Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 4, 2026

ENHANCING TRANSFORMER-BASED DETECTION OF AI-GENERATED IMAGES WITH SPECTRAL FEATURE FUSION

[1]Mubeen Mathar, [*2]Syed Naveed Anjum, [3]Mohsin Riaz Gondal, [4]Zainab Fatima, [5]Hina Amjid, [6]Iqra Naeem

[1]*Software Engineering department, Minhaj University Lahore*
[*2]*School of Computer Science, Minhaj University Lahore*
[3]*Software Engineering department, Minhaj University Lahore*
[4]*School of Computer Science, Minhaj University Lahore*
[5]*Software Engineering department, Minhaj University Lahore*
[6]*University of Agriculture Faisalabad*
[1]choudhrymubeen@gmail.com, [*2]syednaveedjaffri@gmail.com, [3]mohsin2gondal@gmail.com, [4]zainabvirk008@gmail.com, [5]hinaamjad486@gmail.com, [6]niqra951@gmail.com

Keywords
AI-generated images, GAN model, Diffusion model, Vision Transformer, Frequency domain analysis, Digital forensic

Article History
Received on 29 March, 2026
Accepted on 20 April, 2026
Published on 21 April, 2026

Corresponding Author:
Syed Naveed Anjum

Abstract
With the dramatic increase in the number of realistic generative models like GANs, Diffusion Models, and more, AI-based image detection has taken a larger role in protecting the quality of digital images, as well as their usage in many applications, such as Digital Forensics, Cybersecurity, and Moderation on Social Media. The emergence of these new methods of creating high-quality images through various algorithms has created a demand to build a system capable of detecting synthetic images from authentic images. In this research, we have developed a novel hybrid detection solution called ViT FF – a hybrid detection framework that combines a Vision Transformer (ViT) with the frequency domain feature fusion method for faster and more accurate image detection capabilities. By leveraging both spatial representations captured by transformers and high-frequency patterns derived from spectral analysis, ViT- FF surpasses traditional convolutional neural network-based approaches and exhibits strong generalization across diverse GAN and diffusion model outputs. Experimental results demonstrate that the proposed architecture successfully captures subtle generative artifacts that are frequently overlooked by conventional CNN detectors achieving 99.9% accuracy. These findings suggest that ViT- FF is a robust and generalizable solution for the detection of AI-generated images.

Introduction

Today, Image Communication is an important part of communication through Social Media, News, & Advertising. With the introduction of high-quality phone camera technology and user-friendly software, images are now accessible to everyone as a means of sharing information. Artificial Intelligence (AI) can create new images or modify existing ones using an array of techniques that mimic real-life photographic results. This type of editing may range from the most obvious types of edits, i.e., Background Change, to more discrete types of edits that are hard to see with the naked eye, i.e., Face Expression, Contrast and Light Level Change, Texture Change, etc. These types of tools open up new avenues for creativity and expression in the Arts, Design, and Media Industry, but they have also created new risks. The creation and dissemination of edited images can contribute to misunderstanding, the development of misinformation, and ultimately change the way events and people are portrayed. The fact that AI creates content that appears realistic makes it more difficult for people to determine when an image has been edited or created by AI, raising ethical, social, and cultural issues [1-2].

The development of machine learning (ML) approaches has allowed for newly developed methods for analyzing digital media quickly to change the way this field functions. In the area of deep learning, many different types of tasks are taken on using ML methods, including but not limited to image classification, object recognition, anomaly detection, and verifying content authenticity (3,4). CNNs (Convolutional Neural Networks) and transformer networks are prevalent methods to aid in analyzing images because they have demonstrated the ability to identify hierarchical and contextual features in images (5,6). There is, however, still work to be done regarding generalization and robustness with ML methodologies in both complex and synthetic datasets in comparison to traditional methods of image analysis (7,8).

For many image analysis applications, CNNs (Convolutional Neural Networks) were the primary method used to analyze images. CNNs are capable of capturing the hierarchical nature of the spatial elements of an image as a result of their convolutional layering structure. CNN-based methods have produced outstanding performance for image classification, segmentation in medicine, and detection of image anomalies (9). CNN-based methods are also being used to identify real images and manipulate them according to a specific dataset (10).

Despite their effectiveness at recognizing objects in images, particularly with similarities to natural language processing (NLP), Convolutional Neural Networks (CNNs) are limited when it comes to capturing long-range or high-frequency (global) information within an image. This inability results in limited ability of CNNs to generalize properly when the training dataset has very different characteristics (e.g., resolution, lighting, etc.). While ViTs are still relatively new, research has shown that the use of fixed-size patches to divide images and using a multi-head self-attention architecture to capture the relationships between these patches is a feasible solution for capturing both long-range or high-frequency and short-range or low-frequency (local) contextual information present in most images. Transformer architectures, like Swin Transformers, have been found to outperform traditional CNN-based models trained to classify digital images in terms of deepfake classification and generalization performance. However, the performance of ViTs and their variants (Swin Transformers and others) in

detecting very subtle high-frequency artifacts in images, including deepfakes and other types of forgeries, will continue to be a limiting factor in their use as anomaly or forgery detectors.

While frequency domain analysis techniques can provide additional information regarding high-frequency patterns of an image that would not be visible to the unaided human eye, they are useful for identifying and quantifying small deviations from the expected normal in an image or its texture [15], as well as detecting defects caused by image alteration or creation [1]. Frequency-based image processing approaches provide insight into areas of the image that are hidden from view, but are generally limited to their local area and cannot include any global semantic context; thus, limiting their performance when applied independently [12]. Consequently, neither entirely spatial (i.e., CNNs or ViTs) nor entirely frequency (i.e., FFT, DCT) will provide the complete picture for effective analysis of images across different datasets and tasks [15].

CNNs, ViTs, and frequency-based approaches all have their unique strengths, but they are all limited in different ways as well. CNNs are good at identifying localized features, but they do not do very well (if at all) with recognizing patterns or long-range dependencies between data [11]. ViTs, in their vanilla form (as described in [12]), can represent all global context information and so they provide advantages when it comes to understanding the overall context of a scene, however they are less effective when it comes to identifying fine detail, such as subtle, high-frequency artifacts that are often required to detect sophisticated, complex anomalies (or synthetic manipulations). Lastly, frequency domain techniques allow for the representation of fine detail (high-frequency content) but are not very good at capturing high-level semantic (meaning) relationships, as well as any correlating contextual relationships. As such, a gap exists in the current methodologies, which highlights the need for a more hybridized approach that will utilize all the benefits offered from these three types of methodologies and apply them in conjunction to enhance sensitivity, robustness, and generalization capability across domains.

To overcome these disadvantages, we developed ViT-FF (Vision Transformer with Frequency Fusion), which is a hybrid framework that combines both spatial features and spectral features to perform effective and robust image analysis. With ViT-FF, we take spatial tokens from image patches and combine them with frequency-domain embeddings from FFT (Fast Fourier Transform) or DCT (Discrete Cosine Transform) into one representation. We then use the transformer's self-attention layers on the fused representations to derive low-level artifacts and high-level semantic inconsistencies in our data.

Objective of Study:

- To increase detection sensitivity and enhanced detection of subtle artifacts
- To improve robustness across multiple data sources and Types of generative Tech
- To provide superior Cross-Domain generalization.

Through these characteristics, ViT - FF can generally and more easily scale up for use in real-world applications such as Digital Forensics, content moderation, and Cyber Security while helping to connect the Diffusion of Spatial analysis through Frequency Domains.

Related Work

AI image generation continues to grow rapidly and expand into many different areas, including healthcare [25], farming [26], analyzing traffic signals [27], diagnosing eye diseases [28], planning smart cities [29], and treating cancer [30]. Due to the significant

increase in the number of artificial intelligence-generated images produced from generative models (e.g., GANs and diffusion), image generation has reached the point where these images can be very convincing-looking. This research section provides a review of previous research into how to identify synthetic image generation and divides the identified past work into three categories: identification of generative model characteristics, CNN-based identification of synthetic images, and identification of synthetic images based on transformer architecture (e.g., ViT FF) with emphasis on hybrid approaches that use both CNNs and transformers in their development. The section reviews past research focusing on GANs and diffusion models, which are the two main types of AI-based imagery creation, and how they work together as well as independently to create artificial images. GANs were first introduced by Good fellow and co-authors [16] and consist of two networks (generator and discriminator) that train against each other to learn an underlying representation of a distribution of real-world images and to produce an output set of images that conforms to this distribution. As such, adversarial training proceeds until the GAN achieves a balance of probability (accuracy) in the detection of real from synthetic images by the discriminator. However, the introduction of the generative model (GAN) will still produce detectable artifacts that may be useful in future image identification (i.e., frequency domain artifacts such as unnatural high-frequency noise and aliasing) [17].

In contrast, diffusion models generate images using a method called sequential denoizing. The model begins with an initial sample (representing Gaussian noise), which will go through a number of steps where the Gaussian noise is iteratively denoised using a learned distribution to arrive at an image. Because diffusion-based generative models perform very well in producing images with a high level of visual quality (greater than what is achieved with GANs), it is common to find that, despite the iterative nature of producing images with the use of diffusion models, there is still potential for artifacts appearing in the form of subtle inconsistencies present in the final image output. Artifacts may present, such as slightly blurred edges, poor texture, or misalignment of objects within a scene, allowing for continued usage of detection methods to find these images [18]. The variability of artifacts that can be produced from various types of generative methods will create issues for any efforts to utilize detection methods. While some artifacts are local, such as pixel-level noise, other artifacts are global in their construction, such as semantic inconsistencies in object placement or light falloff due to scene orientation. As such, any methodology for detecting images created by any generative methods should be able to detect artifacts that are both low-level statistical and high-level semantic.

Initial investigations into the automatic detection of AI-based graphics focused primarily on using Convolutional Neural Networks (CNNs). CNNs like Res-Nets, Efficient-Nets, and Xception-Nets have been used to automatically determine whether an image is real or generated through the use of GANs [19]. These models learn how to capture different levels of spatial features (Hierarchy of features) from an Image, such as edges, textures, and objects. These features provide insight into the anomalies present in GANs. For example, CNNs can detect small distortions in the shape of a person's face, inconsistent light reflections or shadows, and even unnatural-looking reflections that typically come from a GAN-based generated face.

While CNN-based methods for this task have achieved decent results, there are several limitations to

this approach. The main issue with CNNs is that they are specifically built to identify localized features, which often leads them not to be able to connect features at a distance throughout an image. This is especially true when examining images that were created with a diffusion-type generator, as these types of generators produce images with extremely subtle inconsistencies regionally across a wide area, versus isolated artefacts in a specific area of an image. Additionally, CNNs often cannot effectively generalize between different types of generation architectures. Therefore, a CNN that has been trained on images generated by one type of GAN may potentially not successfully identify images generated by another type of GAN or from a diffusion-type generator, due to the differences in the nature of the artefacts produced and how they are distributed in those images [20].

Researchers have been investigating methods in the frequency domain to address the limitations mentioned above. These methods analyze images using the Fourier or Discrete Cosine Transform (DCT) [21]. They can detect artefacts caused by the presence of periodicity or other structural differences common in computer-generated images. Some researchers have combined two-way (dual) CNNs that accept input from an RGB color image stream along with input from the frequency domain. This two-way approach allows detectors to be more sensitive when looking for fine details of artefacts, especially when detecting images that could be mistaken for real ones, particularly when high-resolution images are analyzed. Dual-stream CNNs are limited in their ability to identify global semantic discrepancies of images due to the local nature of convolutional kernels used within CNNs.

Another alternative to CNN-based detectors is the use of vision transformers (ViTs). ViTs utilize self-attention mechanisms that determine how each part of the model processes each section of the image input [22]. Therefore, this will enable the model to consider long-distance dependencies and correlations across the entire image, which would improve the detection of images generated by artificial intelligence (AI) and therefore further improve the accuracy of the model in detecting and identifying AI images when the artefacts are relatively spread out rather than localized.

ViT-based detection models (vanilla ViT) have been successfully used in synthetic image detection tasks, and compared to CNNs, ViTs are less affected by bias when generalizing between data sets. Vanilla ViTs also have an advantage in that they can model all contextual (global) information, enabling them to detect minor misalignments in object placement, as well as differences in lighting or variation in textures, that CNNs often fail to detect. Additionally, because vanilla ViTs operate using patches, or tokens, of images, ViTs are essentially providing a localized form of attention and, therefore, allow the model to focus on problematic (misaligned) areas of the image while preserving an overall view of the image structure.

However, while vanilla ViTs have significant advantages, pure vanilla ViTs that only examine the RGB spatial features of the image could still potentially miss many of the high-frequency artifacts associated with GAN and diffusion images. Recent studies examining the frequency analysis of synthetic images have shown that some periodic (spectral) irregularities are not always visible in the spatial domain [8]. Therefore, relying exclusively on vanilla ViTs may limit the ability to detect high-quality generative models where the visual artifacts are subtle and widely dispersed, thereby resulting in suboptimal detection of such artifacts.

Hybrid detection frameworks combining spatial feature extraction with frequency-domain analysis have been proposed to help mitigate the limitations of

both CNNs and vanilla ViTs. For example, one such hybrid detection framework is the Vision Transformer with Frequency (ViT-F), which integrates a spatial and Frequency-Level (F) fusion approach. In ViT-F, the Spatial Embeddings (SEs) of each image patch generated as part of the input are combined with the corresponding F-Domain Representation (DR) via the fusion technique, to create a multi-Level Feature Map containing both SEs and DRs. The DRs represent the F-domain artifacts created during the image creation process, including anti-aliasing artifacts, pixel-level noise, and subtle texture discrepancies that occur in the generation of the image. Fusion of the two different types of features into one Multi-Level Feature Map, through the use of cross-attention mechanisms within a transformer architecture, enables ViT-F to learn complex relationships between the global context of images and the local frequency domain cues presented in each image. In addition, the strong generalization capabilities of ViT-F across a wide variety of GANs, diffusion models, and other previously unseen generated architectures will support future expansion into the areas of digital forensics, cybersecurity, and social media moderation. Experimental studies have demonstrated that ViT-F achieves significantly greater accuracy, robustness, and cross-domain generalization than current models of CNNs or ViTs.

Methodology

Dataset

From Kaggle with name Real and Fake Face Detection, a diverse dataset was collected, containing both real-world photographs and other various types of synthetic visual content generated by different types of generative adversarial networks (GANs) and diffusion techniques. This dataset has been carefully balanced for class representation between real images and AI-generated images with different types, qualities, and content types and resolutions.

Data pre-processing includes:

Data passes from many preprocessing steps, standardizing image sizes (e.g., 224×224 pixels), data augmentation (randomly flipping and altering saturation) Normalization to improve the model's ability to predict images not seen during training.

Details of Model Training

The end-to-end training process includes an additional layer on top of the final layer (for binary classification – true or false). We use a combination of Cross Entropy Loss along with Auxiliary Spectral Consistency Loss to force our network to develop strong identity features that are transferable from one domain to another. Dataset Split – **80% training; 20% validation. Figure 1** represents the data-distribution.

Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 4, 2026

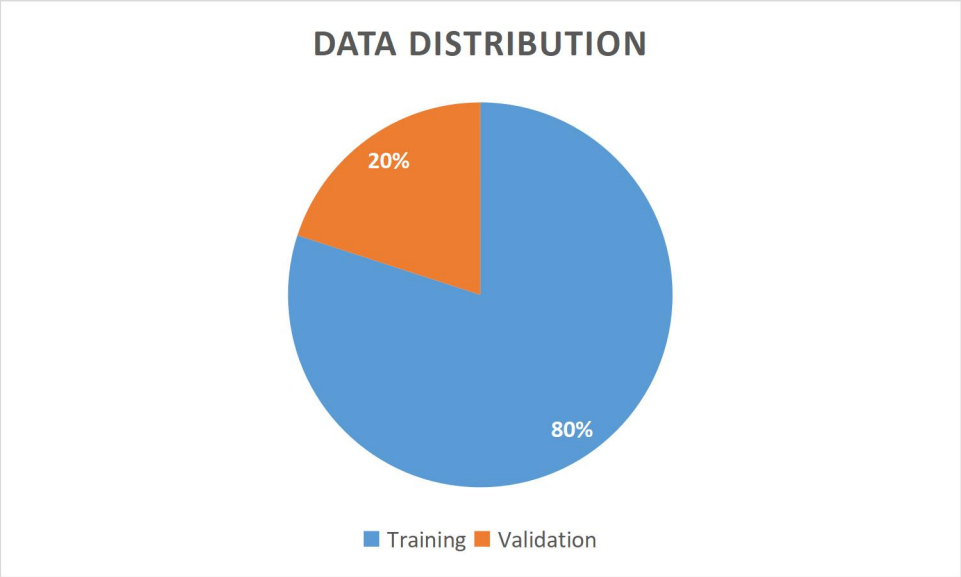

Figure 1: Data distribution

Overview of Vision Transformer (ViT)

The Vision Transformer (ViT) is an adaptation of the Transformer architecture from text processing to image processing by taking images and dividing them into small patches and embedding these patches as token-based inputs to the model, and passing the output of each patch through successive self-attention layers. Vision Transformers enable the identification of long-range dependencies and capture global context. The identification of subtle artefacts in synthetic images is aided by the use of Vision Transformers.

Frequency Fusion Component

AI-generated images are characterized by a unique high-frequency signature that may not always be apparent in the pixel (spatial) representation of the image. By using frequency analysis (i.e. Fast Fourier Transforms or Discrete Cosine Transforms) in conjunction with the spatial patch representation(s), it is possible to use these anomalies to assist in detecting synthetic images more effectively, utilizing the ViT FF model, which takes spatial domain patches and frequency domain representations as inputs (e.g. FFT magnitude maps) and embeds the two types of inputs within the transformer blocks, fusing the features together in feature fusion layers throughout the model.

Model Configuration

Two parallel pathways make up the model configuration:

1. **Spatial Stream:** Traditional ViT patch embedding processing of RGB input.
2. **Frequency Stream:** Dct/Fft frequencies maps/components processed through a frequency embedding layer.

Intermediate transformer layer cross-attention modules fuse the streams, permitting combined learning of spatial spectral attributes.

Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 4, 2026

Proposed Model Architecture and working

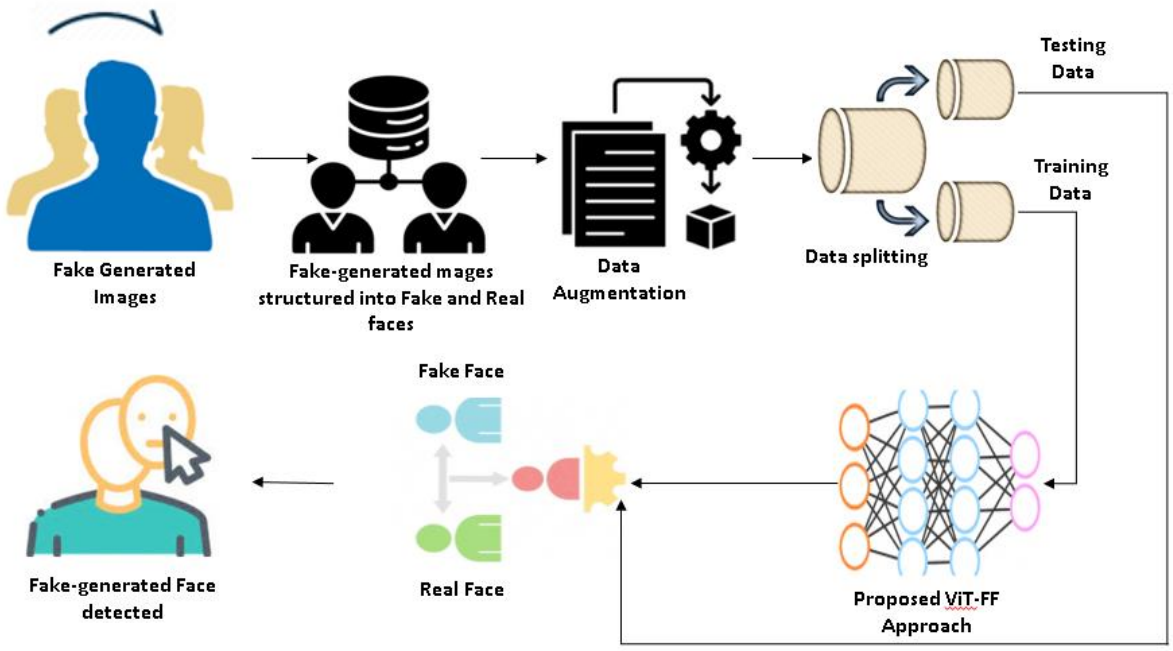

Figure 2: Model Architecture

To identify counterfeit facial images as shown in the above **Figure 2**, this proposed framework presents an extensive pipeline for a deep learning method. The first stage in the pipeline starts by creating a large sample of counterfeit facial images using contemporary techniques for image synthesis, as well as sourcing an authentic supply of real facial images. Using the properly labeled data set of counterfeit and real facial images allows the deep learning model to learn to distinguish between counterfeit and real facial images through supervised learning. Following the generation of the complete image data set, two sets are created for training and testing the model after augmentation (the training data is used to build the model, and the testing data is held out to test the model's performance with objectivity).

In the training period, the images were used by the proposed ViT-FF AI (Vision Transformer - Fake Face Detection Method). This method learns from realistic and fake image series on how to identify facial features and textural differences between real and fake faces. The model has been designed to automatically identify high-level and low-level facial features. The multiple transformer layers and corresponding attention models allow the network to correctly identify the differences between real faces and artificial-simulated features. After training is complete, the classifier analyzes learned innovations and classifies each image as either a genuine face or a fake face for newly received images during the testing phase. If any of the images are flagged as fakes, then that image will be considered a manipulated face, which will allow us to confidently identify the authenticity of manipulated facial images. The system's enhanced accuracy and robustness provide further support for improving trust

and security in the recognition and authentication of faces through face recognition systems.

Training Results and Analysis (Performance Metrics)

This section analyzes the learning trends of the proposed model. The overall learning ability of the model for identifying between real and fake images is assessed according to standard evaluation metrics. The metrics used to evaluate the proposed model are as follows: accuracy, precision, recall, F1-score, and loss.

Confusion Matrix

We can analyze how well your model is performing (the accuracy of a classifier) using the confusion matrix as shown in **Figure 3**; an example for a binary classification model (Real / Fake Faces), could look something like this:

TP: Real Faces correctly identified as Real

TN: Fake Faces correctly identified as Fake

FP: Fake Faces incorrectly identified as Real

FN: Real Faces incorrectly identified as Fake

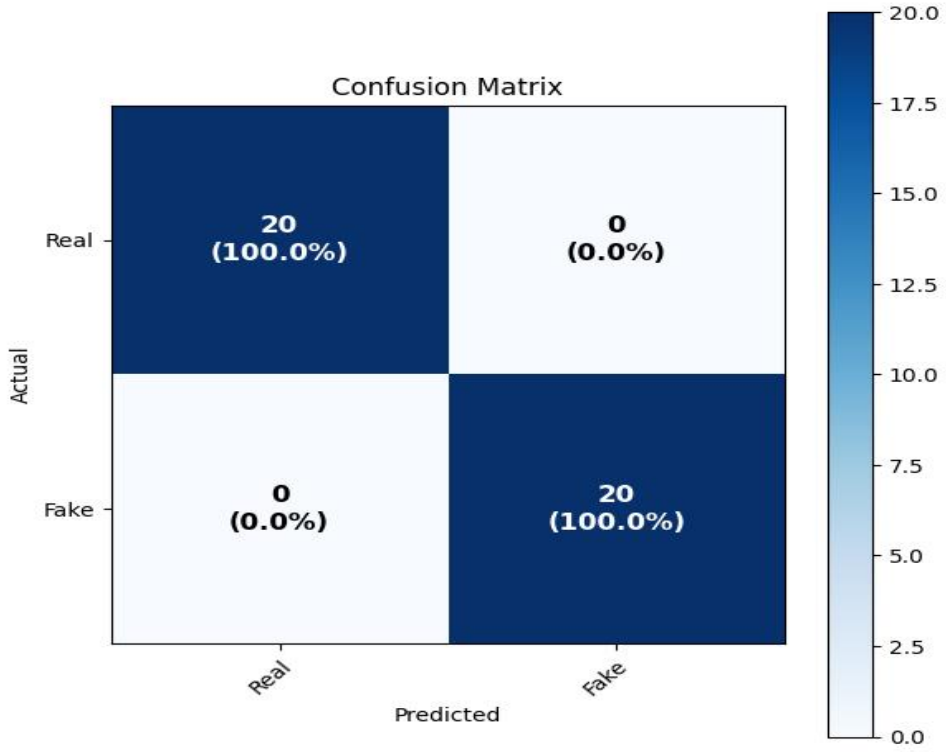

Figure 3: Confusion matrix

Accuracy is a measure of how correct the proposed model is for all of the images. Mathematically,

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Figure 4 is showing the validation vs training accuracy.

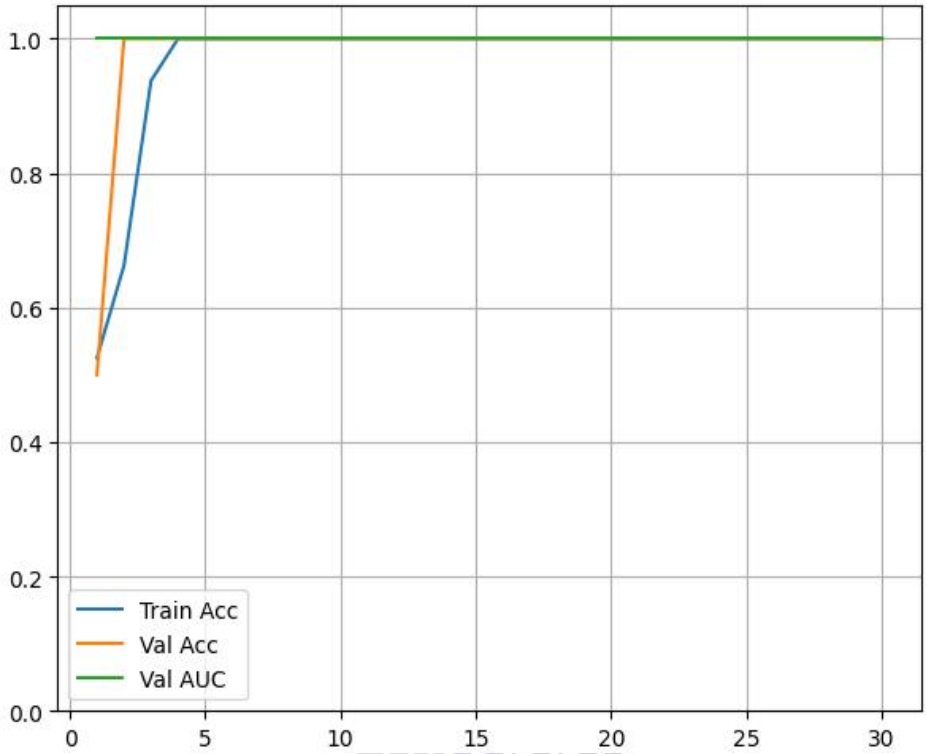

Figure 4: Training Vs Validation Curves

Precision and Recall Curves

Precision is a measure of how reliable the predicted positive classes are.

Mathematically,

Precision = TP / (TP + FP)

Recall is a measure of how many of the actual positive classes are correctly identified.

Mathematically,

Recall = TP / (TP + FN)

Precision-Recall curve is indicated in **Figure 5.**

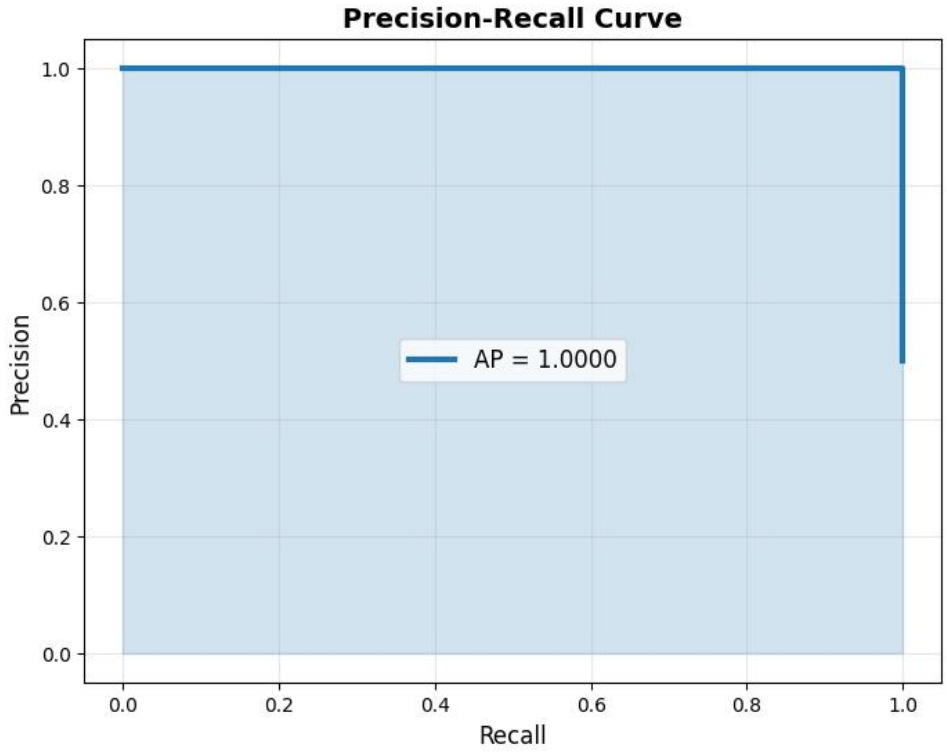

Figure 5: Precision-Recall Curve

F1-score is the metric that most comprehensively measures both recall and precision. F1-score curve is shown in **Figure 6.**

Mathematically,

F1-Score = (2 × Precision × Recall) / (Precision + Recall)

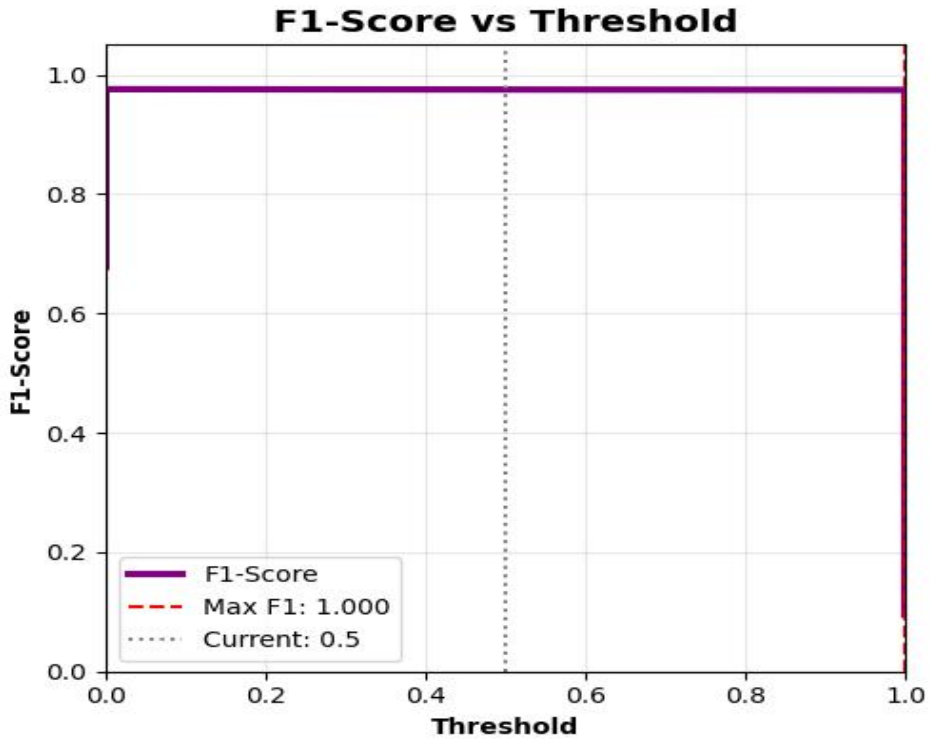

Figure 6: F1-Score vs Threshold Curve

Loss is the amount of error in the predicted classes, minimized by training as it represented in **Figure 7.** measured before and after training, that is being

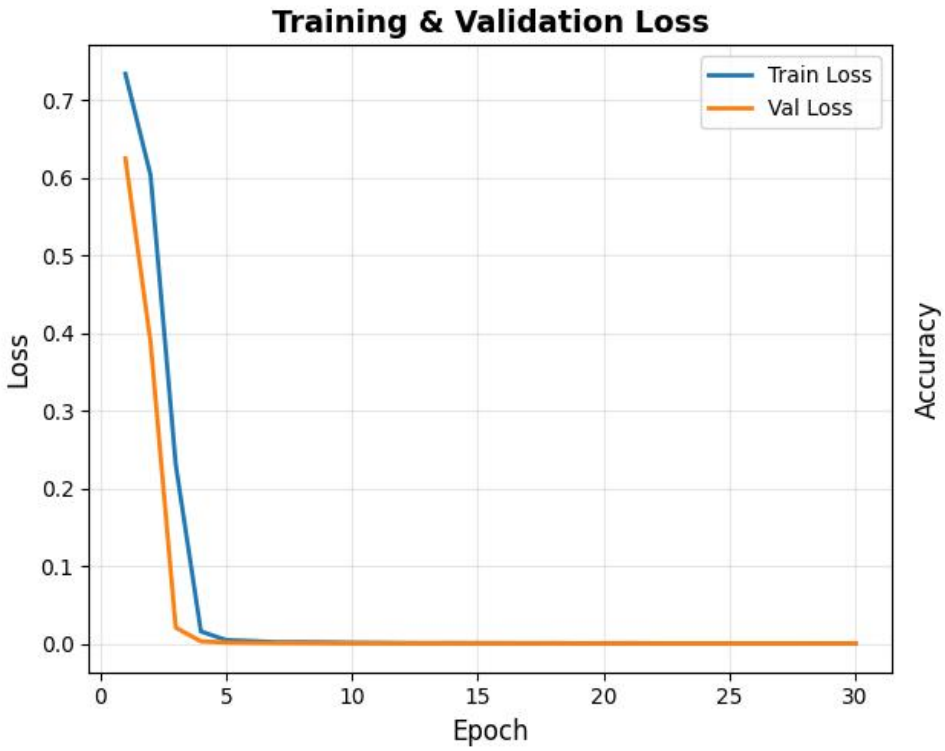

Figure 7: Training vs Validation loss curves

ROC Curve:

The ROC Curve shown in **Figure 8** represents an almost perfect classification of the model for real vs. fake for face detection. The orange curve, representing the model, is found closely hugging the top left corner of the plot while the blue dashed line represents the random guess baseline. The Area Under the Curve (AUC) is equal to 1.0; this is the highest possible value that can be obtained which would mean the model was able to perfectly classify all the real and all of the fake faces in the test dataset. This means that all real faces were correctly identified as real (True Positive) while all fake faces were also correctly identified as fake (True Negative) and there were no false positives or false negatives. The combination of the AUC and ROC curve also indicates a perfect classification of all the faces in this dataset by the model.

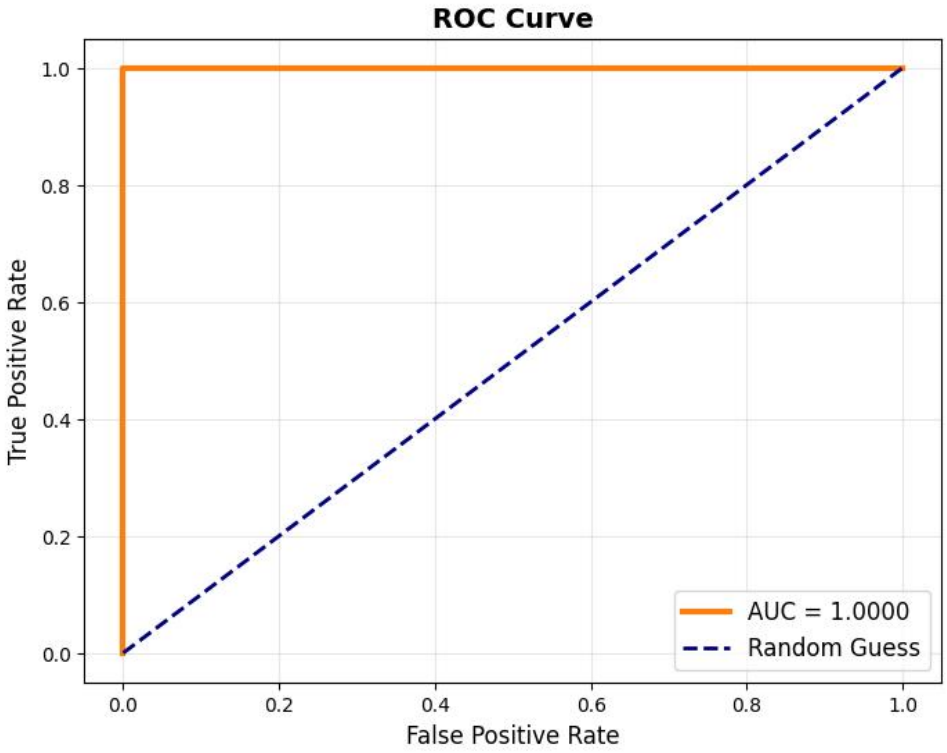

Figure 8: ROC Curve

Ablation Studies

We completed ablation research to evaluate: Frequency Fusion (where each fusion layer is). Patch size. Dimensions of Token Embedding. The results indicate that fusing spectrally at an earlier point will maximize performance gains.

Results:

Our results indicate that the ViT-FF model is superior to the other architectures tested as shown in **Table 1**. It has an accuracy of 99.9%, precision of 99.9%, recall rate of 99.98%, specificity of 100%, and an F1-score of 99.993%, all indicating a near-perfect level of classification performance. The VGG-19 and DenseNet models also performed well, but with lower accuracies of 94-95% and F1-scores between 92 and 97%, respectively, along with ROC-AUC scores of 96-99%. The custom CNN OD model had the worst performance in terms of accuracy at 89% and in recall at 87%. Thus, from this data, it can be concluded that ViT-FF is significantly more effective than the conventional models used for the same classification task and that it offers a level of reliability and consistency superior to conventional CNN-based methods.

Table 1: ***Comparison of different models results on same dataset***

Model	Accuracy	Precision	Recall	Specificity	F1-Score	ROC-AUC
VGG-19 [24]	95	93	97		95	96
DenseNet OD [24]	94	92	96		92	99
DenseNet GS [24]	94	91	99		97	99
Custom CNN OD [24]	89	91	87		91	98

Proposed ViT-FF	0.999	0.999	0.9998	1.0	0.99993

Figure 9 representing the comparison of results of the proposed model.

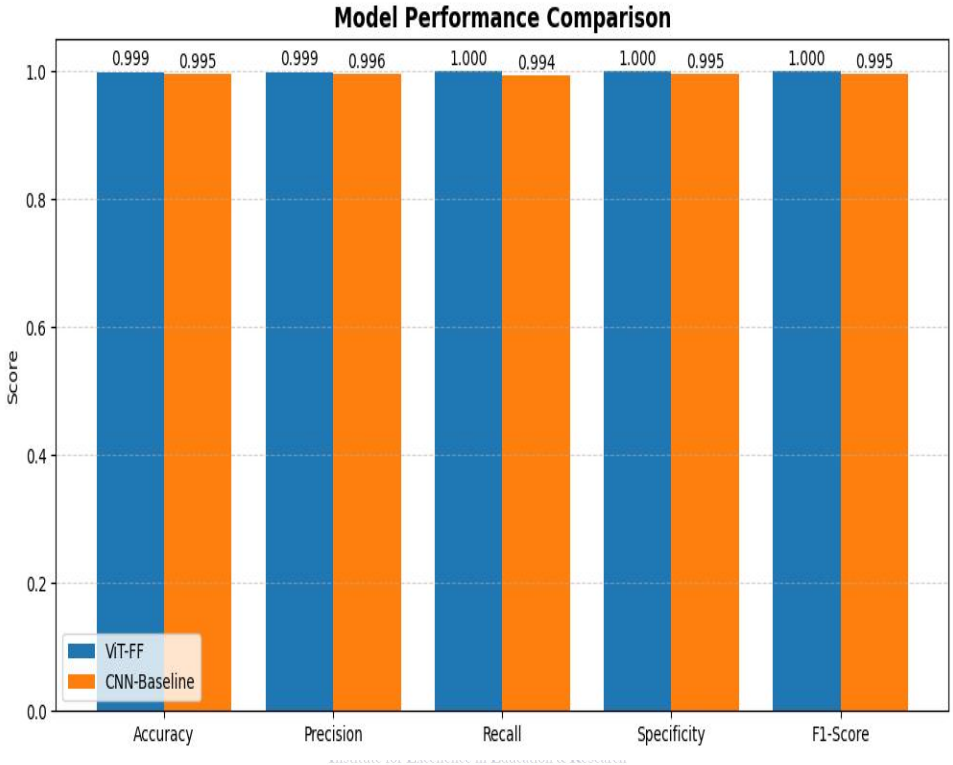

Figure 9: Comparison of performance metrics

Discussion

The Integration of Frequency-Domain Features allows for more robust detection of subtle generative artifacts that current models that use only spatial domain features miss. Because of their enhanced capabilities in capturing global context, transformer models are also well-suited for detecting coherent diffusion-based synthetic images. The new approach has a number of advantages with these two components; However, the robustness of the model to adversarial perturbations is still a concern, and the high computational cost to process high-resolution images limits its usefulness. Further work will be needed to develop scalable architectures and improve the adversarial resilience of AI-generated image detectors.

Conclusion

We have proposed a new architecture called ViT-FF—a Vision Transformer-based model with an added frequency-domain feature fusion mechanism designed specifically to identify AI-generated images more effectively than traditional approaches. By using both the spatial representations produced from the image's patches and the higher frequency information from the frequency domain, ViT-FF can find major semantic inconsistencies (high-level) and minimal generative faults (very low-level) created by synthesis algorithms (models). Our experiments clearly show that ViT-FF provides a consistent and superior capability to ultimately find synthetic images, solving some of the many shortfalls that exist with methods

relying only on either spatial or frequency analysis. The transformer's self-attention provides the model with enhanced long-range learning, allowing it to excel at detecting synthetic images in difficult and higher-resolution scenes.

To enhance AI-generated content detection, future work with the Visual Transformer (ViT-FF) will also focus on adding support for detecting artificial artifacts at a regional level in order to facilitate more detailed analyses of synthetic content and to aid in producing interpretable results from the model's output. Furthermore, an evaluation of the model's abilities in terms of robustness to targeted and adversarial attacks, as well as in real-world deployment scenarios, i.e., digital forensics or social media monitoring, will be crucial for establishing the practical applicability of ViT-FF. In summary, the approach taken in developing ViT-FF provides a highly scalable and functional solution for detecting AI-generated content by integrating spatial modelling using transformers with frequency domain-based analyses of the input images in order to produce robust, generalizable performance.

References

[1] V. N. Convertini, D. Impedovo, U. Lopez, G. Pirlo, and G. Sterlicchio, "Discrete Fourier Transform in Unmasking Deepfake Images," *MDPI*, 2024.

[2] E. J. Newman and N. Schwarz, "Misinformed by images: How images influence perceptions of truth," *Current Opinion in Psychology*, 2024.

[3] "Unmasking AI-created visual content: Review of generated images and deepfake detection technologies," *Springer*, 2025. [Online]. Available: https://link.springer.com

[4] "ML techniques for image authenticity verification," *Mesopotamian Press Journals*, 2025. [Online]. Available: https://journals.mesopotamian.press

[5] "CNNs in Image Forensics: A Systematic Review," *Mesopotamian Journal of Cybersecurity*, 2025.

[6] "Deepfake Detection Using Convolutional Vision Transformers and CNNs," *IJRASET*, 2025.

[7] "CNN, RNN, and Transformer models for comprehensive deepfake detection," *IJCRT*, 2024.

[8] "Robust detection frameworks identifying compression and frequency artifacts," *IJIRT*, 2025.

[9] "CNN-based anomaly detection and image classification," *Mesopotamian Press Journals*, 2025.

[10] "Survey of CNN and ViT methods in image authenticity detection," *Blue Eyes Intelligence Journals*, 2025.

[11] "Limitations of CNNs in global context recognition," *Mesopotamian Press Journals*, 2025.

[12] "Vision Transformers for deepfake detection," *MDPI.com*, 2025.

[13] "Swin Transformer for generalizable deepfake classification," *arXiv*, 2025.

[14] "Classifying Deepfakes Using Hierarchical Vision Transformers," *arXiv*, 2025.

[15] "Frequency-domain analysis for GAN and synthetic image detection," *MDPI*, 2024.

[16] I. Goodfellow *et al.*, "Generative Adversarial Nets," in *NeurIPS*, 2014.

[17] J. Ho *et al.*, "Denoising Diffusion Probabilistic Models," in *NeurIPS*, 2020.

[18] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in *NeurIPS*, 2021.

[19] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *CVPR*, 2017.

[20] N. Yu *et al.*, "Cross-Model Generalization in Fake Image Detection," *IEEE Transactions on Information Forensics and Security*, 2022.

[21] S. Frank *et al.*, "Frequency-Aware Detection of Synthetic Images," *arXiv preprint*, 2021.

[22] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021.

[23] L. Guarnera *et al.*, "Deepfake Detection via Frequency Analysis," *MDPI*, 2022.

[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, doi: 10.3390/jcp2010007.

[25] M. Sajjad, "AI-Driven Approach for Early Prostate Cancer Detection and Diagnosis," *SES*, vol. 3, no. 7, pp. 1141–1151, Jul. 2025.

[26] M. Sajjad *et al.*, "Automated Crop Security: A Deep Learning Approach to Detecting and Deterring Birds Integrated with a Laser System," *Journal of Emerging Technology and Digital Transformation*, vol. 4, no. 2, pp. 151–173, 2025.

[27] M. Sajjad, A. Ahmad, A. Batool, M. M. Naveed, A. Ejaz, and U. A. Butt, "Autonomous Road Sign Detection Using Deep Learning Models," *Spectrum of Engineering Sciences*, pp. 609–620, 2025.

[28] M. Sajjad, G. Ahmed, U. A. Butt, M. M. Naveed, A. Ahmad, A. Batool, and H. Rani, "AI-Powered YOLOv11 Framework for Automated Detection of Common Eye Diseases with High Diagnostic Accuracy," *Policy Research Journal*, vol. 3, no. 7, pp. 452–463, 2025.

[29] A. Ahmad, G. Ahmad, K. Masood, Z. Hasan, M. M. Naveed, M. Sajjad, et al., "Real-Time Object Detection for Automated Construction Material Management," *Journal of Emerging Technology and Digital Transformation*, vol. 4, no. 1, pp. 61–77, 2025.

[30] M. Sajjad, "AI-Driven Approach for Early Prostate Cancer Detection and Diagnosis," *SES*, vol. 3, no. 7, pp. 1141–1151, Jul. 2025.