generative Adversarial Networks (GANs)[12], autoencoders[13], NeRFs[14], and diffusion models. These models aid in producing deepfake text data, voices, and human images (facial and entire body) [15], [16]. To execute face swapping and full body or face reenactment, these models are often trained using a large dataset that captures fine details of speech, motion, and facial expression [17],[18],[19],[20]. Deepfakes were once simple to identify with a human eye, but in the past few years, as computer hardware and software have advanced, it has become increasingly challenging for both people and machines to distinguish the difference between the real and the fake. Now it has become possible to generate deepfake with a few or one image of the target individual [21]. Furthermore, real-time deepfake synthesis, driven by efficient neural networks, has broadened its use in live streaming and interactive media. While these advancements have enhanced creativity in filmmaking, gaming, and virtual communication, they have also heightened concerns about digital deception, cybercrimes involving deepfakes, and declining trust in visual content, underscoring the importance of stringent regulations and advanced detection methods.

2.2 Medical deepfake generation

Tools for Generating Medical Deepfakes.

To generate medical deepfake, researchers utilize the most advanced deep learning models like GAN, Variational Autoencoders (VAE), diffusion model, transformers and so on. Style GAN and CycleGAN are widely used GAN models in medical image manipulation or generation. GAN-based frameworks like CycleGAN and StyleGAN can effortlessly generate or manipulate medical data like MRI, CT, X-ray, mammography, and much more since they are capable of learning the features of medical images and scans in an unsupervised manner [1]. Similar to GAN models, VAE plays a major role in medical image reconstruction and anomaly detection, which makes them useful for producing controlled modifications to artificial medical datasets. The most advanced deep learning developments like transformers and diffusion models have recently become an excellent substitute, providing high-quality medical image synthesis with reduced artifacts and enhanced feature retention [22].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

Categories of Medical Deepfakes.

Medical deepfakes fall under three major categories. First is the generation of an entirely new medical image or scan called synthetic medical image generation. Second is image-to-image translation, where models like Pix2Pix are utilized to modify the existing medical data. When there is very little data available, as in rare diseases cases, this is most frequently utilized [23]. The final technique is inpainting based modification, in which these manipulations emphasize specific modifications like adding or removing lesions, tumors, or scars. These manipulations are often made possible by attention-based technology, which ensures seamless interaction with the surrounding anatomical structures [24]. These realistic synthetic data create issues of disinformation in medical diagnosis, underscoring the necessity for careful validation and ethical measures to prevent abuse in clinical settings, even though they have promise for medical research and education.

2.3 Deepfake detection

Fernandes et al. investigated the application of Neural Ordinary Differential Equations (Neural-ODEs) for deepfake detection by estimating heart rates, revealing a notable distinction between authentic and manipulated videos[25]. Although their approach utilizes physiological signals, its effectiveness may be impacted by inconsistencies in video quality and subject movement, potentially reducing heart rate estimation accuracy. In [26] the research introduced a deep learning-based convolutional neural network designed to automatically detect diabetic retinopathy and macular edema in retinal fundus images, demonstrating strong sensitivity and specificity. However, its effectiveness depends on extensively annotated datasets and high-resolution images, which may restrict its usability across varied clinical environments. Solaiyappan et al.[1] examined medical deepfake detection using eight machine learning models, demonstrating high accuracy in detecting manipulated CT scans. However, their dependence on pre-trained models and feature extraction may restrict adaptability to emerging manipulation techniques.

In [27] the author examines the effectiveness of different YOLO models in identifying medical deepfakes within Knee Osteoarthritis X-rays and lung CT scans. The results indicate promising performance, though variations exist across datasets. While the study underscores the potential of YOLO models, inconsistencies in detection accuracy highlight the need for further refinement. In [28] the study investigates deep learning models, such as CNNs and patch-based networks, for identifying deepfake medical images, focusing on skin cancer images generated via stable diffusion. The findings demonstrate the models' effectiveness in differentiating real and synthetic images, with histogram analysis uncovering significant color distribution shifts. However, challenges remain in establishing a consistent classification threshold, and the models exhibit limitations in generalizing across datasets.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

3 Methodology

The proposed DWConv-ViT + FFT architecture ensures enhanced medical deepfake detection by capturing both local and global features through integrating depth-wise convolution module in vision transformer. The DWConv is integrated with vision transformer in a plug and play concept, without modifying any of the internal components of the transformer, including MHSA and FFN as shown in Fig. 3 with minimal computational load. Additionally, as GAN synthesized images suffer from spectral artifacts, a lightweight FFT module is incorporated in the detection network. The fusion of spatial and spectral characteristics guarantees excellent accuracy while preserving computational economy, making it lightweight, flexible, and ideal for clinical applications.

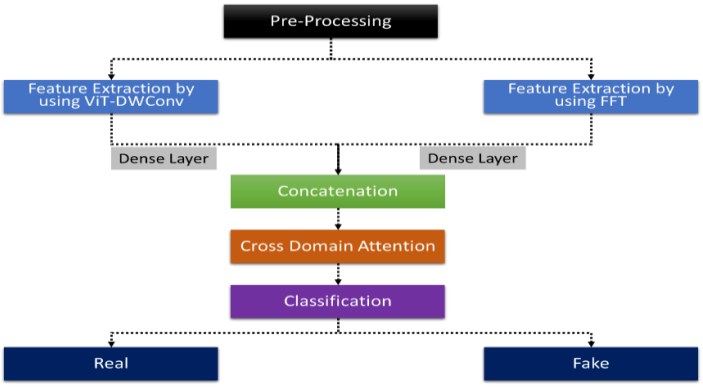

Fig. 1. Proposed Hybrid DWConv-ViT + FFT flowchart

3.1 Preprocessing

It is important to maintain the anatomical consistency of the medical image to ensure high-quality detection. By standardizing knee X-ray images via the preprocessing workflow, it is made possible. To avoid the distortion and to maintain the original aspect of the X-ray image, it was compressed to 256×256 pixels with padding. To account for real-time variation between different X-ray scanners, values of pixel intensity are standardized within a range of [0,1]. Even while performing augmentation, excessive rotation and flipping are avoided, as this could produce inaccurate and misleading medical data for the model's training. To bridge the gap between natural images and grayscale X-rays, the input X-rays are converted into pseudo-RGB images and undergo histogram matching to normalize their intensity distribution. In Section 4.1, preprocessing procedures were described in depth. Fig.1 illustrates the workflow of the proposed Hybrid DWConv-ViT + FFT model.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

3.2 Feature Extraction in DWConv-ViT Variant

The DWConv-ViT model enhances feature extraction by utilizing a pretrained Vision Transformer (DINOv2 ViT-S/14)[1], which is specifically adapted to handle the unique characteristics of medical images. To retain the characteristic of the vision transformer, the first seven transformer layers are frozen, allowing the remaining layers to be fine-tuned for capturing domain-specific details in medical images, including bone textures and subtle anatomical structures.

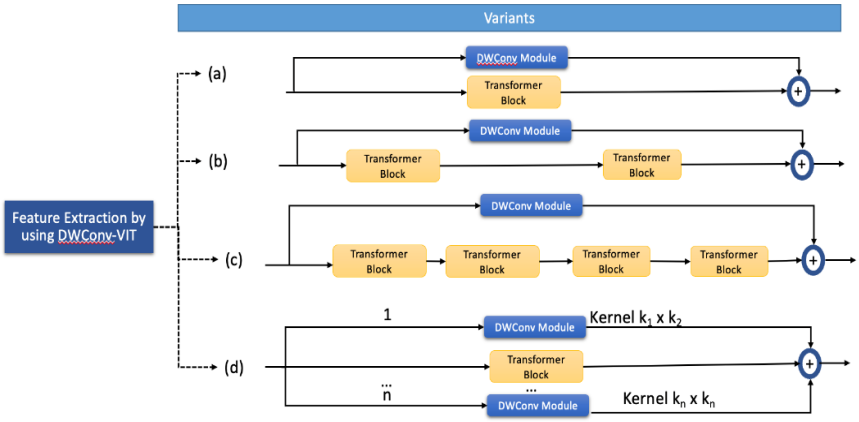

Fig. 2. Feature extraction using different variants of DWConv-VIT

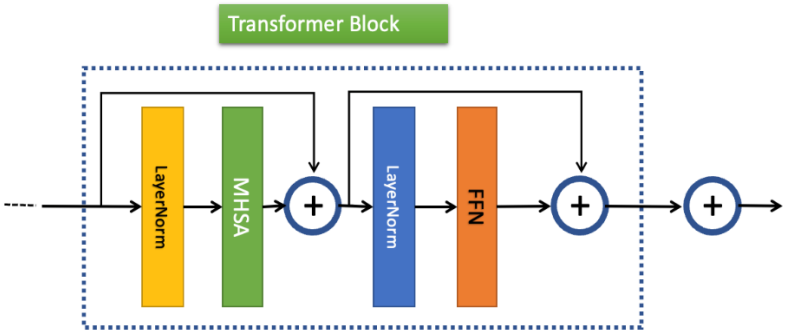

Fig. 3. Architecture of the Vision transformer used DWConv-VIT model

[1] https://dl.fbaipublicfiles.com/dinov2/dinov2_vits14/dinov2_vits14_pretrain.pth

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

DWConv-ViT improves vision transformers by incorporating depth-wise convolutions (DWConv), which enhance local spatial feature extraction while preserving the global contextual modeling of transformers. In the base variant, each transformer block is paired with a DWConv module, where 1D patch tokens are temporarily reshaped into 2D feature maps. These maps undergo 3×3 depth-wise convolution, batch normalization, and GELU activation before reintegrating with the transformer's global self-attention output. This mechanism ensures the model captures fine-grained textures like bone structures and synthetic noise, which standard self-attention may overlook. In addition to the base variant, three optimized variants shown in Fig. 2 refine this approach:

1. Base Variant: Each transformer block is individually paired with a DWConv module.

2. 2-Block Bypass: A single DWConv module serves two transformer blocks, striking a balance between parameter efficiency and sensitivity to intricate medical features.

3. 4-Block Bypass: One DWConv module spans four blocks, maximizing computational efficiency while maintaining spatial awareness, making it suitable for edge deployment.

4. Parallel Multi-Kernel: Multiple DWConv branches with different kernel sizes (e.g., 3×3 and 5×5) operate simultaneously, improving detection of both small-scale GAN artifacts and larger anatomical distortions.

All variants maintain the transformer's multi-head self-attention (MHSA) and feed-forward network (FFN) layers, ensuring seamless compatibility with pretrained weights and minimal computational overhead. In bypass variants, the shared DWConv module acts as a persistent local memory, mitigating the tendency of deep transformers to lose fine-grained details. This hybrid approach enhances global anatomical coherence while effectively detecting local synthetic artifacts, making it particularly effective for medical deepfake detection.

3.3 Feature Extraction Using Fast Fourier Transform (FFT)

The FFT-based extraction module processes medical images in the frequency domain, uncovering synthetic artifacts that might be imperceptible in spatial analysis. Applying a 2D FFT to an input X-ray produces a magnitude spectrum that highlights high-frequency patterns linked to generative models (e.g., GANs, diffusion models), such as repetitive edges, grid-like distortions, and irregular texture harmonics. This transformation shifts the image into its frequency representation, recenters the zero-frequency component, and employs logarithmic scaling to amplify subtle anomalies.

The FFT module strengthens medical deepfake detection by identifying high-frequency artifacts that spatial analysis may overlook. The process starts with applying a 2D Fast Fourier Transform (FFT) to the grayscale X-ray image, followed by generating a log-scaled magnitude spectrum to highlight subtle inconsistencies in the frequency domain. A lightweight convolutional neural network (CNN) then extracts spectral

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

features, incorporating batch normalization, GELU activation, and adaptive pooling for effective dimensionality reduction. These spectral features are then fused with spatial representations from the Vision Transformer with Depth-Wise Convolution (DWConv-ViT) through a dense transformation layer, ensuring a comprehensive and robust analysis.

3.4 Feature Fusion and classification

The fusion and classification stages combine spatial (DWConv-ViT) and spectral (FFT) features through cross-domain attention, dynamically balancing their influence based on input properties. Spatial features that capture anatomical consistency are refined by spectral features that highlight synthetic artifacts, enhancing the model's focus on medically significant patterns. The fused representation is processed through a compact dense network with Mish activations and strong regularization to minimize noise while retaining subtle synthetic markers. Focal loss emphasizes difficult cases, while weight normalization ensures stable training on limited data. The final sigmoid layer produces calibrated probabilities, boosting AUC-ROC and reducing false positives for dependable medical application.

3.5 Loss Function

The proposed loss function combines multiple elements to improve deepfake detection by tackling class imbalance, challenging samples, and overly confident predictions. At its core, it utilizes Binary Cross-Entropy (BCE) loss from Eq. (1):

$$\mathcal{L}_{\text{BCE}} = -[y \log(p) + (1-y) \log(1-p)] \quad (1)$$

where y denotes the actual class (real or synthetic) and p represents the predicted probability. To emphasize hard-to-classify cases, Focal Loss in Eq. (2) extends BCE with a modulation factor $(1\text{-pt})^{\gamma}$, giving more importance to misclassified instances:

$$\mathcal{L}_{\text{Focal}} = -\alpha_t (1-p_t)^{\gamma} \log(p_t) \quad (2)$$

where γ regulates this effect. To balance real-class distributions, KL-Grade Weighting assigns a weighted factor α_{Real}^{KL} in Eq. (3) to real samples based on their frequency:

$$\alpha_{Real}^{KL} = \alpha_{\text{Real}} \times (N_{\text{total}} / (N_{\text{KL-grade}})) \quad (3)$$

Where $N_{\text{KL-grade}}$ denotes the sample count for a specific severity level. Additionally, Label Smoothing in Eq. (4) prevents the model from making excessively confident predictions by adjusting the target labels:

$$y_{\text{smooth}} = y \times (1-\epsilon) + \frac{\epsilon}{2} \quad (4)$$

where ϵ defines the smoothing intensity. The final loss function integrates these components into a unified formulation in Eq. (5):

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{n} \alpha_t^{(i)}(1\text{-}p_t^{(i)})\gamma . \mathcal{L}_{BCE}(y_{smooth}^{(i)}, p^{(i)}) \quad (5)$$

allowing the model to effectively learn from both real and synthetic data while minimizing bias and enhancing generalization.

4 Experimental Result

4.1 Dataset

The process of constructing a clinically meaningful and artifact-rich osteoarthritis (OA) X-ray dataset starts by collecting 9,786 authentic knee X-rays from Chen's dataset[29]. All KL grades are treated uniformly and labeled as "real" to focus solely on image authenticity. A matching set of 9,786 synthetic X-rays is then randomly drawn from a pool of 320,000 GAN-generated images by Prezja et als [3], resized to 256×256 pixels, normalized, and labeled as "fake". To introduce controlled visual distortions, 30% (2,936) of these synthetic images are modified with artifacts such as grid patterns (simulating GAN upsampling flaws), Fourier-based high-frequency noise (to mimic spectral irregularities), and Gaussian blur patches over joints (replicating copy-paste errors). Augmentation strategies differ by image type. Real images undergo transformations like elastic warping, KL progression simulation, and realistic noise; synthetic ones are further modified with channel dropout, overlayed DICOM tags, and the artifact injections mentioned earlier.

To enhance variability in synthetic data, StyleGAN3 is trained on 50,000 OAI X-rays [2] to generate 4,893 new synthetic samples. These can be blended with either the original or artifact-injected Prezja images to reduce reliance on a single GAN source. Hybrid images are then synthesized by embedding 64×64 synthetic patches (e.g., knee joints) into real X-rays using seamless blending through imgaug, with the results still marked as fake. A stratified train/validation/test split is performed: real samples are grouped by KL grade, and synthetic samples by type (original, artifact-injected, or StyleGAN3-based). The dataset is divided into 6,850 real and 6,850 synthetic images for training, 1,468 of each for validation, and 1,468 of each for testing. An extra 500 synthetic images from an unseen diffusion-based model are appended to the test set to test generalization. This pipeline ensures a dataset rich in artifact variety, realistic hybrid cases, balanced class distributions, and resilience to novel synthetic image types.

4.2 Analysis

A comparison of the model variants with other models in table 1 highlights the 2-Block Bypass as the top performer, achieving the highest accuracy (91.5%), AUC-ROC (92.0%), and synthetic recall (92.5%), along with a low false positive rate (1.8%) and fast inference time (24 ms). The Base Model delivers stable results but falls short in both recall and accuracy. In contrast, the 4-Block Bypass records the lowest metrics for

[2] https://nda.nih.gov/oai

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

accuracy (84.0%) and recall (83%), though it benefits from the quickest inference speed. The Parallel Multi-Kernel variant presents a solid compromise, offering competitive AUC and recall performance, albeit with a slightly slower processing time. The 2-Block Bypass is the most effective in balancing performance, detection reliability, and speed.

Table 1. Comparison of Variants on X-Ray Dataset with different models

Variant	Acc	AUC-ROC	Synthetic Recall	FP Rate (Real)	Inference Speed
Base Model	88%	89%	86%	2.5%	28 ms
2-Block Bypass	**91.5%**	**92.0%**	**92.5%**	**1.8%**	24 ms
4-Block Bypass	84.0%	87.0%	83%	3.2%	**20 ms**
Parallel Multi-Kernel	89.5%	91.5%	89%	2.3%	29 ms
VGG19	82%	84%	78%	6.5%	35 ms
InceptionV2	85%	86%	81%	5.0%	30 ms

Table 2. Confusion Matrices for Different Model Variants (Balanced Test Set, N = 19,572)

Model Variant	TN (Real Detected)	FP (Real as Fake)	FN (Fake as Real)	TP (Fake Detected)	FP Rate (%)	Synthetic Recall (%)
Base Model	9,541	245	1,370	8,416	2.5%	86%
2-Block Bypass	9,610	176	734	9,052	1.8%	92.5%
4-Block Bypass	9,473	313	1,664	8,122	3.2%	83%
Parallel Multi-Kernel	9,561	225	1,076	8,710	2.3%	89%

As shown in Table 2 and the confusion matrices in Figure 4, the 2-Block Bypass model outperforms the others, delivering the highest synthetic recall of 92.5% and the lowest false positive rate of 1.8%, highlighting its strong suitability for clinical applications. In contrast, the 4-Block Bypass emphasizes speed but compromises on accuracy, evidenced by a much higher number of false negatives (1,664) and a lower recall rate of 83%. While the Parallel Multi-Kernel model improves artifact detection with an 89% recall, it introduces greater computational demands. Overall, the 2-Block Bypass offers

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

the most effective trade-off between performance and efficiency, making it the preferred option for detecting medical deepfakes.

Table 3. Ablation Configurations

Model Variant	AUC-ROC	Synthetic Recall	FP Rate(Real)	Grade4 Recall
Base ViT	84.2%	79%	5.1%	72%
ViT+DWConv	88.5%	85%	3.5%	80%
ViT + FFT	86.1%	82%	4.2%	75%
DWConv-ViT + FFT (2 block bypass)	92.0%	92.5%	1.8%	89%

Table 3 illustrates that the combination of Depthwise Convolution and Fast Fourier Transform markedly enhances the effectiveness of deepfake detection in medical imaging. While DWConv targets localized texture irregularities, FFT focuses on high-frequency signal anomalies. Together, they boost synthetic recall by 13.5% (rising from 79% to 92.5%) and improve AUC-ROC by 7.8% (from 84.2% to 92.0%) over the base ViT model. Additionally, incorporating KL-grade weighting significantly improves performance, raising Grade 4 recall by 17% (from 72% to 89%), thus supporting consistent detection across various osteoarthritis grades. With an excellent trade-off between precision (92.0% AUC-ROC) and a low false positive rate (1.8%), the DWConv-ViT + FFT model stands out as the most reliable option for clinical deployment

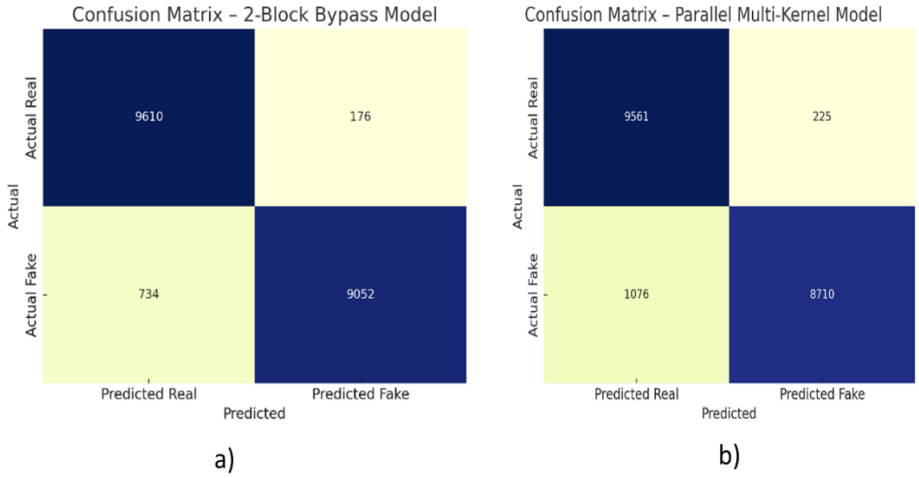

Fig. 4. Confusion Matrix: a) 2-Block bypass model and b) Parallel Multi-Kernel

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

5 Conclusion and Future Directions

The proposed DWConv-ViT+FFT-based deepfake detection model demonstrates outstanding performance in identifying medical deepfakes, owing to its ability to extract both spectral and spatial features. It effectively uncovers subtle inconsistencies that conventional methods often miss, thereby strengthening defenses against malicious manipulations. This research presents a robust hybrid deep learning architecture—DWConv-ViT combined with FFT, specifically tailored to detect synthetic osteoarthritis X-ray images, addressing the rising threat of medical image forgery. The model leverages depthwise convolutions and vision transformers to capture intricate texture patterns along with larger anatomical structures. Incorporating the Fast Fourier Transform (FFT) module further refines its sensitivity to frequency-domain artifacts commonly introduced by GAN-based generation techniques. Among various configurations evaluated, the 2-Block Bypass variant emerged as the most effective. This approach holds promise for broader application in detecting deepfakes across diverse medical imaging formats, including CT and MRI, where adversarial anomalies may differ. Additionally, integrating this model with explainable AI tools in real-time clinical environments could boost interpretability and foster greater trust in AI-assisted diagnostics.

[1] S. Solaiyappan and Y. Wen, "Machine learning based medical image deepfake detection: A comparative study," *Mach. Learn. with Appl.*, vol. 8, no. April, p. 100298, 2022, doi: 10.1016/j.mlwa.2022.100298.

[2] N. Waqas, S. I. Safie, K. A. Kadir, S. Khan, and M. H. Kaka Khel, "DEEPFAKE Image Synthesis for Data Augmentation," *IEEE Access*, vol. 10, pp. 80847–80857, 2022, doi: 10.1109/ACCESS.2022.3193668.

[3] F. Prezja, J. Paloneva, I. Pölönen, E. Niinimäki, and S. Äyrämö, "DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification," *Sci. Rep.*, vol. 12, no. 1, pp. 1–16, 2022, doi: 10.1038/s41598-022-23081-4.

[4] H. C. Yang, A. R. Rahmanti, C. W. Huang, and Y. C. Jack Li, "How Can Research on Artificial Empathy Be Enhanced by Applying Deepfakes?," *J. Med. Internet Res.*, vol. 24, no. 3, pp. 1–8, 2022, doi: 10.2196/29506.

[5] K. Falahkheirkhah *et al.*, "Deepfake Histologic Images for Enhancing Digital Pathology.," *Lab. Invest.*, vol. 103, no. 1, p. 100006, Jan. 2023, doi: 10.1016/j.labinv.2022.100006.

[6] A. S. Coyner *et al.*, "Synthetic Medical Images for Robust, Privacy-Preserving Training of Artificial Intelligence: Application to Retinopathy of Prematurity Diagnosis.," *Ophthalmol. Sci.*, vol. 2, no. 2, p. 100126, Jun. 2022, doi: 10.1016/j.xops.2022.100126.

[7] C. Stokel-Walker, "Deepfakes and doctors: How people are being fooled by social media scams," *Bmj*, 2024, doi: 10.1136/bmj.q1319.

[8] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici, "CT-GAN: malicious tampering of 3D medical imagery using deep learning," in *Proceedings of the 28th USENIX Conference on Security Symposium*, in SEC'19. USA: USENIX Association, 2019, pp. 461–478.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

[9] R. Azad *et al.*, "Advances in medical image analysis with vision Transformers: A comprehensive review," *Med. Image Anal.*, vol. 91, p. 103000, 2024, doi: https://doi.org/10.1016/j.media.2023.103000.

[10] Y. Guo, Y. Li, L. Wang, and T. Rosing, "Depthwise convolution is all you need for learning multiple visual domains," *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 8368–8375, 2019, doi: 10.1609/aaai.v33i01.33018368.

[11] J. Gao, Z. Xia, G. L. Marcialis, C. Dang, J. Dai, and X. Feng, "DeepFake detection based on high-frequency enhancement network for highly compressed content," *Expert Syst. Appl.*, vol. 249, no. March, 2024, doi: 10.1016/j.eswa.2024.123732.

[12] L. Zhang, H. Yang, T. Qiu, and L. Li, "AP-GAN: Improving Attribute Preservation in Video Face Swapping," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2226–2237, 2022, doi: 10.1109/TCSVT.2021.3089724.

[13] Z. Li *et al.*, "Identity-Aware Variational Autoencoder for Face Swapping," *IEEE Trans. Circuits Syst. Video Technol.*, vol. PP, no. July 2023, p. 1, 2024, doi: 10.1109/TCSVT.2024.3349909.

[14] X. Liu, Y. Xu, Q. Wu, H. Zhou, W. Wu, and B. Zhou, "Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, Berlin, Heidelberg: Springer-Verlag, 2022, pp. 106–125. doi: 10.1007/978-3-031-19836-6_7.

[15] T. Sha, W. Zhang, T. Shen, Z. Li, and T. Mei, "Face , Pose and Cloth Synthesis," *J. ACM*, vol. 37, no. 4, 2018.

[16] T. Li, W. Zhang, R. Song, Z. Li, and J. Liu, "PoT-GAN : Pose Transform GAN for Person Image Synthesis," *IEEE Trans. Image Process.*, vol. 30, pp. 7677–7688, 2021, doi: 10.1109/TIP.2021.3104183.

[17] Y. Nirkin, Y. Keller, and T. Hassner, "FSGANv2 : Improved Subject Agnostic Face Swapping and Reenactment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 560–575, 2023, doi: 10.1109/TPAMI.2022.3155571.

[18] X. Chen, B. Ni, Y. Liu, N. Liu, Z. Zeng, and H. Wang, "SimSwap++: Towards Faster and High-Quality Identity Swapping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 576–592, 2024, doi: 10.1109/TPAMI.2023.3307156.

[19] S. Waseem, S. A. R. S. Abu Bakar, B. A. Ahmed, Z. Omar, T. A. E. Eisa, and M. E. E. Dalam, "DeepFake on Face and Expression Swap: A Review," *IEEE Access*, vol. 11, pp. 117865–117906, 2023, doi: 10.1109/ACCESS.2023.3324403.

[20] X. Wen, M. Wang, C. Richardt, Z.-Y. Chen, and S.-M. Hu, "Photorealistic Audio-driven Video Portraits," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 12, pp. 3457–3466, 2020, doi: 10.1109/TVCG.2020.3023573.

[21] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural Voice Puppetry : Audio-driven Facial Reenactment," pp. 1–16.

[22] A. Dash, J. Ye, and G. Wang, "A Review of Generative Adversarial Networks

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

(GANs) and Its Applications in a Wide Variety of Disciplines: From Medical to Remote Sensing," *IEEE Access*, vol. 12, no. December 2023, pp. 18330–18357, 2024, doi: 10.1109/ACCESS.2023.3346273.

[23] J. S. Chen *et al.*, "Deepfakes in Ophthalmology: Applications and Realism of Synthetic Retinal Images from Generative Adversarial Networks.," *Ophthalmol. Sci.*, vol. 1, no. 4, p. 100079, Dec. 2021, doi: 10.1016/j.xops.2021.100079.

[24] C.-H. Yeh, H.-F. Yang, M.-J. Chen, and L.-W. Kang, "Image inpainting based on GAN-driven structure- and texture-aware learning with application to object removal," *Appl. Soft Comput.*, vol. 161, p. 111748, 2024, doi: https://doi.org/10.1016/j.asoc.2024.111748.

[25] S. Fernandes *et al.*, "Predicting Heart Rate Variations of Deepfake Videos using Neural ODE," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1721–1729. doi: 10.1109/ICCVW.2019.00213.

[26] V. Gulshan *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs.," *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, doi: 10.1001/jama.2016.17216.

[27] M. Karaköse, H. Yetiş, and M. Çeçen, "A New Approach for Effective Medical Deepfake Detection in Medical Images," *IEEE Access*, vol. 12, no. March, pp. 52205–52214, 2024, doi: 10.1109/ACCESS.2024.3386644.

[28] M. A. Arshed, S. Mumtaz, Ștefan C. Gherghina, N. Urooj, S. Ahmed, and C. Dewi, "A Deep Learning Model for Detecting Fake Medical Images to Mitigate Financial Insurance Fraud," *Computation*, vol. 12, no. 9, p. 173, 2024, doi: 10.3390/computation12090173.

[29] P. Chen, "Knee Osteoarthritis Severity Grading Dataset," *Mendeley Data*, 2018, doi: 10.17632/56rmx5bjcr.1.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97632-2_16