Mesopotamian Journal of Cybersecurity
Vol. 5, No.2, **pp**.863-885
DOI: https://doi.org/10.58496/MJCS/2025/049; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/cybersecurity

Research Article

Adversarial Attacks on Hybrid Attention Integrated Transfer Learning for Lung Cancer CT Classification

Omar Ibrahim Obaid[1,*], Abdulbasit ALazzawi[2]

[1]*Department of Computer, College of Education, Al-Iraqia University, Baghdad, Iraq*

[2]*Diyala University, College of Science, Computer Department, Diyala, Iraq*

ARTICLEINFO

Article History

Received 20 May 2025
Revised 15 Jul 2025
Accepted 9 Aug 2025
Published 27 Aug 2025

Keywords

Lung cancer
Classification
Attention Mechanisms
Adversarial attacks
Transfer learning

ABSTRACT

Deep learning–based classification of lung cancer from CT images can achieve high accuracy but is vulnerable to adversarial attacks that introduce imperceptible perturbations, potentially leading to misdiagnoses. Paying attention to neural networks in transfer learning could improve both the effectiveness and resistance to change. In this paper, we propose a hybrid framework that combines a MobileNetV2 backbone with channel–spatial attention modules and white-box adversarial testing via the fast-focused gradient sign method (FFGSM) and projected gradient descent under an L_2 norm constraint ($PGDL_2$). The model was trained end-to-end on a stratified CT dataset of 3,451 images (normal, benign, malignant) with adversarial examples injected during training (ε=4/255; $PGDL_2$: α=1/255, 7 steps). The evaluation of the performance was based on accuracy and precision, recall, the F1- score and the reduction in model performance due to adversarial attacks. For clean inputs, the attention-augmented model achieved 86% accuracy (FFGSM) and 95% accuracy ($PGDL_2$), with balanced F1-scores >0.90 across classes. Under adversarial attack, the accuracy decreases to 78% (FFGSM) and 86% ($PGDL_2$), indicating a smaller robustness drop for $PGDL_2$-augmented training. Having attention modules in the model significantly enhanced the ability to discern features, saving up to 9% in performance compared with the models without attention. These methods made it clear that the model was more focused on clinically important parts. Incorporating hybrid channel–spatial attention into transfer-learning pipelines substantially improve the accuracy and resilience of lung cancer CT classification to strong adversarial attacks. The results of our study can help guide the development of strong AI tools for examining medical images.

1. INTRODUCTION

Lung cancer classification from CT scans requires immediate and correct assessment because this disease continues to kill many people worldwide [1]. The high-resolution imaging provided by CT technology demands tedious manual interpretation, which also suffers from observer variations in interpretations [2]. Deep learning models with convolutional neural networks excel at extracting hierarchical features from CT images, which yields better effectiveness than traditional diagnosis systems do [3]. Deep model training on small CT datasets usually results in poor generalization ability along with overfitting because medical data labelling remains a significant challenge [4]. The practice of transferring existing knowledge from natural image repositories such as ImageNet through fine-tuning becomes possible with transfer learning (TL) methods. The early network layers master basic features such as edges and textures in generic settings before later network layers specialize in lung-related patterns during fine-tuning [5]. The implementation of transfer learning for CNNs leads to decreased training duration and leaner data requirements and yields superior results, with more than 90% accurate nodule malignancy assessments through the use of 200 labelled scans [6]. Lung cancer applications record successful results from two TL methods: freezing convolutional weights while training a new classifier or conducting complete network fine-tuning [5]. CT-based lung cancer classification now uses DL with TL as its main operational practice. The combination of artificial intelligence technologies with test-time and learner variations enhances both early disease identification and robust diagnostic equipment for radiologists to better detect diseases early before mortality rates rise [3][6]. Deep learning has revolutionized healthcare while producing its most substantial effects in radiology [7, 8]. Deep learning-based tools have appeared in large numbers and have moved efficiently through development and deployment stages, securing several FDA approvals [9–11]. Healthcare institutions maintain an increasing pace of implementation for these technologies within their clinical operations [12–15]. The advance of information technologies increases the degree of security risk for potential cyberattacks. The security flaw called an adversarial attack has become the greatest concern in research [16, 17].

*Corresponding author. Email: omar.obaid@aliraqia.edu.iq

An adversary uses small changes in original input data combined with fabricated false information to force model mistakes during these attacks. The imperceptible disturbances that cannot be detected by humans can nonetheless misguide algorithms into mistaken decisions. Executing adversarial attacks requires certain specific steps that must be followed. Running an adversarial attack requires designing precise alterations that specifically strike at vulnerabilities of an algorithm and typically demands knowledge of the model design or extensive testing [18]. The modification of medical images in radiological contexts through attacks deceives deep learning systems so that they produce incorrect interpretations that lead to incorrect diagnoses. Artificial intelligence system security problems become evident because of these attacks, which affect safety-critical domains. Recent research indicates that model robustness evaluation against such attacks has become critical because AI systems drive healthcare and autonomous vehicles along with cybersecurity [19].

Deep learning models yield impressive results when applied throughout various applications. White-box attack approaches: FFGSM and $PGDL_2$ represent substantial threats to model systems. Two white-box attack strategies known as FFGSM and $PGDL_2$ have gained popularity because they threaten model security [20]. The FFGSM improves upon the fast gradient sign method through a single-step algorithm that maximizes loss functions via gradient-based adjustments [21]. The algorithm first calculates loss gradient values regarding input parameters and then executes one movement along the resulting gradient direction, which maximizes this loss function. The high efficiency of the FFGSM raises concerns about its use in practical real-time systems that have limited computational capacity [22]. $PGDL_2$ utilizes successive small perturbation steps that adhere to L2-norm constraints for its implementation. Studies have demonstrated that $PGDL_2$ operates as an effective first-order attack that bypasses numerous defense mechanisms [23]. The iterative gradient-based strategy in $PGDL_2$ produces more significant adversarial examples than single-step approaches do because it performs gradient-based movements while projecting input vectors along L2-norm constraint spheres [24]. The core distinction among these methodologies lies in their computational performance and their ability to perform attacks. The FFGSM attains high speeds through one-step attacks yet comes with lower power than $PGDL_2$, which demands more computer resources.

Furthermore, attention mechanisms have been implemented in convolutional neural networks (CNNs) so that they can improve performance by adapting weights between channel and spatial features to concentrate on relevant clinical zones [25, 26]. The channel attention modules optimize feature-map responses to focus on important imaging biomarkers—such as CA-Net's channel module, which increased segmentation precision during ISIC 2018—while spatial attention modules detect diagnostic areas on CT or MRI slices [26]. The sequential execution of attention strategies through hybrid attention modules (HAMs) delivers stable performance improvements in multiorgan analysis as well as skin lesion identification processes via reduced parameter resources [27, 28]. Current convolutional neural network frameworks could be enhanced by using lightweight attention mechanisms to both increase accuracy and achieve performance improvements.

This study proposes a new framework that integrates fast-focused gradient sign methods (FFGSMs) with projected gradient descent under an L_2 constraint ($PGDL_2$) in a transfer-learning platform that includes hybrid channel–spatial attention. Two white-box adversarial attacks were applied to the lung cancer CT classification models to test their weakness levels and understand how attention mechanisms affect feature representation during attacks. Our contributions are threefold: (1) We quantify the impact of FFGSM and $PGDL_2$ attacks on diagnostic accuracy, revealing critical failure modes. This research measures the diagnostic reliability degradation caused by FFGSM and $PGDL_2$ assaults and investigates hybrid attention system impacts on resistance to manipulated signals. In this study, we closed the gap between adversarial testing of attention-augmented classifications by conducting laboratory experiments that yield practical recommendations for building stronger lung cancer diagnostic systems.

2. RELATED WORK

2.1. Deep Learning and Transfer Learning for Lung Cancer CT Classification

Deep learning models, that is, convolutional neural networks (CNNs), have made breakthroughs in lung cancer classification via chest computed tomography (CT). Initial work used trained ImageNet networks such as ResNet-50 and VGG19 [29] to extract features followed by fine-tuning them to datasets of pulmonary nodules and achieved significant improvement over traditional machine learning methods. The following works introduced the concept of transfer learning in an effort to overcome the problem of labelled medical image availability, illustrating that, in addition to reducing training times, the use of pretrained weights significantly improves on small training sets [30, 31]. However, current studies continue to focus on the single-label context, even though reports based on multilabel CT are so prevalent in clinics. In addition, explainability tools are not used to a large extent, which creates interpretability drawbacks for diagnostic decision support [32]. By presenting a multilabel CT classification model, the proposed method eliminates such deficiencies by providing an embedded Grad-CAM++ visualization of the decision regions.

2.2. Attention Mechanisms in Medical Imaging

Attention mechanisms have become fundamental in deep learning models, especially in medical imaging analyses. This instillation of channel and spatial attention modules in models, as in CBAM [33], allows the network to focus on clinically relevant areas and therefore contributes to finding any minor pathologies. Initial research focused on segmentation tasks only [34], but later research has shown that attention can be used in classification, where it has also been shown to perform significantly better and be more resistant to image noise [35]. Despite such successes, most of the current methods focus only on one network level and do not consider hierarchical interactions between features across the layers. In addition, few articles have discussed how attention modules are vulnerable to adversarial perturbations [36]. The current study combines the idea of multilevel attention and the application of Grad-CAM++ visualization, increasing both the classification accuracy and interpretability.

2.3. Adversarial Attacks and Defences in the Medical Domain

Through adversarial attacks, it is possible to expose dangerous vulnerabilities in deep learning-based diagnostic systems, as they can be distorted by tiny perturbations to elicit incorrect diagnoses [37, 38]. Szegedy *et al*. [39] first demonstrated that CNN misclassification is caused by small perturbations over its inputs, and Goodfellow *et al*. [40] later developed an efficient adversarial example generation procedure known as the fast gradient sign method (FGSM). Further developments in attack tactics—one-pixel attacks [41, 42] and propagation map analyses—have shown that perturbations are applied in receptive fields and not individual neurons. In some recent research carried out within the medical domain, it was confirmed that CT-based classifiers are vulnerable to such attacks [43]. Although this has already been achieved significantly, the current body of knowledge has not been fully incorporated into attention-augmented, multilabel CT classifiers, which are subject to real-world inference limitations. This gap is bridged by the current study, which assesses the strength of attention modules in the lightweight multioutput framework and explains how to establish an adequate defense against attacks that are specific to medical CT use. A potential vulnerability caused by adversarial radiation, which is beyond radiological imaging, extends to other safe critical life-saving biomedical applications, especially electroencephalography-based brain–computer interfaces (BCIs). The work of Sayah Ben Aissa *et al*. [44] is similar to the current research, given that it examines the resilience of attention-based architectures to attacks by assessing the robustness of attention-based architectures to adversarial attacks in the field of motor imagery classification via EEG data. This proves that the interplay of the attention mechanisms and the security of the model represents an upstart area of concern. Related literature on the issue therefore supports the need to consider what such attention techniques imply in model security: as they find their way into deep learning systems to increase their performance, it becomes imperative to assess their implications for the security of these systems. This external environment supports the current analysis time and needs to focus on the particular application of the problem of lung cancer CT classification.

3.DATASET

In a three-month period during the autumn of 2019, the IQ-OTH/NCCD acquired chest CTs from 110 patients with lung cancer at various stages [45], reciprocating 1,190 axial slices. All the images were labelled by specialist oncologists and radiologists in the IQ-OTH/NCCD. Stratification was based on three diagnostic groups of cases. normal (n = 55), benign (n = 15) and malignant (n = 40). The present study used an available Kaggle dataset [46] to address the problem of imbalanced classes in lung cancer imaging and, consequently, the categories that are underrepresented.

3.1 Dataset overview

A comprehensive dataset consisting of 3,451 CT-scan images was collected and organized into 3 different classes: Benign, Malignant, and Normal. These images were first retrieved from a Google drive repository and later copied to Google Colab for processing. Table 1 shows an overview of the characteristics of the entire dataset.

TABLE I. DATASET OVERVIEW

Metric	Value
Total images loaded	3,451
Number of classes	3
Unique duplicate hashes found	157
Duplicate images excluded	314
Invalid images found	0

3.2 Data Integrity Assessment

To guarantee dataset quality, we introduce two main integrity checks. First, the images were checked for corruption and size adequacy, with the images that were less than 10 × 10 pixels being flagged to eliminate them. No invalid images were detected in this process. Second, a custom hashing function with the help of the SHA technique was used to find specific duplicates within the dataset. This procedure revealed 157 unique duplicate hashes for 314 images, which were then removed from further processing.

3.3 Class distribution

The class distribution in the dataset was balanced, as each category represented roughly one third of the total dataset. Table 2 presents the actual distributions for the three classes.

TABLE II. OVERALL CLASS DISTRIBUTION

Class	Images	Percentage
Benign	1,074	31.12%
Malignant	1,186	34.37%
Normal	1,191	34.51%

The class distribution balance was carefully preserved across all the splits, with variations limited to ±0.1%, as shown in Table 3.

TABLE III. CLASS DISTRIBUTION PER SPLIT

Split	Benign (%)	Malignant (%)	Normal (%)
Training	687 (31.11%)	759 (34.38%)	762 (34.51%)
Validation	172 (31.16%)	190 (34.42%)	190 (34.42%)
Test	215 (31.11%)	237 (34.30%)	239 (34.59%)

3.4 Dataset partitioning

The dataset was then split and stratified by class into a training set, a validation set, and a test set. We used an initial 80/20 split to split the test set, and then we allocated 20% from that remainder for validation purposes. The number of images partitioned under this scheme was 2,208 training images, 552 validation images, and 691 test images, as shown in Table 4.

TABLE IV. DATA SPLITS

Split	Images
Training	2,208
Validation	552
Test	691

3.5 Normalization Parameters

To normalize the model training pixel values, we computed the normalization statistics of a random set of 1,000 training images. This method avoids possible data leakage from test or validation sets. Before computation, the intensity of all the pixels was rescaled to the [0, 1] range. The obtained channelwise statistics are shown in Table 5.

TABLE V. NORMALIZATION STATISTICS

Statistic	Channel R	Channel G	Channel B
Mean	0.4038	0.4038	0.4038
Std	0.3004	0.3004	0.3004

3.6 Data Augmentation

Using the Albumentations library, we applied an extensive data augmentation method to increase model generalizability. For the validation and test images, only resizing and normalization were applied to maintain the integrity of the evaluation process. For the training dataset, all the images were passed through the following sequence of transformations:

1. Resizing to 299 × 299 pixels
2. Random horizontal and vertical flips
3. Random rotations
4. Brightness and contrast adjustments
5. Application of elastic, grid, or optical distortions
6. Occasional Gaussian or median blurring
7. Color adjustments (CLAHE or sharpening)
8. Normalization via previously calculated statistics
9. Conversion to tensors via ToTensorV2

3.7 Implementation details

To ensure experimental reproducibility, all the random operations, regardless of hash, sampling and augmentation, were initialized with a seed value of 42. Data loading was implemented with PyTorch's dataloader, a batch size of 32,4 worker processes and pinned memory to speed up the data transfer to the GPU. The training data loader was set to be shuffled at the start of every epoch, but the validation and test loaders kept the data order in check.

3.8 Dataset samples

The dataset consists of various medical images that reflect the variety of radiological manifestations of lung cancer. Expert radiologists practice systemic sampling and provide detailed annotations to reflect the diagnostic variability measured in the course of routine clinical practice. In this section, we outline the structures of the given dataset by demonstrating its classwise distribution and examining critical image parameters. One cannot overstate the importance of a full examination of every sample to inform the following preprocessing steps and build strong classification models.

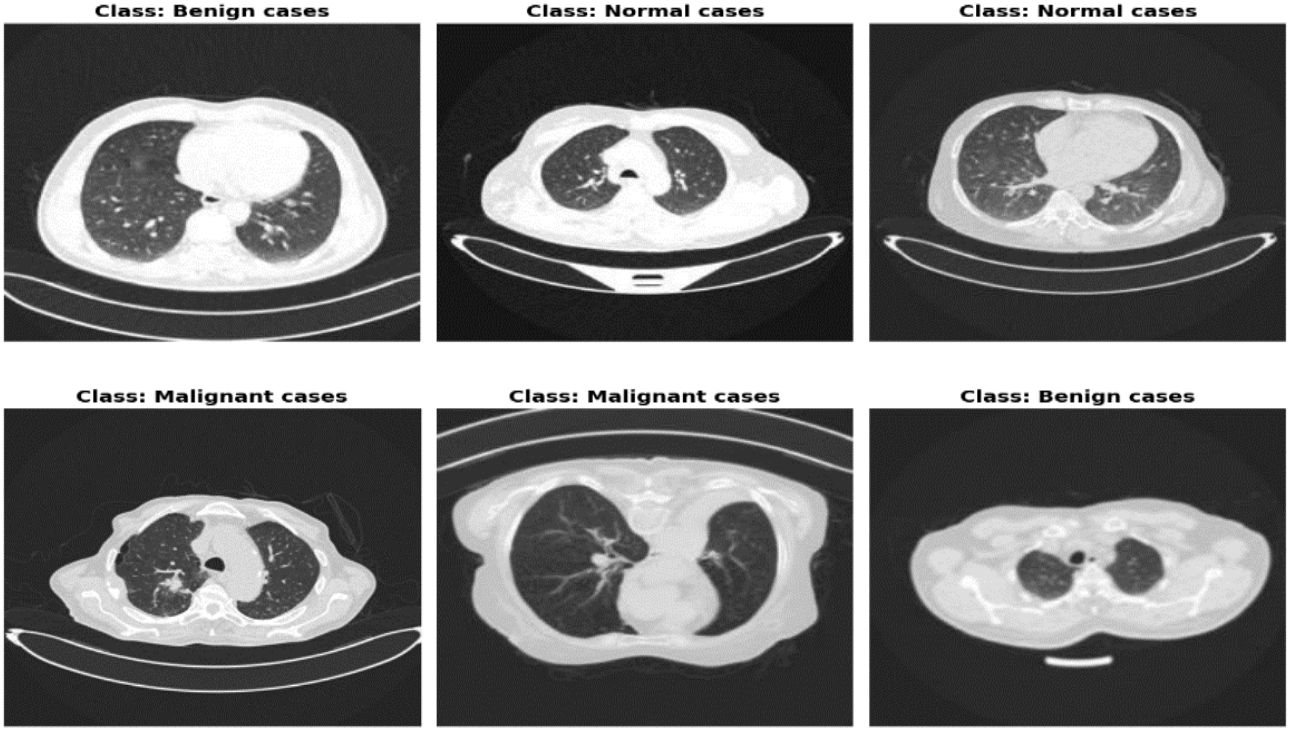

Fig. 1. Sample of the dataset

4. METHODS

4.1 Model Architecture

The developed model of medical image classification combines a backbone feature extractor with a hybrid attention mechanism to improve feature representation. The architecture is modular so that the backbone choice and attention integration can vary. The main building blocks are as follows in the model, as illustrated in Figure 2.

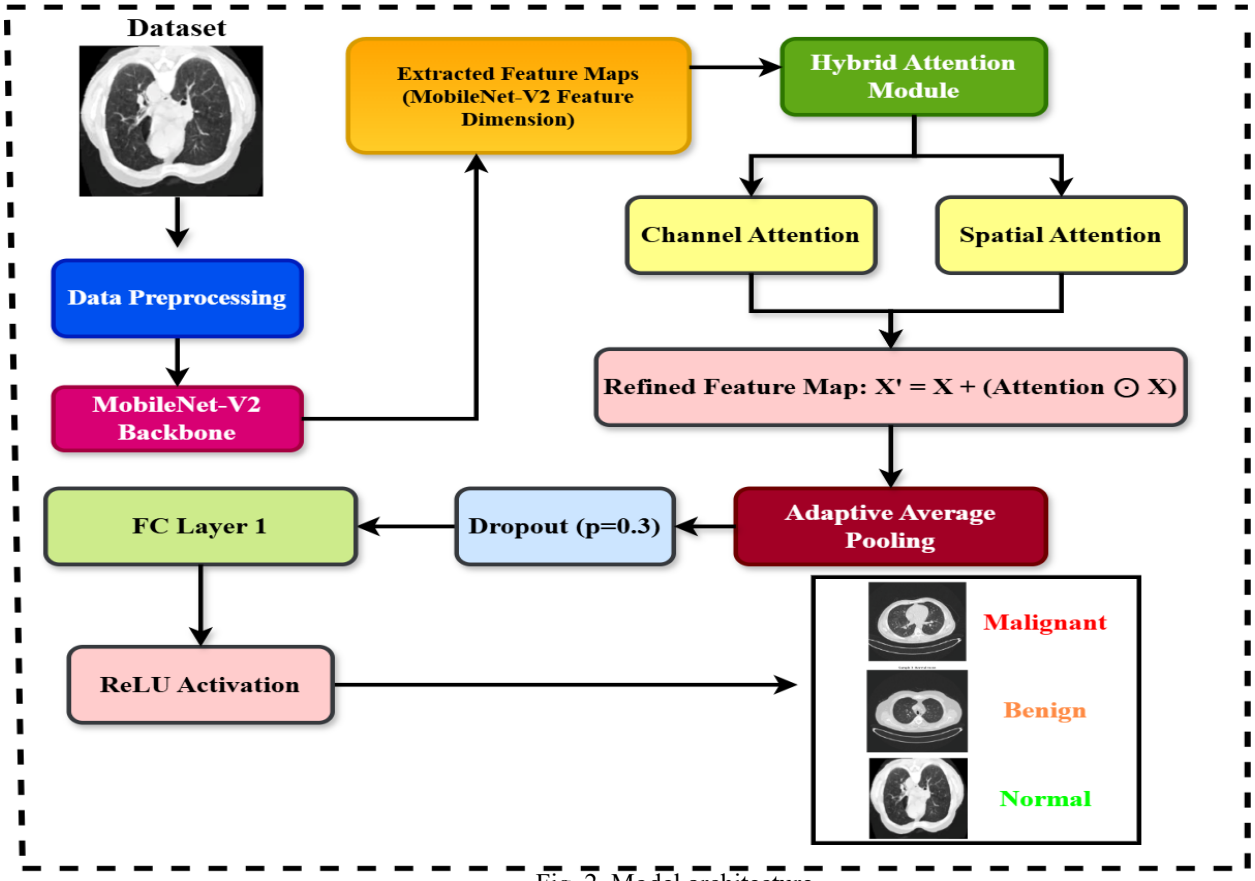

Fig. 2. Model architecture

4.1.1 Backbone Feature Extractor

The model works with a MobileNetV2 backbone. This backbone is initialized with pretrained weights (e.g., ImageNet) to use transfer learning. The feature extraction layers are modified to produce a fixed-dimensional feature representation. MobileNetV2 is commonly selected as the target model in adversarial-attack research because of the lightweight and depthwise-separable design of the network, which achieves the right trade-off between computational efficiency and representational capacity; therefore, it is both a reasonable target for edge deployments and a demanding benchmark for vulnerability analysis. MobileNetV2 has also been shown to work well for fine-tuning in medical imaging environments for applications such as pulmonary nodule detection and magnetic resonance imaging (MRI) classification, where high baseline accuracy at these tasks makes it possible for researchers to examine the impact of adversarial effects, as opposed to performance artifacts.

4.1.2 Hybrid Attention Module

A hybrid attention block is added to improve the feature maps. Both channel attention and spatial attention are used in this module to improve discriminative features. The attention mechanism is used after feature extraction but before the classification head, ensuring that the model will orient to the most important parts of the input image. The following figure illustrates the channel and spatial attention module.

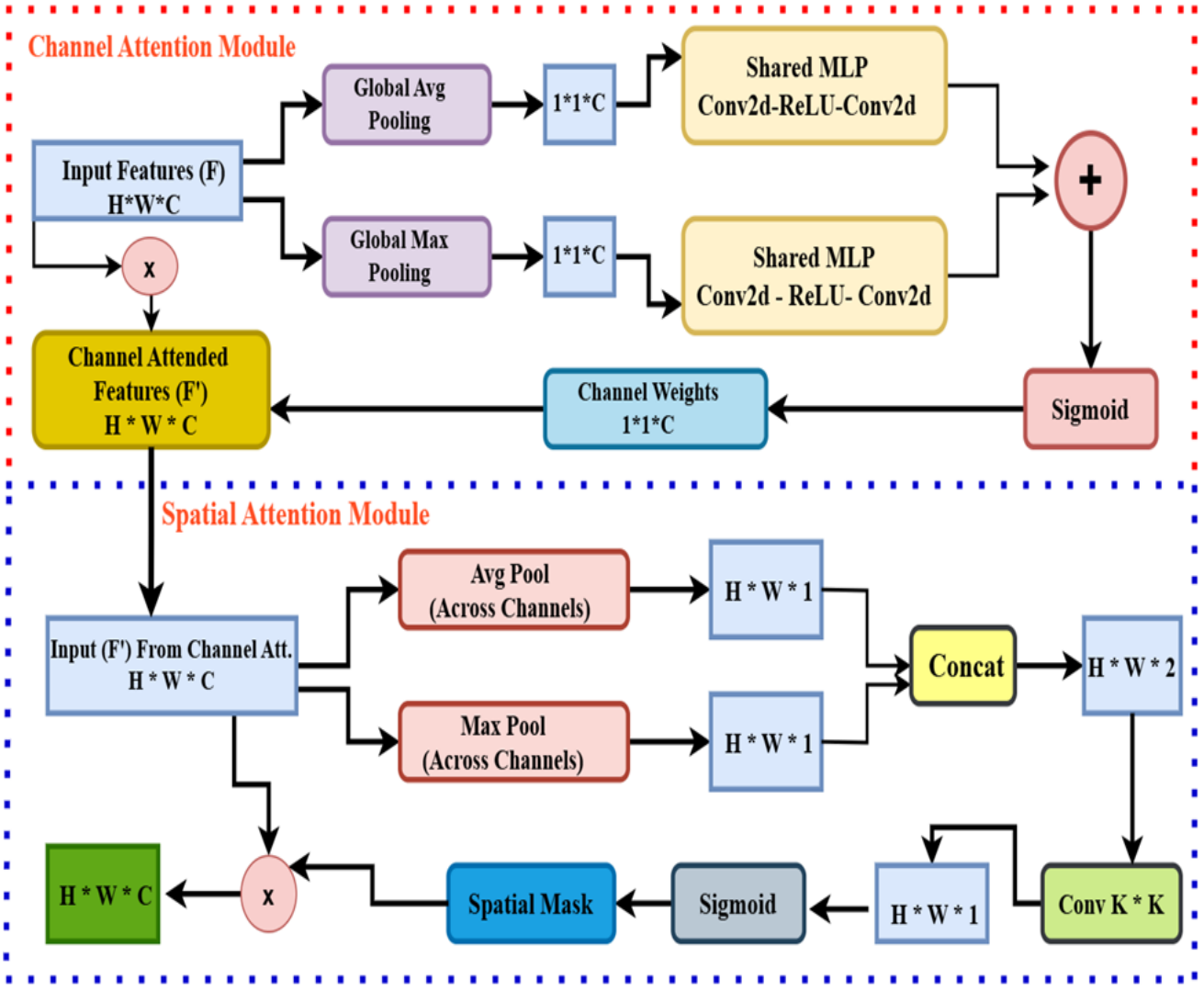

Fig. 3. Hybrid attention module

The diagram shows a modified two-stage attention mechanism that aims to narrow the attention to channels and, finally, spatial locations. The channel attention module determines which of the feature channels carry the most information for the given input. It works by:

- The module sums every channel twice over all spatial positions by taking the average and the maximum.
- Both summaries go through the same small neural network ("shared MLP") learning to accentuate or detract channels.
- Both outputs are combined and squashed into values between zero and one, thus obtaining one attention weight per channel.
- These channels re-scale the feature map of the original image, amplifying rewarding channels and attenuating less important channels.

Furthermore, the spatial attention module determines what locations of the feature map contain the most important information. It works by:

- The refined feature map from the channel stage is collapsed along the channel dimension twice: once by computing average activation at each position and once by computing the maximum activation.
- These two 2D activation maps are stacked and then fed into a lightweight convolutional layer that learns to produce a single 2D attention mask.

- After being subjected to a sigmoid, this spatial mask has values that range from zero to one per pixel and show the importance of each position.
- The mask multiplies the feature map, which accentuates interest in critical spatial areas while muting background or noise.

The attention mechanism adopted in this experiment is based on the convolutional block attention module (CBAM), but with slight alterations that improve performance and increase flexibility. In contrast to the original CBAM, which makes use of fully connected layers, as part of its MLP architecture, our implementation replaces them with a 1 × 1 convolutional stack. Compared with CBAM, which has a kernel of only 7 × 7, the current design can have a variable kernel size (named K × K) depending on the task. Table 6 describes the main differences between the proposed method and the common CBAM.

TABLE VI. DIFFERENCES BETWEEN STANDARD CBAM AND PROPOSED VARIANT

Component	Standard CBAM	Proposed Variant	Justification
Channel Attention MLP	Fully Connected Layers	Conv2D → ReLU → Conv2D (1×1 convolutions)	More lightweight and compatible with fully convolutional networks.
Spatial Attention Kernel	Fixed 7×7 convolution	Flexible K×K convolution	Allows task-specific tuning of the receptive field.

The key benefits of this hybrid module are as follows: First, it identifies which channels are most informative, and afterwards, it determines where to look within those channels. Second, there is a minimal increase in both the parameter count and computational complexity with each attention mechanism. Third, with adaptive feature reweighting, the network places more focus on important details, which increases the performance in downstream applications such as classification and detection.

4.1.3 Classification Head

The classification head is composed of a global average pooling (GAP) layer that reduces spatial dimensions to 1×1, a dropout layer (0.3 dropout rate) that mitigates overfitting on small medical datasets, and a fully connected layer of end predictions.

4.2 Adversarial Attack Implementation

We use a threat model in which the adversary knows all aspects of the classifier's setup, parameters, architecture, and training and can therefore utilize gradient-based perturbation techniques. Two attacks are carried out in our study by using the functions provided by the torchattack library. The assaults include the FFGSM, which extends the basic FGSM by using random initialization for each sample, and $PGDL_2$, an iterative strong variant of the PGD penalized with the L_2 norm. Both methods depend on ε (maximum perturbation), α (step size), and the number of iterations for $PGDL_2$. A white-box adversary model is assumed to be able to:

- Measure gradients of the loss J (θ, x, y) with respect to the input x for an arbitrary x.

- Use knowledge of both the model weights and its architecture to produce perturbations that lead to the greatest errors.

- Carry out attacks within the training procedure during adversarial training.

From a robustness point of view, this is the hardest possible situation in medical imaging because trivial modifications can still result in incorrect diagnoses.

The current study analyses the resilience of the model against two different but widespread white-box adversary threats, which are characterized by discrepant optimization goals and a pixel-level approach. In particular, we test the FFGSM attack, which works in the L-infinity norm and adds uniformly scalable perturbations to all the pixels, and the progressive iterative $PGDL_2$ attack, which also works in the L 2 norm but restricts the perturbations to a total magnitude less than a user-specified amount, produces modifications that are predominantly localized and concentrated but typically imperceptible. This is not an attempt to obtain a quantitative view of which of the two attack algorithms is the most powerful but an assessment of comparative strength against each of the two offensive strategies by training individual models specifically against the two offensive strategies.

4.2.1 Fast FGSM (FFGSM)

In FFGSM, the process includes a uniform random selection inside the ε-ball before performing a gradient step, which improves upon the original FGSM. By randomly perturbing the input prior to the gradient step, this "free" randomization strategy stops models from overfitting to one particular type of perturbation and makes adversarial training more effective.

$$\begin{aligned} \delta^{(0)} &\sim \mathcal{U}(-\epsilon, \epsilon), \\ x_0^{adv} &= clip\big(x + \delta^{(0)}, [0,1]\big), \\ x^{adv} &= clip\left(x_0^{adv} + \alpha\, sign\left(\nabla_x J(\theta, x_0^{adv}, y)\right), [x-\epsilon, x+\epsilon]\right) \end{aligned} \tag{1}$$

where ϵ bounds the Linf perturbation and α (here set equal to ϵ) is the step size.

4.2.2 PGDL₂ attack

PGDL₂ is the modification of the PGD that uses the L₂ norm, involving updates along the gradient and projection back onto a sphere of radius ϵ. Starting from either the clean input or a random point in the L₂ ball:

$$x_0^{adv} = \begin{cases} clip\big(x + \delta^{(0)}, [0,1]\big), & \delta^{(0)} \sim \epsilon \dfrac{r}{\| \delta \|_2},\ r \sim \mathcal{U}(0,1) \\ x & \text{(no random start)} \end{cases}$$

for $t = 0 \ldots T-1$:

$$\begin{aligned} g_t &= \nabla_{x_t^{adv}} J(\theta, x_t^{adv}, y), \quad \hat{g}_t = \frac{g_t}{\| g_t \|_2 + \eta}, \\ x' &= x_t^{adv} + \alpha\, \hat{g}_t, \\ \delta &= x' - x, \quad \delta \leftarrow \delta \min\left(1, \frac{\epsilon}{\| \delta \|_2}\right), \\ x_{t+1}^{adv} &= clip(x + \delta, [0,1]). \end{aligned} \tag{2}$$

4.3 Training procedure

The model is fully trained end-to-end under a white-box threat model, making use of cross-entropy loss together with an Adam optimizer and a StepLR scheduler, with the option of enabling adversarial training via the addition of FFGSM/PGDL₂ examples to each batch for protection. Each batch is augmented with FFGSM (ε = 0.0157, α = ε) or PGDL₂ (ε = 0.25, α = 1/255, 7 steps) examples produced via full gradient information whenever adversarial training is on, and the total loss consists of both clean and adversarial cross-entropy. The hyperparameters of each adversarial training regimen (FFGSM and PGDL₂) were set independently to the point where a good compromise between clean accuracy and robustness to adversarial attacks was achieved by that specific process. The goal was not to make a normalized comparison between the individual attack algorithms but rather to analyse the performance of the resulting models. Table 4 lists the hyperparameters used in this situation.

TABLE VII. THE USED HYPERPARAMETERS

Hyperparameter	Value
Optimizer	Adam (lr = 1e-4, wd = 1e-5)
Scheduler	StepLR (step_size = 7, γ = 0.1)
Loss Function	Cross-Entropy
Epochs	30
Batch Size	16
Image Size	299×299
FFGSM Step Size (α)	ε (4/255)
PGDL₂ Step Size (α)	1/255
PGDL₂ Iterations	7
Adv Loss Weight (λ)	0.5
White-Box Assumption	Full model & gradient access

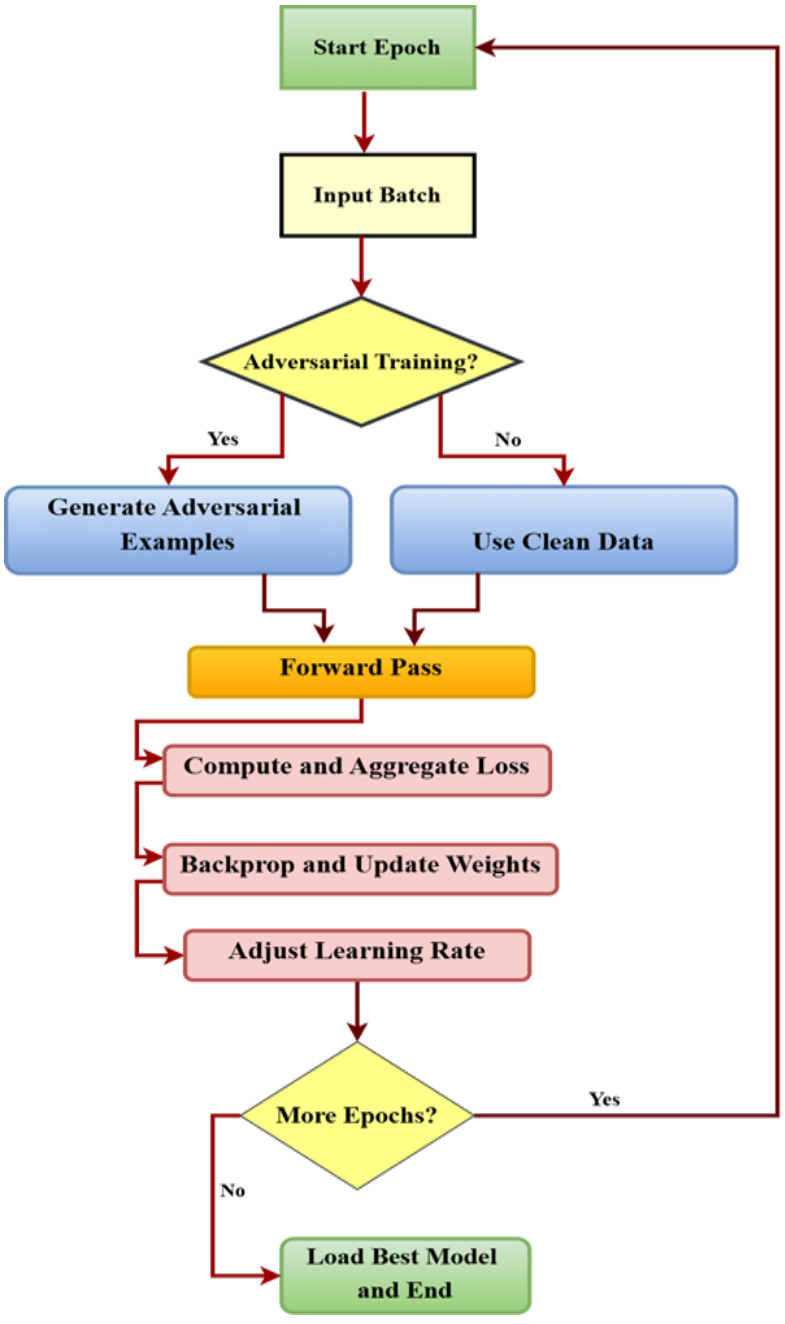

Fig. 4. Training process diagram

4.4 Evaluation Metrics

Our performance assessment for white-box adversaries involves measuring the model's performance with overall accuracy; a detailed classification report containing precision, recall, and F1-scores per class; a confusion matrix for label comparisons; the adversarial accuracy for perturbed inputs; and the amount of accuracy loss caused by adversarial perturbations. All the measurements are performed on both clean and adversarial examples to show the model's worst-case performance. Table 5 shows the performance measurement details.

TABLE VIII. PERFORMANCE MEASURING DETAILS

Metric	Definition
Overall Accuracy	$(TP + TN)/(TP + TN + FP + FN)$
Classification Report	Precision, recall, F1-score per class
Confusion Matrix	Matrix of true vs. predicted label counts
Adversarial Accuracy	Accuracy on adversarial perturbed inputs (FFGSM/PGDL2)
Drop in Accuracy (Δ)	$\mathbf{Acc}_{\mathbf{clean}} - \mathbf{Acc}_{\mathbf{adv}}$

Performance was assessed on distinct clean test data. We calculated the cross-entropy loss as well as standard performance metrics. The metrics calculated were accuracy, weighted precision, recall, the F1 score, confusion matrices and comprehensive classification reports. We evaluated resilience by computing adversarial perturbations and measuring model behaviour on perturbed samples. Adversarial loss and our chosen metrics were recorded for each variably perturbed sample. We quantified robustness by calculating the difference between model accuracy on adversarial inputs and that on clean inputs. We conducted each experiment with at least three random seeds so that our results would be representative of the model's ability. We present the results from the experiment with the training that yielded the highest average performance across repeat runs.

5. RESULTS AND DISCUSSION

The performance of the classification model is compared across clean and adversarial environments. The results of both methods produced excellent performance on clean inputs from normal high-resolution imagery.

5.1 FFGSM Method

With FFGSM training, the model achieved an overall accuracy of 86% on clean data. Both precision and recall achieved excellent results for malignant cases, with values of 1.0 and 0.91, respectively, and a corresponding F1 score of 0.95. The approaches achieved excellent performance in detecting benign cases, with a recall rate of 99%. For adversarial inputs, the results showed a slight decrease in accuracy to 78%. Nevertheless, the recognition of malignant cases continued to be reliable, with a balanced score of 0.93.

TABLE IX. CLEAN TEST RESULTS (FFGSM)

Class	Precision	Recall	F1-Score	Support
Benign Cases	0.71	0.99	0.83	215
Malignant Cases	1.00	0.91	0.95	237
Normal Cases	0.95	0.70	0.81	239
Accuracy			**0.86**	691
Macro avg	0.89	0.87	0.86	691
Weighted avg	0.89	0.86	0.86	691

TABLE 10. ADVERSARIAL TEST RESULTS (FFGSM)

Class	Precision	Recall	F1-Score	Support
Benign Cases	0.75	0.62	0.68	215
Malignant Cases	0.93	0.93	0.93	237
Normal Cases	0.66	0.76	0.70	239
Accuracy			**0.78**	691
Macro avg	0.78	0.77	0.77	691
Weighted avg	0.78	0.78	0.77	691

Figure 5 shows the accuracy and loss curves associated with training and validating the proposed MobileNet-V2–based classifier. During the first round of training, the model can quickly detect the most important data points, increasing accuracy and decreasing loss. As time progresses during training, the graphs plateau, and learning loss decreases more slowly, which means that the network improves its inner workings steadily, without major changes in performance. Finally, the learning curves mostly standstill, which indicates that the model is doing very well, with little indication of overfitting. Regularization works well throughout the model's training, which can be seen in how closely the training and validation accuracies track one another.

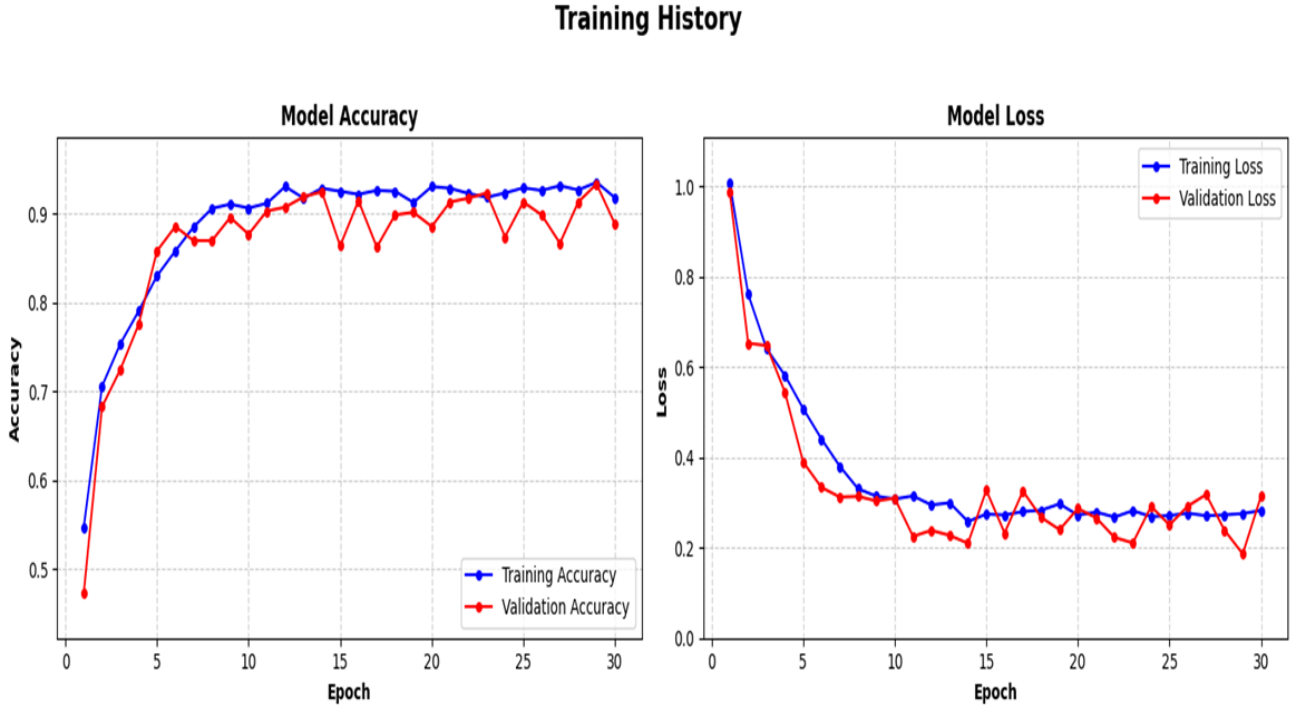

Fig. 5. Training history (FFGSM)

According to the clean confusion matrix, 70 instances of "normal cases" were falsely marked as "benign cases," and 15 "malignant cases" were incorrectly classified as "benign cases." The adversarial confusion matrix highlights a significant increase in the number of "normal cases" falsely classified as "benign cases" (81 cases).

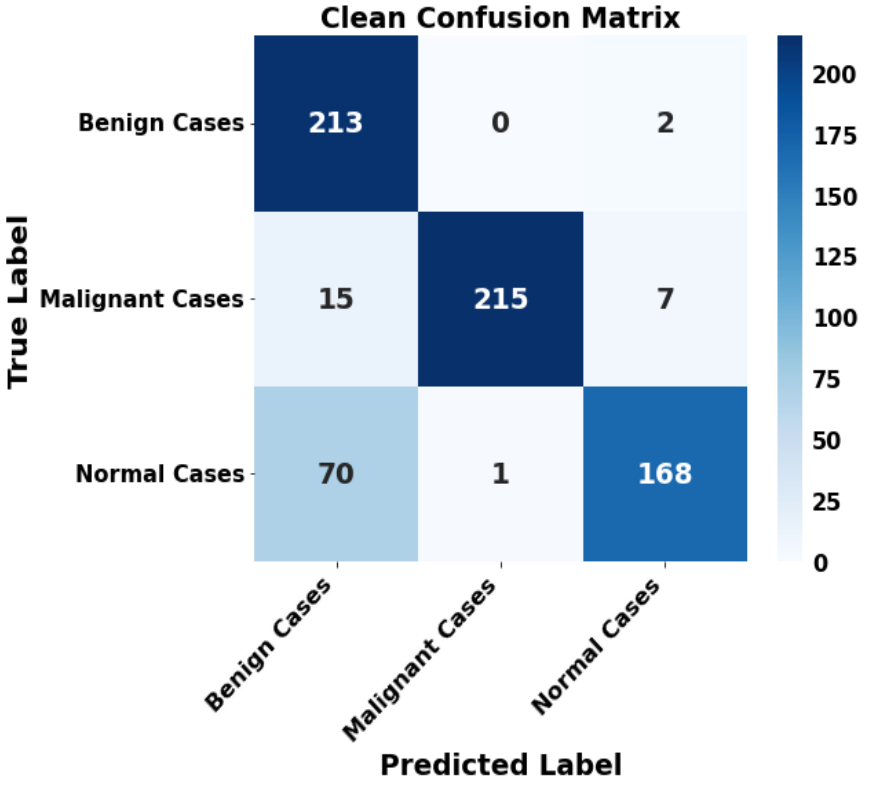

Fig. 6. Clean confusion matrix (FFGSM)

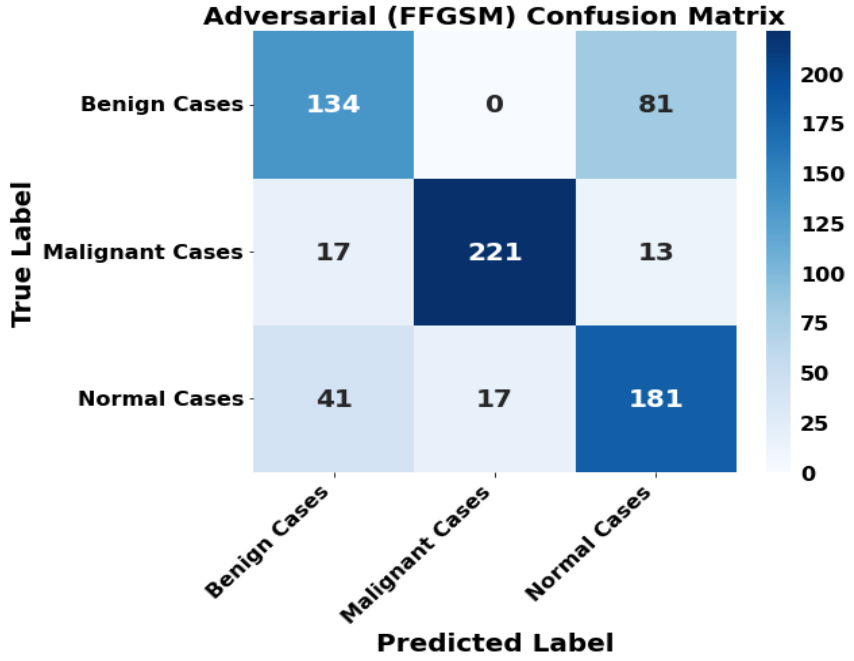

Fig. 7. Adversarial confusion matrix (FFGSM)

Figure 8 shows the visualization that compares three panels: the original image and the FFGSM image are classified with identical labels but different levels of confidence. The figure shows that changes made to the object helped the system reach a more accurate diagnosis. This example shows how small alterations to images can tell the AI model that something is strange without being noticeable to doctors, which poses serious challenges to trusting diagnostic AI systems in real-world medical applications.

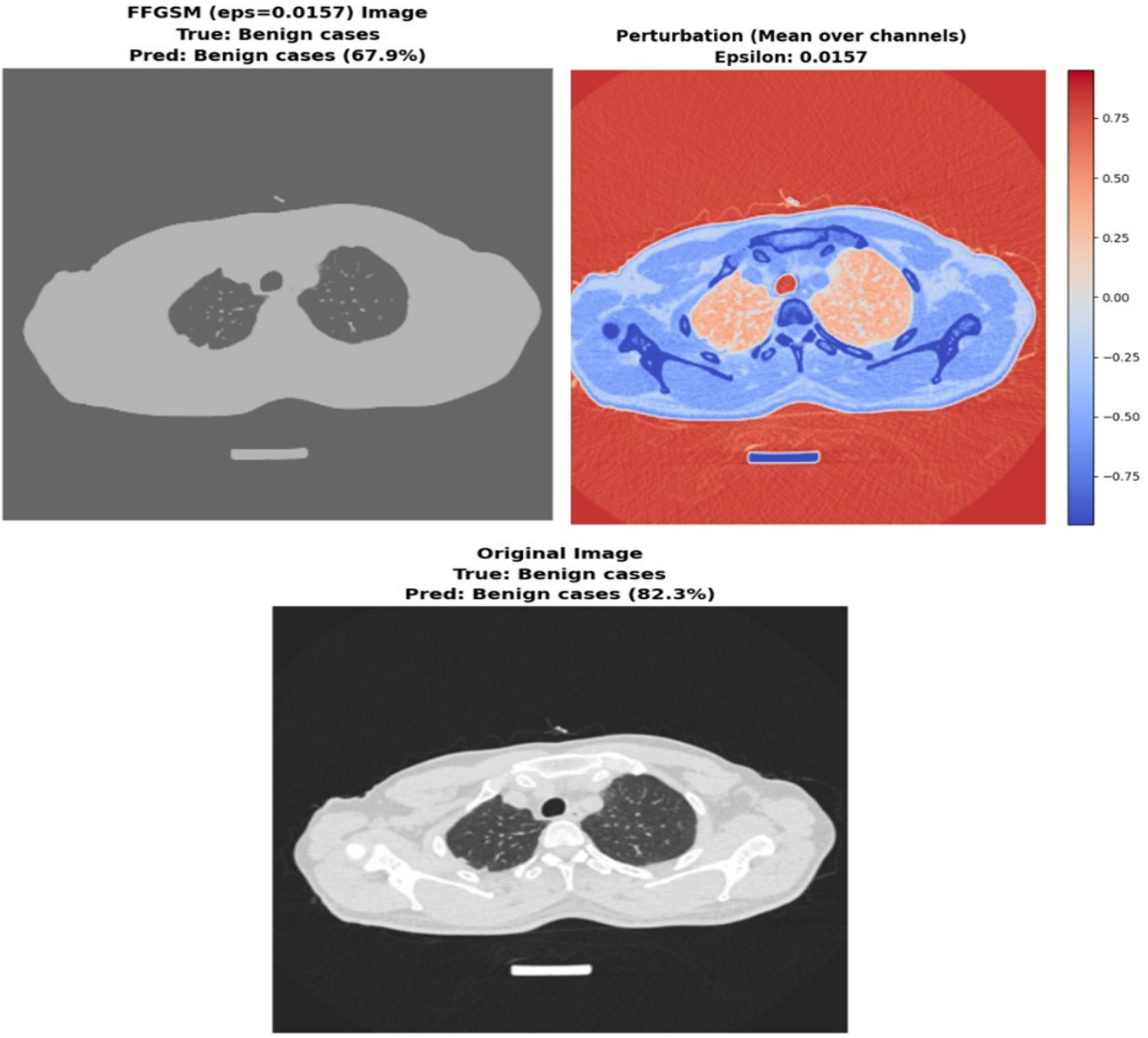

Fig. 8. Adversarial example visualization (FFGSM (eps = 0.0157))

5.2 PGDL₂ Method

PGDL₂ yielded better results on both types of test sets. Overall, the model achieved 95% accuracy and exhibited balanced F1 scores over 0.94 for each class on the clean data. When attacked under PGDL₂, the model was still able to predict correctly 86% of the time. It also had an F1 score of 0.96 for malignant cases and good accuracy in detecting other instances, with values of 0.82 for benign cases and 0.81 for normal cases.

TABLE XI. CLEAN TEST SET RESULTS (PGDL$_2$)

Class	Precision	Recall	F1-Score	Support
Benign Cases	0.90	0.98	0.94	215
Malignant Cases	1.00	0.96	0.98	237
Normal Cases	0.96	0.92	0.94	239
Accuracy			**0.95**	691
Macro avg	0.95	0.95	0.95	691
Weighted avg	0.95	0.95	0.95	691

TABLE XII. ADVERSARIAL TEST SET RESULTS (PGDL$_2$)

Class	Precision	Recall	F1-Score	Support
Benign Cases	0.70	1.00	0.82	215
Malignant Cases	1.00	0.92	0.96	237
Normal Cases	0.99	0.68	0.81	239
Accuracy			**0.86**	691
Macro avg	0.90	0.87	0.86	691
Weighted avg	0.90	0.86	0.86	691

Figure 9 shows the accuracy and loss curves associated with training and validating the proposed model. In the beginning, the model learns substantially, and its accuracy ranges from 0.55 to 0.90 in the first epochs, but then it levels out at usually between 0.92 and 0.95. Notably, the model's accuracy on the validation data starts to surpass its accuracy on the training data around epoch 5, suggesting that regularization is effective and that the gap between the two metrics remains small throughout training, hinting at limited overfitting. Because there is very little difference between the metrics, the model can carry out its task well and is unlikely to fail in new situations. This means that training past epoch 10 does not result in large benefits, as the model has grabbed most of the knowledge it can within the limits of the available information and structure.

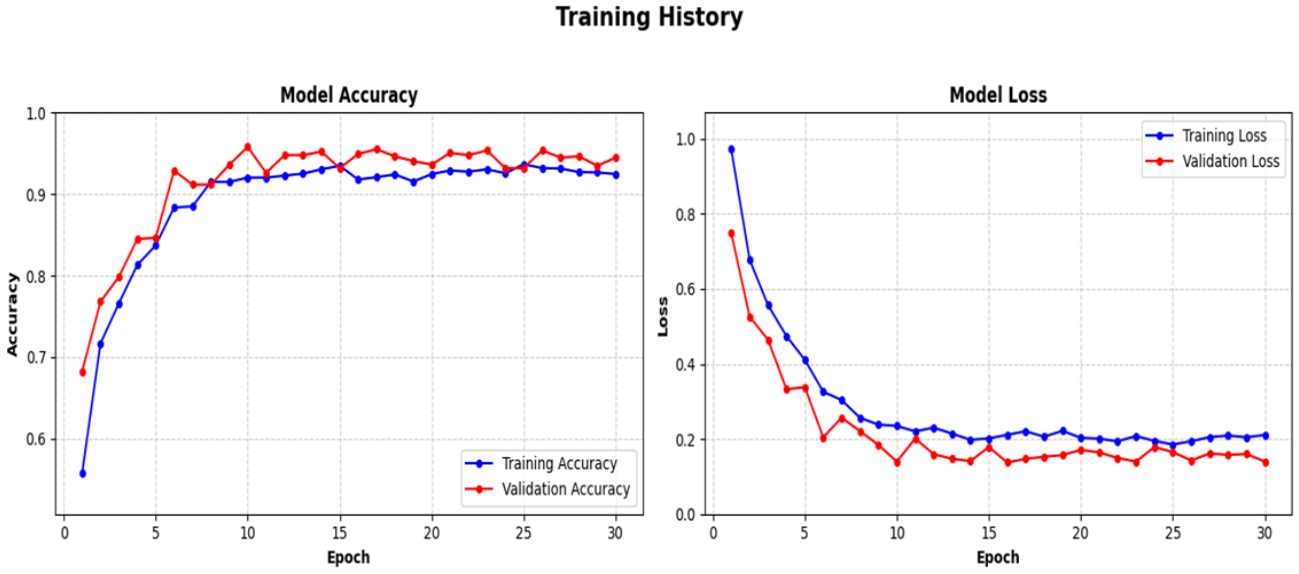

Fig. 9. Training history (PGDL$_2$)

The data in the clean confusion matrix revealed that the number of misclassifications was much lower in the proposed model than in the FFGSM. One and most of the incorrectly assigned cases were "normal cases" recognized as "benign cases" (20 instances). Even with strong attacks, the model managed to retain 86% accuracy. Importantly, "benign cases" had a perfect recall score (1.00) but had a decrease in precision from 0.90 to 0.70, so the F1 score was 0.82. Because of this, a high number of true normal cases (76) and a lower number of malignant cases (17) were not correctly identified, as seen in the adversarial confusion matrix. The fact that the training and validation curves closely match one another in Figure 9 can be used to confirm that the regularization techniques adopted, which include intensive data augmentation and dropout, prevent the model from overfitting the training subset of data.

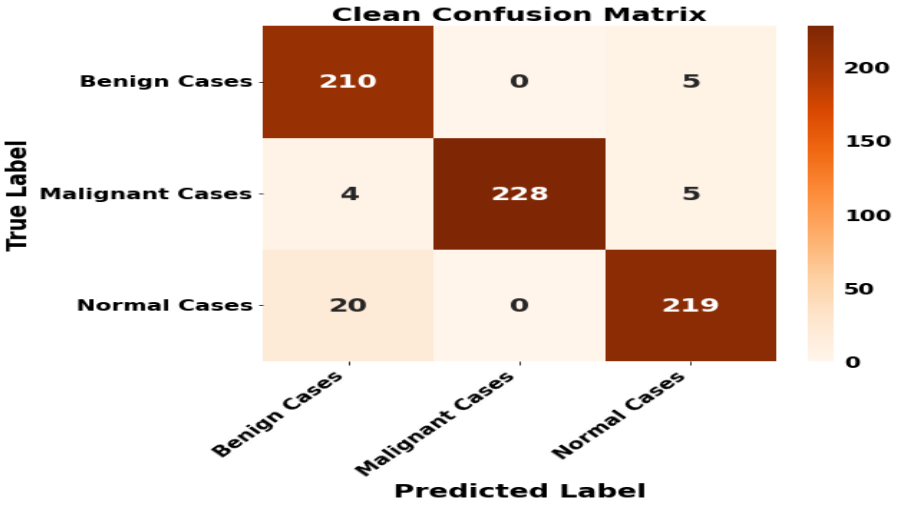

Fig. 10. Clean confusion matrix (PGDL$_2$)

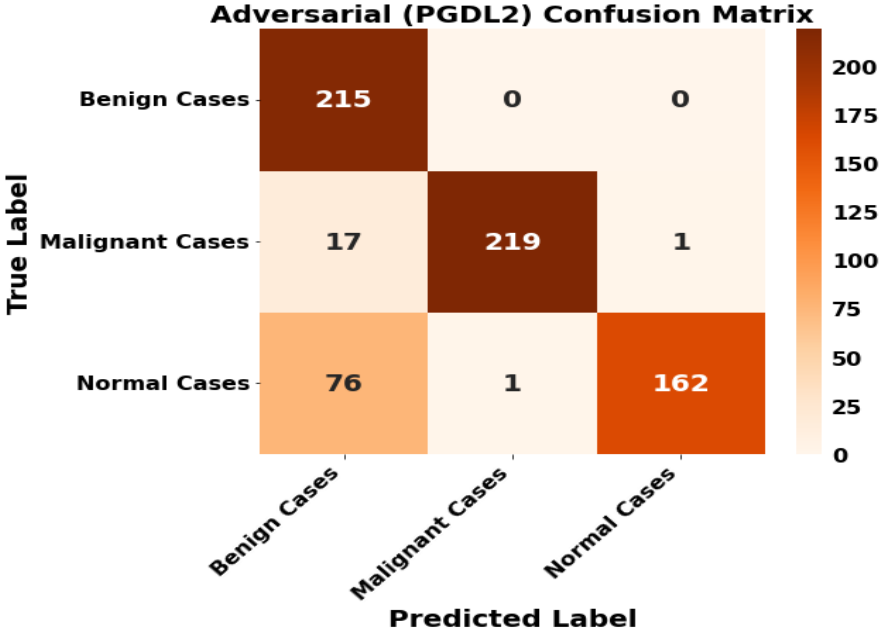

Fig. 11. Adversarial confusion matrix (PGDL$_2$)

Figure 12 shows the difference in prediction made by a medical image classification system after being hit by a projected gradient descent (PGD-L2) attack. The three-panel display shows the following: First, the chest CT scan was correctly classified as 'benign cases' satisfactorily at 97.0% accuracy. Then, by using perturbation techniques such as PGD-L2 with parameters epsilon=0.2500 and 10 optimization steps, both "benign cases" were still correctly identified but with a confidence level of up to 99.7%. Despite increasing the epsilon value used in the attack and running it several times, the PGD-L2 attack does not manage to alter the decision or confidence of the model, which proves that this model is especially robust against this attack technique. As seen from the changes in the model's parameters, the attacker focused on the part of the model connected to the lungs and surrounding parts, indicating that these regions were important for classification.

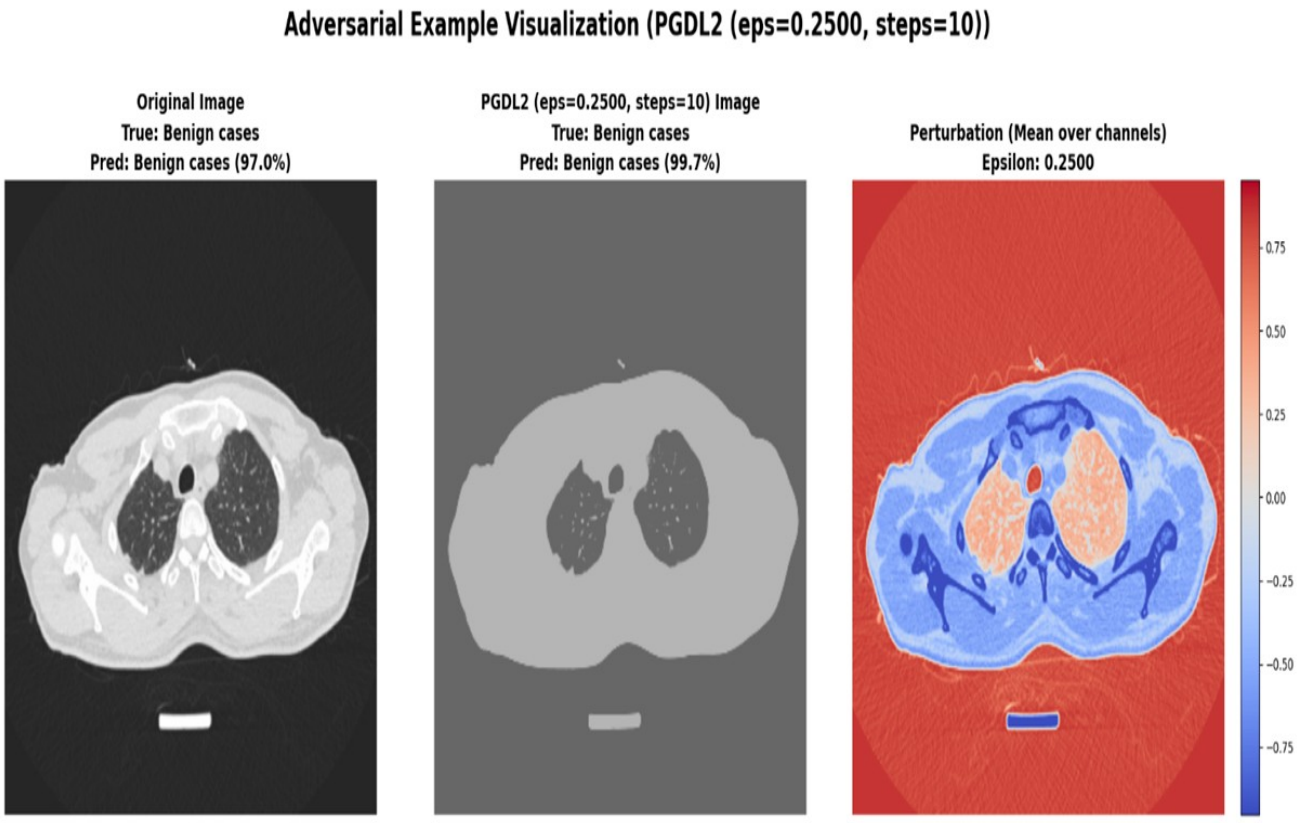

Fig. 12. Adversarial example visualization (PGDL$_2$ (maximum perturbation budget (eps)) = 0.2500, steps = 10))

5.3 Ablation Study: Impact of Attention Mechanisms on Adversarial Robustness (FFGSM Case)

To clearly understand the unique benefit of attention mechanisms to adversarial robustness, a controlled ablation procedure was conducted by comparing two models with the same architecture but different features: one has the proposed attention mechanism, and the other has not. We compared the same experimental conditions via two solutions that were created via MobileNetV2: (1) Model A, the original model focused on the baseline design in which no attention mechanism is provided, and (2) Model B, which included the proposed attention mechanism. The model adopted adversarial training via FFGSM (epsilon equals 4/255) over 30 epochs, whereas both networks shared hyperparameters. The test data were adversarially perturbed and cleaned, and an evaluation was conducted on both to measure the accuracy robustness trade-off. The empirical findings in this paper align with the whole set of findings and further confirm the importance of attention-guided feature-learning schemes as critical architecture-level solutions for engineering more reliable and robust deep neural networks. In Figure 13, a comparison of the performance of the baseline model (Model A) and the attention-enhanced model (Model B) under clean and adversarial evaluation conditions is given. The visualization also confirms the trade-off accuracy robustness, which is a characteristic of adversarial defense mechanisms.

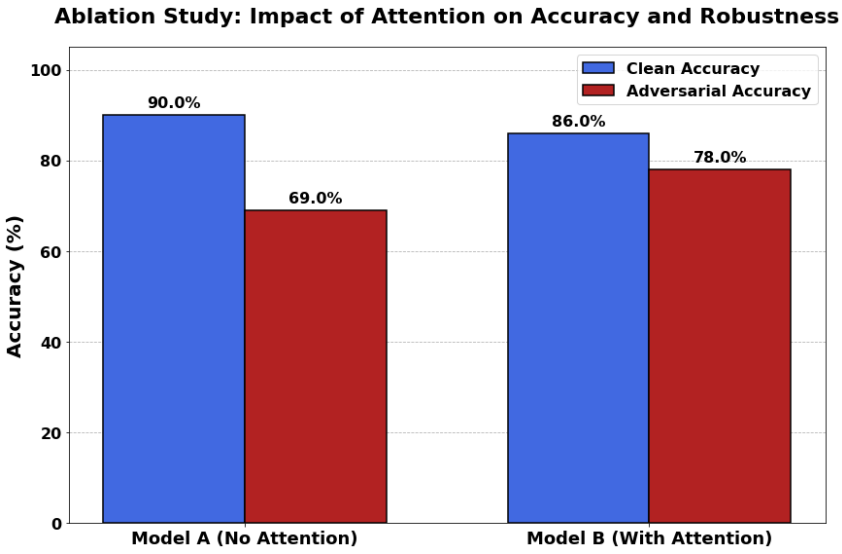

Fig. 13. Impact of attention on accuracy and robustness

Clean Performance Analysis: When samples composed entirely of clean examples were sent to either the attention-free baseline (Model A) or its attention-equipped variant (Model B), significant disparity was observed, as the clean-data accuracy of Model A was 90.0%, whereas that of Model B was 86.0%. The 4.0%-point gap between these two models shows that the attention mechanism gives rise to a built-in trade-off favouring stable feature selection over pattern completeness of the data at rest.

Comparison of Adversarial Robustness: Relative perturbations of such models shifted considerably under adversarial conditions. Model B also showed great improvement in robustness, reaching 78.0% accuracy on adversarially perturbed inputs to the opponent Model A (69.0%). This 9.0%-point gain is a 13.0% relational increase based on adversarial robustness, which is an accurate realization of the implementation of attention-guided feature learning in a test.

Trade-off Visualization: This trade-off is brought to the forefront by plotting clean accuracy, against adversarial robustness, in two dimensions. Because it is impossible to optimize both clean accuracy (x-axis) and adversarial robustness (y-axis) simultaneously, the extent to which adversarial robustness in one model can be degraded on the basis of clean accuracy varies. Even though Model A achieved a relatively small 4.0%-point improvement in clean accuracy, Model B achieved a much greater 9.0%-point improvement in adversarial robustness, which suggests a positive exchange ratio.

Stability of Performance Determination: In addition to absolute accuracy, the number also reflects significant parameters of reliability. Model A showed sharp volatility, ranging from 90.0% accuracy to 69.0% adversarially labelled inputs, a decrease of 21.0%. In comparison, however, Model B was much more resistant to sliding and slid only 8.0% down, dropping to 78.0% accuracy under this adversarial perturbation, with an accuracy of 86.0% on clean examples. The 62% decrease in the volatility of the performance figures indicates that attention-based regularization not only results in increased adversarial adversity but also greater predictability in the presence of malevolent attackers.

Conclusion: The geometric interpretation proves that the attention mechanism was indeed a robust regularizer since it compromised some minor benefits on the clean accuracy to gain enormous returns in the adversarial robustness and overall system robustness, which is quite sensible in terms of security-critical applications when robust performance under various possible attack settings is the top priority.

5.4 Discussion

5.4.1 Principal findings

In this study, the effects of deep learning models that include adversarial training mechanisms on the classification of medical images were tested, and a significant increase in model robustness was observed. Specifically, the findings showed that the PGDL$_2$ framework yielded better robustness than FFGSM-based training in both the clean and adversarial regimes:

the model, obtained with the help of $PGDL_2$, maintained higher accuracy when attacked (0.86 vs. 0.78) and had better baseline accuracy on the clean data (0.95 vs. 0.86). These results indicate that further adversarial training methods can be used to increase the quality of the generalizability and robustness of medical imaging models.

5.4.2 Interpretation of the Key Results

The significant advantage of $PGDL_2$ over FFGSM on the clean input data implies essential differences in the way the corresponding training algorithms shape the feature space. These iterative optimums of $PGDL_2$ force the model to build more robust and movable models, contributing to the subjecting of the models to a series of robust heterogeneous adversarial perturbations. This progressive improvement seems to reduce the dependency on spurious correlations and noise-sensitive traits, which might be adequate for performing innocent classification but would be too weak to collapse in an adversarial scenario. The balance (F1-scores) of these classes Benign, Malignant, and Normal, which were found by the $PGDL_2$ model, further proves that through this training regime, categories are represented more equitably and thereby avoid being biased toward those categories with clear boundaries at the expense of less distinct categories.

5.4.3 Vulnerability Patterns and Feature Dependencies

Adversarial attacks reveal characteristic patterns of confusion, which can reveal dependency within the model. These two models fail to classify most of the “Normal” cases into the “Benign” category, meaning that trained feature representations are based on subtle textural characteristics, which are easily disrupted by perturbations. These results imply that the models call attention to low-level patterns of statistics rather than the higher-level details that radiologists use to make a diagnosis. The significant decrease in benign case recall in the FFGSM model reveals the fatal flaw in the feature hierarchy that the model employs. This trend implies that its decision boundaries between benign and normal cases are vaguely defined and overly sensitive to changes in inputs, which may be caused by the lack of diversity of training data or improper regularization.

5.4.4 Implications of attention mechanism integration

The empirical evidence also suggests that the resilience of the $PGDL_2$ structure may be largely explained by the fact that it employs attention mechanisms. By limiting the attention of the model to only clinically important areas, the network tends to concentrate on discrete and localized features as opposed to diffuse and uniform representations of the image. This type of strategy is likely to produce a more structured decision-making strategy that is inherently less exposed to adversarial disturbances. Since alterations made in non-attended areas have little effect on the overall classification accuracy, the model is comparatively less prone to noise introduced in the non-attended parts of the input.

5.4.5 Clinical Implications, Model Trustworthiness and Patient Safety Considerations

The most concerning consequence of the $PGDL_2$ adversarial attack on the classification models of chest radiography is the misclassification of 17 malignant cases as benign. These misclassifications create a false-negative rate, which might be disastrous in practice: unnoticed cancer results in belated treatment and, therefore, possibly worse patient outcomes. The clinical cost of failing to detect a malignant case hence far exceeds the cost of false positives, which makes such a weakness especially problematic for autonomous diagnostic systems.

The success of the attack also proves that even the most advanced adversarial training approaches may not provide full protection against powerful attacks. In this way, the given finding points to the principal difficulty of using machine learning systems in safety-critical medical systems, where adversarial inputs could arise owing to intentional attacks, equipment failures, or unanticipated imaging scenarios.

5.4.6 Integration into the Clinical Workflow

The current study shows that although adversarial training can make models more robust, it is not enough to learn autonomously diagnostic systems completely. The existing models are hence described as decision-help tools that increase the efficiency of radiologists but do not replace human knowledge. A hybrid structure combines the pattern recognition abilities of machine learning algorithms and the critical decision-making skills of clinicians, especially when the low-confidence or unexpected attention patterns of the models are involved. An interface based on attention mechanisms acts

as a building block of explainable AI, allowing a clinician to know which regions of images affect model predictions. This kind of transparency is essential in the establishment of trust and successful integration of models in the right clinical context because radiologists are in a position to establish whether model focus aligns with easily noticed anatomical features.

5.4.7 Toward Robust Clinical Deployment

The development of clinically feasible AI systems requires stronger robustness that cannot be limited to adversarial training. With this aim, future research needs to focus on ensemble approaches by combining the outcomes of different training paradigms, together with systems that compute confidence estimates in real time and can raise alarms with potentially malicious input. Steady monitoring models that are able to identify distributional harm in clinical statistics are also necessary. The determination of these systems has hereby demanded comprehensive procedures that test model behaviour across a wide set of stress circumstances, in contrast to how they will be deployed, owing to these stress circumstances: adversarial robustness, competence under various imaging requests, various patient groups, and moisture of natural clinical apparatus fluctuation.

5.5 Limitations and Future Directions

This study is conducted by applying FFGSM and $PGDL_2$ attacks, along with a set of training schedules. The epsilon for FFGSM was set at 0.0157, whereas that for $PGDL_2$ was 0.2500; thus, it becomes difficult to compare how strong the attacks are, as they do not use the same scale. The results are taken from the median performance of three independent runs, making them somewhat reliable but not covering the whole range of possible outcomes. New research should look into more types of attacks and solutions designed to protect against them. Investigating a possible conflict between clean accuracy and how well a model resists adversarial attacks is still a main interest of researchers. The choice of each of the norms was made to maximize the stability of a particular model with its training process. The design does not allow any direct and standardized comparison of the performance impairment inflicted by each attack. In turn, the fact that the model trained with $PGDL_2$ had higher absolute accuracy on both clean and adversarial data justified its conclusion that this model is more robust than the other one but not the reduced relative decrease in performance under a unified threat model. It is therefore not unlikely that future work should attempt to re-examine these architectural contributions with a unified specification of the threat (e.g., both attacks use the L infinite norm with a specific value of epsilon) and thereby allow the comparison to be undertaken more directly. The main weakness of this research is that we relied on only one dataset. Although the IQ-OTH/NCCD dataset was suitable for the controlled experimental design of the study, it has not been proven to have high external validity to other unseen datasets collected in different clinical settings.

Experiments carried out in the future should focus on cross-dataset validation to assess the model in a domain that spans various patient groups and imaging modalities. In the context of the current research, we have only tested our robustness against white-box attacks, thus providing a basis on which to base our security level. However, a comprehensive evaluation must necessarily consider adversaries more likely to occur in practice, e.g., black-box attacks and a transferability study where the adversary is denied maximum—and possibly all—knowledge about the design structure of the model. The investigation of model resilience under such conditions is an obligatory line of additional investigation. Another weakness of this study is the lack of analysis by medical experts. Even though the quantitative results seem promising, a crucial next step will be a further reader study among radiologists to identify whether the feature attributions of the model match the human rationale behind diagnosis and to understand how applicable the tool will prove in a practical clinical environment. Finally, the considered study intentionally uses different training and testing threat models: An L- infinity norm (one for the FFGSM) and an L 2 norm (the other one used in $PGDL_2$). Such a methodological decision is bound to confuse any direct evaluation of the resulting models' performances. The advantage of the model trained with $PGDL_2$ may, however, be due to its more iterative training defense in addition to its superior geometric properties in relation to the geometry of the L2-norm-based attack on which it was trained. Although the results unambiguously demonstrate the effectiveness of each defense in its own threat model, future work must consequently examine the performance of each defense in such a unified threat model (e.g., by comparing FFGSM with $PGDL_2$ to provide a more straightforward, normalized comparison of adversarial training strategies).

6. CONCLUSION

This study used a hybrid of attention mechanisms and transfer learning combined with two adversarial training techniques, and the results showed improved effectiveness and robustness within the IQ-OTH/NCCD dataset compared with the

baseline model in recognizing lung cancer. The model was found to have 95 % accuracy when using clean test data and achieved F1-scores above 0.94 for all classes. By sharpening the channels and focusing on key regions, the attention modules steered the network to identify aspects relevant to clinical care. The technique that equally weighs losses from both the training data and adversarial data, along with proper adjustments of hyperparameters, performs well in both avoiding overfitting and making the model safer to white-box attacks. Because of this balance, the system can still yield reliable results even when exposed to tiny changes. The results of this study can assist in the design of medical tools that rely on AI so that they remain reliable in adversarial situations. This framework will be further developed to support 3D networks and examine various other attack scenarios to identify more possible security risks. Using uncertainty quantification when predicting with models and checking data from different hospitals will make the models more effective in clinical settings. Studying the connection between clean-data generalization and adversarial robustness in medical imaging AI is a key area for improving reliability.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgment

The authors would like to thank their Universites for their valuable support and for providing essential facilities for this research.

References

[1] M. A. Thanoon et al., "A review of deep learning techniques for lung cancer screening and diagnosis based on CT images," *Diagnostics*, vol. 13, no. 16, art. 2617, 2023.

[2] M. Rawashdeh, M. A. Obaidat, M. Abouali, D. E. Salhi, and K. Thakur, "An effective lung cancer diagnosis model using pre-trained CNNs," *C. – Comput. Model. Eng. Sci.*, vol. 143, no. 1, pp. 1129–1155, 2025, doi: 10.32604/cmes.2025.063765.

[3] I. S. Ahmad, J. Dai, Y. Xie, and X. Liang, "Deep learning models for CT image classification: a comprehensive literature review," *Quant. Imaging Med. Surg.*, 2024. [Online]. Available: https://qims.amegroups.org/article/view/132929

[4] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Med. Imaging*, vol. 22, no. 1, p. 69, 2022, doi: 10.1186/s12880-022-00793-7.

[5] H. Liang, M. Hu, Y. Ma, L. Yang, J. Chen, L. Lou, C. Chen, and Y. Xiao, "Performance of deep-learning solutions on lung nodule malignancy classification: a systematic review," *Life (Basel)*, vol. 13, no. 9, p. 1911, Sep. 14, 2023, doi: 10.3390/life13091911.

[6] R. Schäfer et al., "Overcoming data scarcity in biomedical imaging with a foundational multi-task model," *Nat. Comput. Sci.*, vol. 4, pp. 495–509, 2024.

[7] S. Soffer, A. Ben-Cohen, O. Shimon, M. M. Amitai, H. Greenspan, and E. Klang, "Convolutional neural networks for radiologic images: A radiologist's guide," *Radiology*, vol. 290, no. 3, pp. 590–606, 2019.

[8] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, et al., "A guide to deep learning in healthcare," *Nat. Med.*, vol. 25, no. 1, pp. 24–29, 2019.

[9] V. Sorin, Y. Barash, E. Konen, and E. Klang, "Creating artificial images for radiology applications using generative adversarial networks (GANs) – a systematic review," *Acad. Radiol.*, vol. 27, no. 8, pp. 1175–1185, 2020.

[10] Y. A. Hamad, J. Kadum, A. A. Rashid, A. H. Mohsen, and A. Safonova, "A deep learning model for segmentation of COVID-19 infections using CT scans," in *AIP Conf. Proc.*, vol. 2398, no. 1, art. no. 050005, Dec. 2022, doi: 10.1063/5.0093739.

[11] B. T. Al-Nuaimi, R. A. Suhail, S. A. Abbas, and E.-S. M. El-Kenawy, "Adaptive feature selection based on machine learning algorithms for lung tumors diagnosis and the COVID-19 index," *J. Intell. Syst. Internet Things*, vol. 11, no. 2, pp. 42–51, 2024, doi: 10.54216/JISIoT.110204.

[12] E. Klang, "Deep learning and medical imaging," *J. Thorac. Dis.*, vol. 10, no. 3, pp. 1325–1328, 2018.

[13] M. D. McCradden, E. A. Stephenson, and J. A. Anderson, "Clinical research underlies ethical integration of healthcare artificial intelligence," *Nat. Med.*, vol. 26, no. 9, pp. 1325–1326, 2020.

[14] X. Liu, S. Cruz Rivera, D. Moher, M. J. Calvert, A. K. Denniston, A.-W. Chan, et al., "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," *Nat. Med.*, vol. 26, no. 9, pp. 1364–1374, 2020.

[15] N. Bhatia, H. Trivedi, N. Safdar, and M. E. Heilbrun, "Artificial intelligence in quality improvement: reviewing uses of artificial intelligence in noninterpretative processes from clinical decision support to education and feedback," *J. Am. Coll. Radiol.*, vol. 17, no. 11, pp. 1382–1387, 2020.

[16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv* preprint arXiv:1412.6572, 2014. [Online]. Available: http://arxiv.org/abs/1412.6572

[17] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recogn.*, vol. 84, pp. 317–331, 2018.

[18] B. Badjie, J. Cecilio, and A. Casimiro, "Adversarial attacks and countermeasures on image classification-based deep learning models in autonomous driving systems: A systematic review," *ACM Comput. Surv.*, vol. 57, no. 1, Oct. 2024, doi: 10.1145/3691625.

[19] Y. L. Khaleel, M. A. Habeeb, and H. Alnabulsi, "Adversarial attacks in machine learning: Key insights and defense approaches," *Applied Data Science and Analysis*, vol. 2024, pp. 121–147, Aug. 2024, doi: 10.58496/ADSA/2024/011.

[20] M. Macas, C. Wu, and W. Fuertes, "Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems," *Expert Syst. Appl.*, vol. 238, p. 122223, 2024, doi: 10.1016/j.eswa.2023.122223.

[21] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[22] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *CoRR*, vol. abs/1810.0, 2018. [Online]. Available: http://arxiv.org/abs/1810.00069

[23] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," *CoRR*, vol. abs/2003.01690, 2020. [Online]. Available: https://arxiv.org/abs/2003.01690

[24] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, San Jose, CA, USA, 2017, pp. 39–57, doi: 10.1109/SP.2017.49.

[25] G. Papanastasiou et al., "Is attention all you need in medical image analysis? A review," *arXiv*, Jul. 2023. [Online]. Available: https://arxiv.org

[26] R. Gu et al., "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 40, no. 2, pp. 699–711, Feb. 2021, doi: 10.1109/TMI.2020.3035253.

[27] C. Hu, N. Cao, H. Zhou, and B. Guo, "Medical image classification with a hybrid SSM model based on CNN and Transformer," *Electronics*, vol. 13, no. 15, 2024, doi: 10.3390/electronics13153094.

[28] O. Bin Naeem and Y. Saleem, "CSA-Net: Channel and spatial attention-based network for mammogram and ultrasound image classification," *J. Imaging*, vol. 10, no. 10, 2024, doi: 10.3390/jimaging10100256.

[29] V. Kumar, C. Prabha, P. Sharma, et al., "Unified deep learning models for enhanced lung cancer prediction with ResNet-50–101 and EfficientNet-B3 using DICOM images," *BMC Med. Imaging*, vol. 24, no. 63, 2024. [Online]. Available: https://doi.org/10.1186/s12880-024-01241-4

[30] P. M. Bruntha, S. I. A. Pandian, J. Anitha, S. S. Abraham, and S. N. Kumar, "A novel hybridized feature extraction approach for lung nodule classification based on transfer learning technique," *J. Med. Phys.*, vol. 47, no. 1, pp. 1–9, Jan.–Mar. 2022, doi: 10.4103/jmp.jmp_61_21.

[31] M. Romero, Y. Interian, T. Solberg, and G. Valdes, "Targeted transfer learning to improve performance in small medical physics datasets," *Med. Phys.*, vol. 47, no. 12, pp. 6246–6256, Dec. 2020, doi: 10.1002/mp.14507.

[32] M. Hammad, M. ElAffendi, A. A. A. El-Latif, et al., "Explainable AI for lung cancer detection via a custom CNN on CT images," *Sci. Rep.*, vol. 15, p. 12707, 2025, doi: 10.1038/s41598-025-97645-5.

[33] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," 2018. [Online]. Available: https://arxiv.org/abs/1807.06521

[34] C. B. Vanaja and P. Prakasam, "Convolutional block attention gate-based U-Net framework for microaneurysm segmentation using retinal fundus images," *BMC Med. Imaging*, vol. 25, p. 83, 2025, doi: 10.1186/s12880-025-01625-0.

[35] Z. UrRehman, Y. Qiang, L. Wang, Y. Shi, Q. Yang, S. U. Khattak, R. Aftab, and J. Zhao, "Effective lung nodule detection using deep CNN with dual attention mechanisms," *Sci. Rep.*, vol. 14, no. 1, p. 3934, Feb. 2024, doi: 10.1038/s41598-024-51833-x.

[36] J. Zhao, L. Xie, S. Gu, et al., "Universal attention guided adversarial defense using feature pyramid and non-local mechanisms," *Sci. Rep.*, vol. 15, p. 5237, 2025, doi: 10.1038/s41598-025-89267-8.

[37] R. Paul, M. Schabath, R. Gillies, L. Hall, and D. Goldgof, "Mitigating adversarial attacks on medical image understanding systems," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1517–1521.

[38] M. Z. Joel, S. Umrao, E. Chang, R. Choi, D. X. Yang, J. S. Duncan, A. Omuro, R. Herbst, H. M. Krumholz, and S. Aneja, "Using adversarial images to assess the robustness of deep learning models trained on diagnostic images in oncology," *JCO Clin. Cancer Inform.*, vol. 6, p. e2100170, Feb. 2022, doi: 10.1200/CCI.21.00170.

[39] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 14–16, 2014.

[40] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 7–12, 2015.

[41] J. Su, D. V. Vargas, and K. Sakurai, "One-pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, pp. 828–841, 2019.

[42] D. V. Vargas and J. Su, "Understanding the one-pixel attack: propagation maps and locality analysis," in *CEUR Workshop Proc.*, vol. 2640, 2020.

[43] H. Javed, S. El-Sappagh, and T. Abuhmed, "Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications," *Artif. Intell. Rev.*, vol. 58, p. 12, 2025, doi: 10.1007/s10462-024-11005-9.

[44] N. E. H. Sayah Ben Aissa, A. Korichi, A. Lakas, C. A. Kerrache, and C. T. Calafate, "Assessing robustness to adversarial attacks in attention-based networks: Case of EEG-based motor imagery classification," *SLAS Technol.*, vol. 29, no. 4, p. 100142, 2024, doi: 10.1016/j.slast.2024.100142.

[45] alyasriy, hamdalla, and Muayed AL-Huseiny, "The IQ-OTHNCCD lung cancer dataset," *Mendeley Data*, v2, 2021, doi: 10.17632/bhmdr45bh2.2.

[46] S. Das, "IQ-OTH/NCCD Lung Cancer Dataset (Augmented)," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/subhajeetdas/iq-othnccd-lung-cancer-dataset-augmented. [Accessed: May 08, 2025].