IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5657~5666
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5657-5666

A hybrid convolutional backbone-transformer model for medical deepfake detection and adversarial defense

Keshavmurthy Shankar Krupa[1], Kiran Chandrappa[2]
[1]Department of Computer Science and Engineering (IoT and Cybersecurity including Blockchain), B.M.S College of Engineering, Affiliated to Visvesvaraya Technological University, Belagavi, India
[2]Department of Information Science and Engineering, Global Academy of Technology, Affiliated to Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Aug 7, 2025
Revised Aug 6, 2026
Accepted Aug 31, 2026

Keywords:

Adversarial defense
Convolutional neural network
Hybrid deep learning
Medical deepfake detection
Vision transformer

ABSTRACT

The advent of deep learning and artificial intelligence (AI) has now made it possible for deepfake technology to produce highly realistic fake images and videos, including fake images that may have edited medical scans. This type of manipulation may have many ramifications, including misdiagnosis, fraudulent insurance claims, as well as danger or harm to patient health and safety, and hospital operations. This paper introduces a novel hybrid deep learning framework for detecting medical deepfakes and defending against adversarial examples. The proposed architecture utilizes efficient network baseline 0 (EfficientNetB0) for extracting spatial features along with vision transformer (ViT) models for capturing the global context. Additionally, an adversarial learning approach is also considered, including stable diffusion-based forgeries, image transformations, and model-specific attacks. The proposed approach is evaluated using the computed tomography–generative adversarial network (CT-GAN) dataset under various adversarial settings. The defended model achieved 99% accuracy on test data and model-based adversarial attacks, 96% accuracy under random transformations, and 98% accuracy on synthetic data via the stable diffusion model.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Keshavmurthy Shankar Krupa
Department of Computer Science and Engineering (IoT and Cybersecurity, including Blockchain)
B.M.S College of Engineering, Affiliated to Visvesvaraya Technological University
Bull Temple Rd, Basavanagudi, Bengaluru, Karnataka 560019, India
Email: krupaks.csi@bmsce.ac.in

1. INTRODUCTION

Imaging techniques such as ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), and X-ray reveal the internal structure of the human body noninvasively and are a crucial tool in the diagnosis and treatment of diseases [1]. The advancement of artificial intelligence (AI) has revolutionized the healthcare industry and opened up a plethora of opportunities in the areas of disease detection, decision-making in clinics, treatment planning, and monitoring patients [2]. The AI solutions offer faster and sometimes even more accurate results compared to human judgments [3]. Despite these advantages of AI, there are also some issues that have emerged. One of the most worrying trends is the misuse of AI to generate deepfakes, i.e., data generated or altered by AI that mimics real data and is often unrecognizable [4]. Deepfakes began in 2014 when researchers trained a generative adversarial network (GAN) on big data [5]. As training on neural networks improved, their use grew and began to be used in areas such as celebrity fake videos, politics, and social media. More recently, deepfake technology has also found its way into the medical domain [6]. As a result, medical scans can be manipulated to conceal diseases or falsely indicate

healthy conditions, which may lead to misdiagnosis or fraudulent activities. Unlike conventional image manipulations, deepfake-generated medical images are often difficult to identify, even for experienced medical professionals [7]. Such forged images can be incorporated into hospital records, insurance claims, or medical research, posing a serious threat to healthcare systems. This may not only affect patient diagnosis and treatment but can also result in legal and financial consequences. Furthermore, conventional image forensic techniques are often insufficient because deepfakes can evade existing detection methods [8], [9], while medical images contain fine anatomical details where even minor modifications can influence clinical interpretation.

Recent state-of-the-art studies have proposed a wide range of deepfake detection methods, starting from traditional handcrafted feature-based approaches to advanced deep learning and transformer-based models [10]. Siegel *et al.* [11] proposed an explainable DeepFake detection approach based on expert-defined image features instead of deep learning. The proposed method achieved an area under the curve (AUC) of more than 0.95, although its performance may decline against newly emerging deepfake generation techniques. Sharafudeen *et al.* [12] combined a conditional GAN with vision transformers (ViTs) for DeepFake detection and reported an accuracy of 97.18%, while also highlighting the computational complexity and practical deployment challenges. Khan and Nguyen [13] presented a comprehensive study on the generalization capability of convolutional neural networks (CNNs) and transformer-based models for DeepFake detection. Raza *et al.* [14] developed a hybrid DeepFake detection model by integrating visual geometry group 16 layers (VGG16) with additional CNN layers, which achieved 94% accuracy and 95% precision. However, the model requires periodic retraining to adapt to newly developed deepfake techniques, thereby increasing computational overhead. Chen *et al.* [15] proposed a privacy-preserving framework for secure DeepFake detection in outsourced environments, where the results demonstrated improved data confidentiality without compromising detection performance. Similarly, Alsaheel *et al.* [16] developed an efficient network (EfficientNet)-based model with an attention mechanism that achieved an accuracy of 85.49% on a lung CT dataset. Also, the emergence of stable diffusion and advanced generative models has brought new risks and research opportunities. Arshed *et al.* [17] showed that both CNNs and patch-based neural networks can perform nearly perfect classification of real and synthetic medical images, but human experts lag behind automated systems. Karaköse *et al.* [18] evaluated different object detection models for identifying deepfake X-ray and CT images and highlighted that dataset-specific model tuning is necessary to obtain the best detection performance. Vivekananda *et al.* [19] demonstrated the effectiveness of EfficientNetV2 for large-scale fake and real image classification by reporting an accuracy of nearly 99.9% on a balanced dataset. Dey *et al.* [20] proposed a hybrid framework that combines image synthesis and deepfake detection, while emphasizing the importance of a comprehensive pipeline covering both image generation and verification. A style feature-based approach is designed by Padhi *et al.* [21] in SafeVision to address privacy issues, but may require more evidence before it can be generalized. Hybrid approaches that combine traditional image processing with deep learning, such as the approach by Awasthi *et al.* [22], have demonstrated improvements in imperceptibility and robustness. Sharafudeen and Vinod [23] developed an interpretable deepfake detection model for biomedical images using dual residual networks and reported high detection accuracy. Hsu *et al.* [24] investigated fast adaptation techniques, whereas Moorthy *et al.* [25] explored multimodal deepfake detection methods to address the evolving threat of deepfakes in healthcare.

According to the literature review, despite significant progress in the field of medical deepfake detection, many problems still exist. Some major gaps are as follows: i) most deep learning models achieve high accuracy but fail to generalize to novel manipulations or adversarial attacks, which limits their applicability in real-world scenarios; ii) existing research primarily focuses on single-step adversarial attacks, and few models are tested against a wider range of manipulations or imperceptible image alterations; iii) GAN-based counterfeits receive more attention, advanced techniques such as stable diffusion, which creates highly believable deepfakes are rarely addressed in current research; iv) the literature is dominated by standard CNN-based approaches, and there has been limited exploration of advanced or hybrid architectures that can enhance detection performance; and v) defensive methods against adversarial threats are often limited or ad hoc, leaving many detection models vulnerable to evolving attack techniques.

To address the above-mentioned research gap, this paper presents a novel hybrid learning model with adversarial robustness and defense. The proposed hybrid model is designed based on the joint approach of a CNN backbone for local feature learning with ViT encoder blocks to capture long-range dependencies and forgery patterns in input medical images. Another significant contribution is the custom adversarial training strategy developed to make the hybrid model resilient against both conventional and advanced image forgery attacks. The novelty of the proposed work is the integration of hybrid feature learning with a comprehensive, dynamic adversarial defense framework in a single windowing approach. Furthermore, for the proof of robustness, an extensive evaluation is done on original, stable diffusion-generated, conventionally transformed, and adversarial test sets.

2. METHOD

The primary aim of the proposed model is to accurately detect medical deepfakes and make the detection process resistant to adversarial attacks. The core module of the proposed computational model is a hybrid learning architecture that combines the spatial learning capability of EfficientNet (a CNN-based backbone) with the global attention modelling power of ViT encoder blocks, which enables the model to detect fine-grained manipulations and global inconsistencies in medical images. To enhance robustness, adversarial training is performed where the model is continuously exposed to clean, transformed, synthetic, and adversarial images during training, which ensures the model learns to generalize across multiple test-time threats. Figure 1 illustrates the complete system design and methodology workflow of the proposed system.

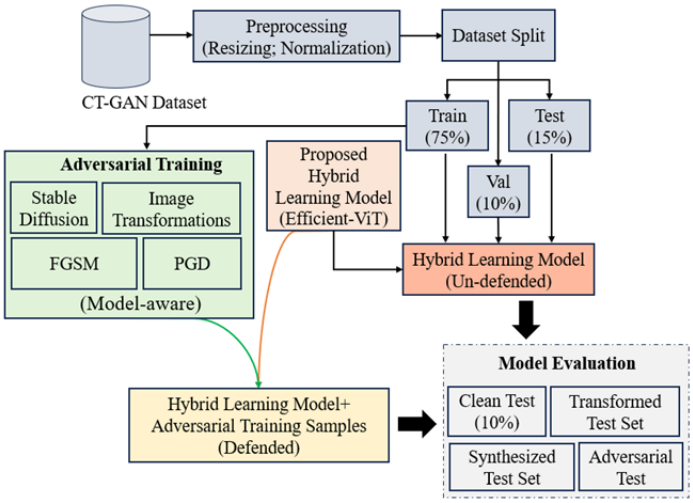


Figure 1. Illustrates the workflow of the hybrid-learning efficient-ViT models with adversarial defense

The proposed system illustrated in the above Figure 1 considers acquiring and preprocessing CT scan data from the computed tomography–generative adversarial network (CT-GAN) medical deepfake dataset. The preprocessing module of the proposed system performs image resizing to 224 × 224 pixels and normalizes pixel values to the range [0, 1] to ensure that the dataset is consistent and compatible with the proposed hybrid learning model. Afterwards, the dataset is split into training, validation, and multiple test subsets. The training data is passed through the adversarial training pipeline, where stable diffusion, image transformations (blur, JPEG compression, and brightness adjustment), and model-aware adversarial attacks such as fast gradient sign method (FGSM) and projected gradient descent (PGD) attacks generate adversarially enriched samples, which are then used to train the defender model (i.e., model with adversarial training). For comparison, an un-defended model (i.e., a model without adversarial training) is also trained using only original training data. This comprehensive evaluation ensures that the model is not only effective under standard conditions but also reliable under adversarial and manipulated scenarios.

2.1. Dataset

This study employs the deepfakes CT image dataset introduced in the CT-GAN study and obtained from a publicly available UC Irvine machine learning (UCI ML) repository [26]. The dataset contains high-resolution 3D CT scans of human lungs in digital imaging and communications in medicine (DICOM) format, consisting of 100–300 slices of 512 × 512 pixels. This dataset highlights the vulnerability of medical imaging to realistic manipulations, where tampering was done using the CT-GAN model by either removal or insertion of lung cancer nodules, with annotations provided at voxel-level coordinates for four classes such as true-benign (TB), true-malicious (TM), false-benign (FB) and false-malicious (FM). One of the most compelling aspects of this dataset is its realism and difficulty. In a blind evaluation trial, three expert radiologists were unable to reliably distinguish real from fake scans, which shows the importance of automated detection systems for deepfake detection models in medical imaging. For the purpose of development and evaluation of the proposed model, the study curated a subset of the dataset consisting of 3,600 training images, 400 validation images, and 800 test images, distributed evenly across the four classes.

2.2. Proposed hybrid learning model

The proposed hybrid learning Efficient-ViT model is designed based on the integration of the EfficientNetB0 CNN backbone [27] with ViT [28] encoder blocks to combine local spatial detail extraction

with global context modelling. The architecture of the proposed predictive model is demonstrated in Figure 2, exhibiting its internal transformer encoding structure and the attention mechanism. Figure 2(a) presents the complete hybrid Efficient-ViT architecture, where Figure 2(b) shows the transformer encoder structure, Figure 2(c) illustrates the multi-head attention mechanism, and Figure 2(d) depicts the self-attention operation used to capture long-range dependencies. The EfficientNetB0 processes an input CT image (224 × 224 × 3) into a high-level feature map of shape (7 × 7 × 1280), which is projected to (7 × 7 × 256) via a convolution operation with a (1 × 1) kernel, followed by a patch embedding process where the feature map is transformed into a sequence of fixed-size tokens, i.e., 49 tokens of dimension 256 each. Here, the extraction of patch is required because the feature map obtained from EfficientNet cannot be directly processed with a self-attention mechanism, and since transformers operate on sequential inputs rather than 2D grids, so the system has used a convolution process to project the high-level feature map to a lower dimension, and in this way, it preserves both local spatial information and global positional relationships (via positional embeddings in the transformer).

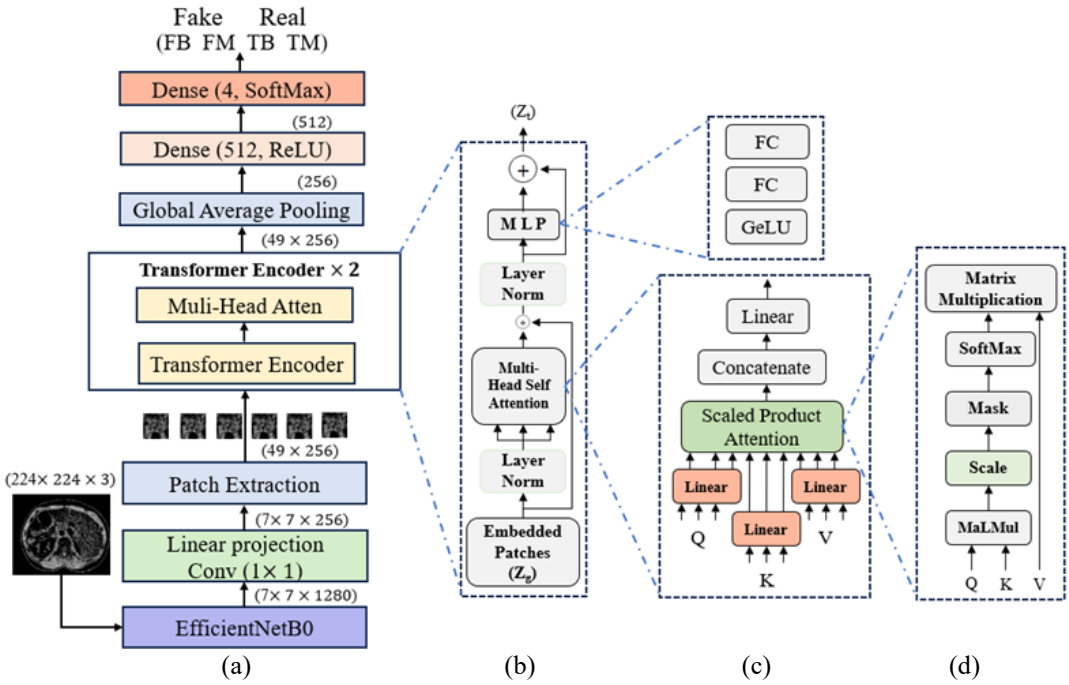


Figure 2. Illustrates the architecture of the proposed hybrid learning model Efficient-ViT: (a) the main architecture of the proposed efficient-ViT model, (b) transformer encode architecture, (c) multi-head attention layer, and (d) self-attention model

The extracted patches Z_g are then fed into a transformer encoder, which includes a multi-head self-attention (MHSA) layer, normalization, and a position-wise feed-forward network multilayer perceptron (MLP) to capture long-range dependencies across spatial regions. In the proposed model, the study has considered two transformer encoder blocks to achieve a balance optimal between model complexity and feature representation power, as a single block may not provide enough multi-layer attention refinement to capture complex inter-patch relationships. The MHSA mechanism splits the input into multiple attention heads, each learning distinct feature subspaces where attention scores such as query (Q), key (K), and value (V) vectors are obtained through separate linear projections. Further, within each attention head, the attention scores are scaled and masked to highlight the important features, followed by the softmax function before it is multiplied with the original feature to produce the output, i.e., a recalibrated feature map. The obtained output is then globally pooled and passed through fully connected layers with dropout before a final softmax classifier predicts one of four classes. This hybrid architecture enables the network to learn both fine-grained textures and semantic inconsistencies in deepfake medical images. The model is trained for 30 epochs using the adaptive moment estimation (Adam optimizer) with cross-entropy loss, a custom learning rate scheduler (factor = 0.5, min-lr = 1e-6), and early stopping to prevent overfitting. Table 1 presents the hyperparameters and configuration details of the proposed model.

Table 1. Model configuration and training hyperparameters

Component	Details
Backbone CNN	EfficientNetB0 (ImageNet pretrained)
Input size	224 × 224 × 3
CNN output shape	7 × 7 × 1280
Patch conversion	Conv2D (1×1, 256 filters) + rearrange to (49, 256)
Transformer blocks	2 blocks, each with 4 heads and FFN size 128
Pooling	GlobalAveragePooling1D
Classifier head	Dense (512 ReLU) → dropout (0.2) → dense (4 softmax)
Loss function	Categorical crossentropy
Optimizer	Adam
Learning rate schedule	Custom callback (factor = 0.5, min_lr = 1e-6)
Batch size	As per training generator (typically 16–32)
Epochs	30 (with early stopping)
Metrics	Accuracy
Regularization	Dropout + batch normalization (within callbacks)

2.3. Adversarial training

The proposed study employs an adversarial training strategy to enhance the robustness of the hybrid learning Efficient-ViT model against both data-level perturbations and model-aware adversarial examples. The goal is to create a defender model that can resist dynamic attacks and manipulations without sacrificing detection accuracy on clean or naturally corrupted images. The implementation details of the adversarial training strategy adopted for making the model defensive are described in Algorithm 1.

Algorithm 1. Iterative adversarial training strategy

Input: D_{clean} (clean training set), M_θ (Hybrid model parameters θ), E (number of epochs), ε_{FGSM} & ε_{PGD} perturbation parameters, α_{PGD} (PGD step size), T_{PGD} (PGD iterations)
Output: M_{θ^*} (Adversarially robust trained model)
Start:

1: Initialize $\theta \leftarrow \theta_0$
2: **For** e = 1 to E do:
$D_{SD} \leftarrow f_1(D_{clean})$ // synthetic sample generation using stable diffusion generator
$D_{trans} \leftarrow f_2(D_{clean})$ // Image Transformation
3: Model-aware iterative adversarial generation
$D_{FGSM} \leftarrow f_3(D_{clean}, \theta) = D_{clean} + \varepsilon_{FGSM} \times \text{sign}(\nabla_x \mathcal{L}(M_\theta(x), y))$ // FSGM attack
$D_{PGD} \leftarrow f_4(D_{clean}, \theta) = \Pi_{B_{\epsilon(x)}}\left[x + \sum_{t=1}^{T_{PGD}} \alpha_{PGD} \times sign\left(\nabla_x \mathcal{L}(M_\theta(x_t), y)\right)\right]$ // PGD attacks
4: Merge adversarial and original train set
$D_{epoch} = D_{clean} \cup D_{SD} \cup D_{trans} \cup D_{FGSM} \cup D_{PGD}$
5: Training Update:
$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \frac{1}{|D_{epoch}|} \sum_{(x,y) \in D_{epoch}} \mathcal{L}(M_\theta(x), y)$
6: **End** of **For**
7: Return output: M_{θ^*}
End

Algorithm 1 considers various input variables, and after successful execution, it returns an adversarially robust trained model. The first step of the algorithm is executed towards initialization of the learnable parameter, i.e., weight (θ) of the proposed Efficient-ViT model. Then, for each epoch (E = 30), the algorithm initiates a process of generating adversarial training samples followed by three core operations such as synthetic data generation, image transformation, and model-aware adversarial attack generation. The obtained adversarial train sample is then merged with the clean training set to form D_{epoch} which is used for model parameter updates during its training phase. This iterative process creates multiple manipulation types and ensures that the model incrementally learns to recognise and resist both semantic-level and pixel-level perturbations. The first operation is carried out towards synthesizing the training samples from D_{clean} using function $f_1()$, which generates highly realistic images that mimic the type of manipulations the deepfake attackers often produce in real-world scenarios, which are more realistic and harder-to-detect fakes. The function $f_1()$ basically uses stable diffusion [29], which is a text-to-image generation AI model, and unlike many GANs, which directly operate in the pixel space, it utilizes a latent diffusion approach that encodes the image into a lower-dimensional latent space, applies the generative process there, and then decodes it back into the image space. In Algorithm 1, the function $f_1()$ is formulated as $z_o = \mathcal{E}(D_{clean})$, where $\mathcal{E}()$ denotes the encoder module, which produces an encoded latent representation of the input images as it operates in a latent space rather than directly on pixel values. The encoded latent representation z_o is progressively perturbed over T timesteps by adding Gaussian noise, as shown in (1) and a closed-form sample of the noised latent variable at timestep t is derived from the initial latent code z_o given in (2).

$$q(z_t|z_t-1) = N\left(z_t; \sqrt{1-\beta_t} \times z_{t-1}, \beta_t I\right) \quad (1)$$

$$z_t = \sqrt{\bar{\alpha}_t} \times z_0 + \sqrt{1-\bar{\alpha}_t} \times \epsilon,\ \epsilon \sim N(0, I) \quad (2)$$

Where, β_t is the variance schedule, $\bar{\alpha}_t = \prod_{i=1}^{t} 1-\beta_t$, and ϵ denotes the noise vector. This process accumulates the noise schedule over all steps and accelerates both training and inference in latent diffusion models by allowing sampling directly at any timestep without simulating all intermediate steps. Afterwards, noise reversal is done using a U-Net-based denoising function $\epsilon_0(\,)$ is applied to z_t to learn and capture noisy patterns at each step, as shown in (3), followed by mean value μ_θ prediction as expressed in (4).

$$p_\theta(z_t - 1|z_t) = N(z_{t-1}; \mu_\theta(z_t, t, y), \sum_\theta z_t, t, y) \quad (3)$$

$$\mu_\theta(z_t, t, y) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(z_t, t, y)\right) \quad (4)$$

Where y, is the conditioning prompt that influences the denoising through cross-attention, thereby enabling targeted forgery synthesis towards deepfake manipulations. In Algorithm 1, the conditioning prompts are customised to medical forgery contexts such as "*injecting artificial cancer nodules*", "*removing actual cancer nodules*" and "*modify latent anatomical textures to appear realistic but incorrect*". Once the denoising process reaches $\hat{z}_0$ the decoder $D(\,)$ reconstructs the image with forgery that maintains high visual quality like the original and is harder to detect. The second operation in the adversarial training process is carried out using the function $f_2(\,)$, which performs different image transformations such as Gaussian blurring, JPEG compression, and brightness adjustment to simulate common distortions and expose the model to non-learned perturbations that an attacker may use to intentionally compromise or degrade the diagnostic details of the CT scan. The third operation is carried out towards performing adversarial attack using the function $f_3(\,)$, which uses the FGSM technique [30] to generate single-step adversarial examples. For each input X in D_{clean} the perturbed sample is computed as shown in Algorithm 1 step (3). Similarly, to generate stronger multi-step perturbations, the study considered the PGD attack [31] technique implemented using function $f_4(\,)$ which perturbs the images by applying iterative updates with projection constraints, and introduces a stronger manipulation that can mislead even the most accurate models. The implementation of both FGSM and PGD is done using a Python library named CleverHans. The epsilon values for both FGSM and PGD are ε_{FGSM} = 0.007 and ε_{PGD} = 0.005, which control the maximum allowable change to each pixel and are chosen based on the empirical analysis to ensure that the perturbations remain effective and undetectable to human observers.

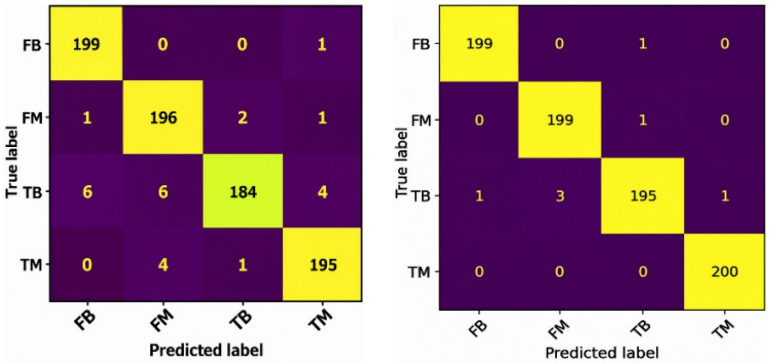
3. RESULTS AND DISCUSSION

This section presents the experimental outcomes and assesses the performance of the proposed hybrid Efficient-ViT model considering extensive experiments under multiple test scenarios. All experiments were carried out using the Python programming language executed within the Anaconda environment on Windows 11 with 16 GB RAM, and an NVIDIA GPU with 6 GB VRAM to accelerate deep learning training. The study considers two different test cases, i.e., with the original testcase an another with adversarial scenarios, where the performance of the model is evaluated using different metrics such as accuracy, F1-score, precision, recall, receiver operating characteristic curve (ROC)-AUC, and confusion matrix.

3.1. Evaluation on original test set

In this evaluation phase, both the undefended model (without adversarial training) and the adversarially trained defended model are evaluated on the unmodified test set. Figure 3 presents the confusion matrix analysis for both models, where Figures 3(a) and 3(b) represent the undefended and defended models, respectively. Both confusion matrices have dimensions of 4 × 4, corresponding to the four classes FB, FM, TB, and TM, with 800 test samples distributed equally across the classes. The rows indicate the true labels, whereas the columns indicate the predicted labels. It can be seen that both have achieved high classification performance for all classes FB, FM, TB, TM, but the defended model exhibits slightly better improvements in true positives for most of the classes. As shown in Figure 3(a), a closer analysis of performance for the undefended model reveals that the most of the misclassifications have taken place in the TB class, with 6 samples of TB being incorrectly predicted as FB. Besides, there were 2 misclassifications for the FM class (the samples were predicted as TB) and 4 misclassifications for the TM class (the samples were predicted as FM). On the other hand, Figure 3(b) shows that the defended model demonstrated better classification accuracy with a significant decrease in the number of misclassifications (from 10 for the undefended model to only 4) and the absence of misclassifications in the TM class. The analysis suggests that

the adversarial training approach not only provides a higher level of robustness against adversarial attacks but also improves the accuracy of classifications. This improvement is further supported by the quantitative results presented in Table 2.



(a) (b)

Figure 3. Confusion matrix of the proposed hybrid Efficient-ViT model: (a) un-defended and (b) defended deepfake detection on the original test set (4 × 4 matrices representing four classes and 800 test samples)

Table 2. Comparison of classification performance

Class	Un-defended			Defended			Support
	Precision	Recall	F1	Precision	Recall	F1	
FB	0.97	0.97	0.98	0.99	0.99	0.99	200
FM	0.95	0.95	0.97	0.99	0.99	0.99	200
TB	0.98	0.98	0.95	0.99	0.97	0.98	200
TM	0.97	0.97	0.97	1.00	1.00	1.00	200
Overall accuracy (OA)	97%			99%			

The quantitative analysis presented in Table 2 demonstrated that the un-defended model achieved an OA of 97%, whereas the defended model achieved 99%. Additionally, the defended model has achieved perfect classification for the TM class, with 100% for both recall and precision rate. Also, a significant improvement can be seen in the F1-score, with 99%, further confirming the effectiveness of the defended hybrid efficient-ViT with adversarial training.

3.2. Evaluation on adversarial test scenario

In the phase of evaluation, both un-defended and defended models are evaluated on adversarial test samples generated using the same operation of Algorithm 1 on the test data, where test samples were altered using via stable diffusion, image-level transformations (Gaussian blur, JPEG compression, and brightness adjustment), and model-aware adversarial attacks (FGSM and PGD). This second test case is considered to assess the robustness of both models against diverse and realistic attack scenarios. Table 3 presents the comparative performance of both models in terms of OA and F1-score under different and dynamic adversarial test scenarios.

Table 3. Comparison of classification performance under adversarial scenarios

Test scenario	Un-defended		Defended	
	Accuracy	F1	Accuracy	F1
Clean test set	0.97	0.97	0.99	0.99
Random transformations	0.72	0.73	0.96	0.96
Synthetic test set	0.61	0.41	0.98	0.97
FGSM adversarial test set	0.65	0.64	0.99	0.99
PGD adversarial test set	0.52	0.45	0.99	0.99

The results in Table 3 demonstrate a clear performance gap between the un-defended and defended models under all adversarial test scenarios. The un-defended model suffers from accuracy and F1-score drops, more under PGD attacks. On the other hand, the defended model achieved 99% accuracy for all conditions, indicating that the proposed adversarial training strategy effectively equips the model to withstand diverse and high-stress perturbations without compromising detection accuracy. Random

transformation involves the use of JPEG compression, Gaussian blur, and brightness adjustment, where the transformation parameters are carefully chosen in a way that makes them different from the training phase in order to make the test condition dynamic and strong. In the case of synthetic test data, the generated fake classes (FB and FM) are done using stable diffusion, whereas the real TB and TM are left the same. The FGSM and PGD attacks are conducted under the same perturbation parameters used in section 2.3. The next analysis in Figures 4 and 5 is shown with respect to varying epsilon values in both FSGM and PGD attacks.

The performance analysis reported in Figure 4, reveals that different perturbation budgets (ϵ) create a different effect on the performance of the defended model when subjected to the FGSM attack. When perturbations are small ($\epsilon \leq 0.00588$), which means that the generated perturbations are nearly imperceptible, all metrics' values stay high and almost unchanged and equal to 0.99. In case of increasing of the perturbation budget up to moderate ($0.00784 \leq \epsilon \leq 0.0156$), there is a slight deterioration of the defended model performance. The greatest deterioration of the defended model performance is observed when the largest perturbation budget is applied ($\epsilon = 0.03137$), with accuracy and F1-score being equal to 92%.

The performance of the defended model under the PGD attack at varying degrees of perturbation is shown in Figure 5. It can be seen that for small to moderate ϵ values ($\epsilon \leq 0.01000$), all evaluation measures are maintained around 0.99 showing resistance towards iterative adversarial perturbations. At the maximum tested level of perturbation ($\epsilon = 0.01500$), there is a little degradation in the performance with both accuracy and F1-score being 0.9825. This analysis shows that the hybrid Efficient-ViT defended architecture is robust and reliable for use in detecting deepfakes in the medical field under adversarial conditions. The robustness of the presented model is further evaluated by comparing its performance with another existing research on the same CT-GAN dataset. Existing research on the CT-GAN dataset reports an accuracy of 91.57% for deep learning-based 3D CNN models [32] and 85.49% for EfficientNet-based attention models [33], showing that deep learning methods are effective in detecting deepfakes in medical applications. However, these techniques mainly dependent on spatial or temporal feature learning and do not assess robustness against adversarial attacks. On the other hand, the proposed hybrid defensive model has achieved an accuracy of 99% over transformed, synthetic, and adversarial test scenarios.

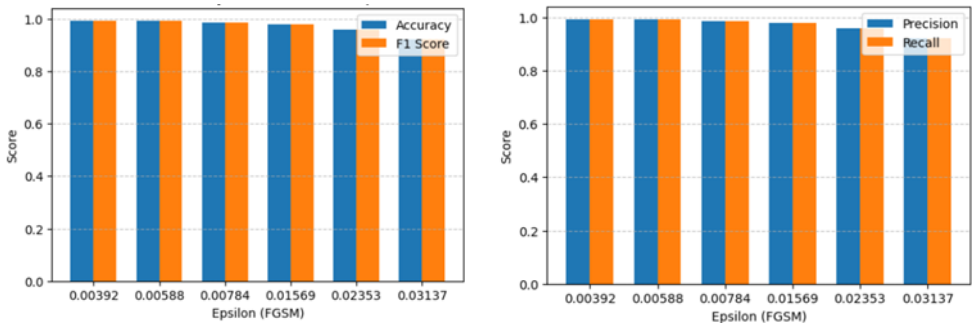


Figure 4. Demonstrates the effect of varying epsilon values under FGSM attack

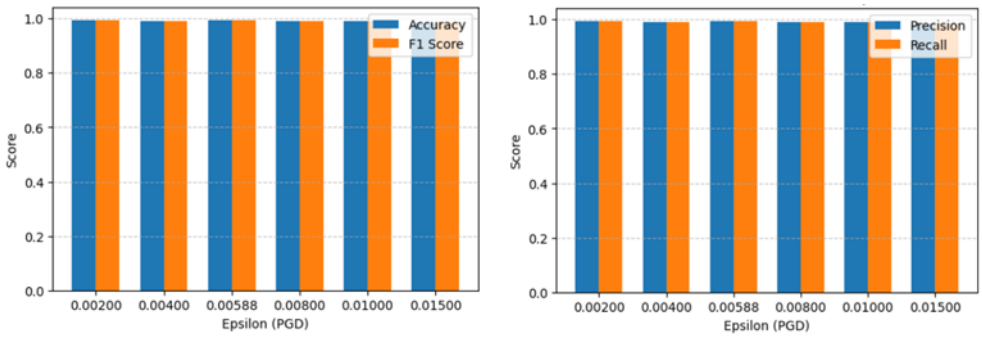


Figure 5. Demonstrates the effect of varying epsilon values under PGD attack

4. CONCLUSION

This paper presented the development of a robust hybrid deep learning framework for medical deepfake detection with improved robustness against adversarial attacks. The main contributions of this work

are threefold. Firstly, a hybrid model is designed with the combination of EfficientNetB0 with ViT to learn image representations from their spatial and semantic perspectives. Second, a novel adversarial training methodology is proposed by incorporating synthetic images generated with stable diffusion, standard image transformations, and FGSM and PGD attack techniques to increase the robustness of the deep learning model against various image manipulations. Third, a detailed evaluation has been performed under synthetic, transformed, and adversarial testing conditions to evaluate the proposed model. The experimental results have shown that the proposed model has achieved an accuracy of 96-99% under all test conditions. In the future, the scope of this research is to extend the proposed framework to other medical imaging modalities and to integrate it into blockchain or federated learning systems.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Keshavmurthy Shankar Krupa	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Kiran Chandrappa		✓				✓		✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [KSK], upon reasonable request.

REFERENCES

[1] S. Hussain *et al.*, "Modern diagnostic imaging technique applications and risk factors in the medical field: a review," *BioMed Research International*, vol. 2022, 2022, doi: 10.1155/2022/5164970.
[2] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: transforming the practice of medicine," *Future Healthcare Journal*, vol. 8, no. 2, pp. e188–e194, 2021, doi: 10.7861/fhj.2021-0095.
[3] S. A. Alowais *et al.*, "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," *BMC Medical Education*, vol. 23, no. 1, 2023, doi: 10.1186/s12909-023-04698-z.
[4] P. Pradeepan, S. G. Raj, and J. George, "A comprehensive review of deepfakes in medical imaging: ethical concerns, detection techniques and future directions," *Applied Computer Science*, vol. 21, no. 2, pp. 139–153, 2025, doi: 10.35784/acs_7054.
[5] A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: a systematic and comprehensive review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 2, 2024, doi: 10.1002/widm.1520.
[6] B. Liu, B. Liu, T. Zhu, and M. Ding, "A review of deepfake and its detection: from generative adversarial networks to diffusion models," *International Journal of Intelligent Systems*, vol. 2025, no. 1, 2025, doi: 10.1155/int/9987535.
[7] L. Stroebel, M. Llewellyn, T. Hartley, T. S. Ip, and M. Ahmed, "A systematic literature review on the effectiveness of deepfake detection techniques," *Journal of Cyber Security Technology*, vol. 7, no. 2, pp. 83–113, 2023, doi: 10.1080/23742917.2023.2192888.
[8] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools and Applications*, vol. 81, no. 5, pp. 6259–6276, 2022, doi: 10.1007/s11042-021-11733-y.
[9] S. Solaiyappan and Y. Wen, "Machine learning based medical image deepfake detection: a comparative study," *Machine Learning with Applications*, vol. 8, 2022, doi: 10.1016/j.mlwa.2022.100298.
[10] I. Amerini *et al.*, "Deepfake media forensics: status and future challenges," *Journal of Imaging*, vol. 11, no. 3, 2025, doi: 10.3390/jimaging11030073.
[11] D. Siegel, C. Kraetzer, S. Seidlitz, and J. Dittmann, "Media forensics considerations on deepfake detection with hand-crafted features," *Journal of Imaging*, vol. 7, no. 7, 2021, doi: 10.3390/jimaging7070108.
[12] M. Sharafudeen, J. Andrew, and V. S. S. Chandra, "Leveraging vision attention transformers for detection of artificially synthesized dermoscopic lesion deepfakes using derm-CGAN," *Diagnostics*, vol. 13, no. 5, 2023, doi: 10.3390/diagnostics13050825.

[13] S. A. Khan and D. T. D.-Nguyen, "Deepfake detection: analyzing model generalization across architectures, datasets, and pre-training paradigms," *IEEE Access*, vol. 12, pp. 1880–1908, 2024, doi: 10.1109/ACCESS.2023.3348450.
[14] A. Raza, K. Munir, and M. Almutairi, "A novel deep learning approach for deepfake image detection," *Applied Sciences*, vol. 12, no. 19, 2022, doi: 10.3390/app12199820.
[15] B. Chen, X. Liu, Z. Xia, and G. Zhao, "Privacy-preserving DeepFake face image detection," *Digital Signal Processing: A Review Journal*, vol. 143, 2023, doi: 10.1016/j.dsp.2023.104233.
[16] A. Alsaheel, R. Alhassoun, R. Alrashed, N. Almatrafi, N. Almallouhi, and S. Albahli, "Deep fakes in healthcare: how deep learning can help to detect forgeries," *Computers, Materials and Continua*, vol. 76, no. 2, pp. 2461–2482, 2023, doi: 10.32604/CMC.2023.040257.
[17] M. A. Arshed, S. Mumtaz, Ş. 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, 2024, doi: 10.3390/computation12090173.
[18] M. Karaköse, H. Yetiş, and M. Çeçen, "A new approach for effective medical deepfake detection in medical images," *IEEE Access*, vol. 12, pp. 52205–52214, 2024, doi: 10.1109/ACCESS.2024.3386644.
[19] G. N. Vivekananda, T. R. Mahesh, M. Gupta, A. Thakur, and A. Sayal, "Refining digital security with EfficientNetV2-B2 deepfake detection techniques," *Egyptian Informatics Journal*, vol. 30, 2025, doi: 10.1016/j.eij.2025.100699.
[20] S. Dey, S. Das, P. Nani, and T. Debbarma, "LUCIDNet: deepfake generation and detection in CT-scan lung cancer imaging via stable diffusion and enhanced EfficientNetV2-L." *Research Square,* May 2025, doi: 10.21203/rs.3.rs-6634989/v1.
[21] S. K. Padhi, H. Kumar, U. Kashyap, and S. S. Ali, "De-Fake: style based anomaly deepfake detection," 2025, *arXiv:2507.033342025.*
[22] D. Awasthi, P. Khare, V. K. Srivastava, A. K. Singh, and B. B. Gupta, "DeepNet: protection of deepfake images with aid of deep learning networks," *Image and Vision Computing*, vol. 158, 2025, doi: 10.1016/j.imavis.2025.105540.
[23] M. Sharafudeen and V. C. Vinod, "Dual residual learning of frequency fingerprints in detecting synthesized biomedical imagery," *Applied Soft Computing*, vol. 173, 2025, doi: 10.1016/j.asoc.2025.112930.
[24] C. C. Hsu, M. Y. Tsai, and C. M. Yu, "Securing healthcare data integrity: deepfake detection using autonomous AI approaches," *IEEE Journal of Biomedical and Health Informatics*, 2025, doi: 10.1109/JBHI.2025.3586963.
[25] T. V. K. Moorthy *et al*., "A novel multi-modality deepfake detection using cross-model fusion–based hybrid harris whale depth-wise separable dense convolutional Bi-GRU deepfake model," *Cognitive Computation*, vol. 17, no. 3, 2025, doi: 10.1007/s12559-025-10465-7.
[26] Y. Mirsky, "Deepfakes: medical image tamper detection," *UCI Machine Learning Repository*, 2020, doi: 10.24432/C5J318.
[27] K. Kansal, T. B. Chandra, and A. Singh, "ResNet-50 vs. EfficientNet-B0: multi-centric classification of various lung abnormalities using deep learning 'session id: ICMLDsE.004," *Procedia Computer Science*, vol. 235, pp. 70–80, 2024, doi: 10.1016/j.procs.2024.04.007.
[28] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: a survey," *ACM Computing Surveys*, vol. 54, no. 10, 2022, doi: 10.1145/3505244.
[29] T. Zhang, Z. Wang, J. Huang, M. M. Tasnim, and W. Shi, "A survey of diffusion based image generation models: issues and their solutions," 2023, *arXiv:2308.13142*.
[30] J. Sen and S. Dasgupta, "Adversarial attacks on image classification models: FGSM and patch attacks and their impact," in *Information Security and Privacy in the Digital World - Some Selected Topics*. London, United Kingdom: IntechOpen Limited, 2023, doi: 10.5772/intechopen.112442.
[31] K. Chen, J. Wang, and J. Zhang, "An overview of adversarial attacks and defenses," *Journal of Information Hiding and Privacy Protection*, vol. 4, no. 1, pp. 15–24, 2022, doi: 10.32604/jihpp.2022.029006.
[32] S. Albahli and M. Nawaz, "MedNet: medical deepfakes detection using an improved deep learning approach," *Multimedia Tools and Applications*, vol. 83, no. 16, pp. 48357–48375, 2024, doi: 10.1007/s11042-023-17562-5.
[33] M. Sharafudeen and S. S. V. Chandra, "Medical deepfake detection using 3-dimensional neural learning," in *Artificial Neural Networks in Pattern Recognition*. Cham, Switzerland: Springer, 2023, pp. 169–180, doi: 10.1007/978-3-031-20650-4_14.

BIOGRAPHIES OF AUTHORS



Keshavmurthy Shankar Krupa is currently serving as an assistant professor in the Department of Computer Science and Engineering at B.M.S. College of Engineering, Bengaluru, India. She received her bachelor of engineering and master of technology degrees in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Belagavi, Karnataka, India. Her research interests include AI and machine learning (ML), cybersecurity, internet of things (IoT), and blockchain technology. She has presented her work at several international conferences and published in peer-reviewed journals. She can be contacted at email: krupaks.csi@bmsce.ac.in.



Dr. Kiran Chandrappa is currently serving as professor and head of the Department of Information Science and Engineering at Global Academy of Technology, Bengaluru, India. He received his Bachelor of Engineering from Kuvempu University, a master of technology degree from Visvesvaraya Technological University (VTU), Belagavi, and Ph.D. from Jain University, Bengaluru, Karnataka, India. His research interests include computer vision, image processing, AIML, and blockchain technology. He has contributed to the academic community by publishing articles at various international conferences and peer-reviewed journals. He can be contacted at email: kiranchandrappa@gmail.com.