İSTANBUL UNIVERSITY CERRAHPAŞA

ORIGINAL ARTICLE
Electrica 2025; 25: 1-20

Robustness Against Adversarial Attacks Through a Hybrid Defense Approach in Medical Imaging

M. Surekha[1], Anil Kumar Sagar[1], Vineeta Khemchandani[2]

[1]Department of Computer Science and Engineering, Sharda University, Uttar Pradesh, India
[2]Department of Computer Science and Engineering, Galgotias University, Uttar Pradesh, India

Cite this article as: M. S., A. Kumar Sagar, and V. Khemchandani, "Robustness against adversarial attacks through a hybrid defense approach in medical imaging," *Electrica*, 25, 0127, 2025. doi: 10.5152/electrica.2025.25127.

WHAT IS ALREADY KNOWN ON THIS TOPIC?

- *For medical imaging deep neural networks, attacks that exploit subtle changes in properties might cause significant issues. Artificial intelligence systems could become confused by hostile inputs that are carefully designed to disrupt them.*
- *This could cause them to make wrong predictions and compromise the accuracy of diagnostics. These challenges make it very difficult to use artificial intelligence solutions in critical healthcare settings [9]. It is essential to develop effective solutions that can withstand important events. To avoid potential dangers, this calls for strong defensive strategies.*
- *These challenges present security holes that require robust safeguards to either eliminate or significantly reduce risks [10].*

Corresponding Author:
M. Surekha

E-mail:
surekhashivakumar@gmail.com

Received: May 30, 2025
Revision Requested: June 28, 2025
Last Revision Received: July 7, 2025
Accepted: July 27, 2025
Publication Date: October 14, 2025

DOI: 10.5152/electrica.2025.25127

ABSTRACT

Medical imaging plays a vital role in clinical diagnosis, yet machine learning models used in this domain are highly vulnerable to adversarial attacks, risking misdiagnosis, which could lead to incorrect diagnoses. Using two benchmark datasets, Pneumonia and BreakHis images, this study assesses the resilience of well-known deep learning architectures, such as VGG16, ResNet50, InceptionV3, and VGG19, against adversarial attacks. For better model robustness, a hybrid defense strategy is suggested that combines adversarial training with autoencoder-based preprocessing. Results indicate that adversarial attacks degrade base model performance, but the hybrid approach enhances accuracy, precision, recall, F1 score, and area under the curve (AUC) score. Autoencoders suit BreakHis data, while adversarial training better supports Pneumonia dataset robustness. Statistical analysis and evaluation metrics such as accuracy, precision, recall, F1 Score, and AUC score on the basis of confusion matrices, and its comparison analysis visualizations support the superiority of the hybrid strategy in improving classification reliability across varying attack types. In situations with limited resources, autoencoders offer a lightweight additional defense, and adversarial training is effective on all architectures. The results demonstrate the critical need for integrated defenses in ensuring trustworthy artificial intelligence–driven medical diagnosis.

Index Terms—Adversarial attack, adversarial defense, adversarial training, autoencoder, deep neural network, medical imaging

I. INTRODUCTION

Deep learning (DL) in medical image (MI) analysis has opened new avenues in healthcare and holds great potential for accurate diagnosis and treatment, such as skin-lesion classification. Indeed, these advanced models have shown impressive efficiency using artificial intelligence (AI) in detecting complicated MIs [1], providing researchers and healthcare practitioners with a new tool. However, counter-attacks are a weakness of deep learning methods, requiring analysis of the adversary's training configuration. Even though deep learning is very promising in many areas, it presents great challenges for deep neural networks (DNNs) because they are subject to various types of adversaries [2]. Especially in safety-critical applications such as MI analysis [3-6], such paradigms, incompatible as they appear initially, would result in a weak, barely noticeable hybrid model, thus implicitly rejecting the evidence. This is a significant challenge of DNN deployment. The accuracy of DNNs in clinical diagnosis is greatly influenced by poor samples of diagnosis, which can potentially lead to misdiagnosis, insurance fraud, and reduced confidence in AI in medicine. It is difficult to defend against such attacks; there are also inconsistencies and frequent revisions in diagnoses that create problems in the way the security apparatus operates. Recent studies have scientifically demonstrated that working on the computer improves cognitive capacity. In a number of situations, including white box (WBA) and black box attacks (BBA), diagnostic models may be susceptible to attacks [4,5,6]. Deep neural networks are widely used in medical imaging and are susceptible to grave threats posed by small distortions to them. Artificial intelligence systems may be thrown off by the byproduct of the deliberate creation of these adversarial inputs, which are designed to lightly perturb.

WHAT THIS STUDY ADDS ON THIS TOPIC?

- *This study evaluates the vulnerability of widely used deep learning architectures, VGG16, VGG19, ResNet50, and InceptionV3, to adversarial attacks in medical imaging using two benchmark datasets: Pneumonia and BreakHis. It applies four prominent attack algorithms, FGSM, projected gradient descent, basic iterative method, and MIFGSM, on these architectures to systematically assess performance degradation under adversarial influence.*
- *To improve model robustness, the study proposes a hybrid defense strategy that combines adversarial training and autoencoder-based preprocessing. Results indicate that the hybrid approach significantly restores classification performance metrics (accuracy, precision, recall, F1 score, and area under the curve score) across all models.*
- *The study highlights that autoencoder-based defense is more effective on BreakHis histopathological data, whereas adversarial training is more robust for pneumonia chest X-ray data.*
- *It emphasizes the importance of incorporating lightweight and scalable defense strategies to safeguard deep learning applications in medical diagnostics under resource-constrained conditions.*

This may cause incorrect forecasts and impair the accuracy of detection. These are all obstacles to the adoption of AI solutions for use in life-dependent healthcare settings. Risks need to be curtailed, though the vulnerabilities have to become a strong defense around. Such adversarial inputs are designed to introduce small distortions that could fool the AI models and cause their prediction errors and diagnostic accuracy to degrade. The use of AI is widely impeded by these issues. In emergency care situations, it is critical to develop efficient services that are resilient against such serious incidents. This demands the development of strong defense systems to prevent such threats.

This work demonstrates weaknesses of AI-based medical imaging systems against adversarial attacks like FGSM or one-pixel-methods observation, with significantly reduced prediction accuracies of CNNs. Applications of DL are widely used in the area of medical diagnostics, such as drug discovery and imaging methods like magnetic resonance imaging, computed tomography, and Positron Emission Tomography (PET). Adversarial perturbations are barely noticeable at the visual level but significantly distort the output of segmentation and detection tasks. As such, robust defenses are required to safeguard the accuracy and reliability of AI in healthcare.

The major contributions of this study are that a hybrid adversarial defense framework was developed by integrating adversarial training with autoencoder-based image reconstruction to enhance the robustness of deep learning models in medical imaging. The framework was applied to two clinically significant datasets: one is Pneumonia chest X-rays, and the other is BreakHis histopathology images, and they were evaluated using four convolutional neural network architectures: VGG16, VGG19, ResNet50, and InceptionV3. To simulate real-world threat scenarios, four gradient-based adversarial attacks (FGSM, projected gradient descent [PGD], basic iterative method [BIM], and MIFGSM) were employed. The results demonstrated that adversarial attacks significantly compromise the performance of deep learning models in medical diagnostics, particularly on Pneumonia chest X-ray images. VGG16 and VGG19 showed greater resilience to gradient-based adversarial attacks (with noise $\varepsilon = 0.03$), establishing them as robust CNN architectures. In contrast, ResNet50 and InceptionV3 exhibited notable performance degradation under attacks; however, adversarial training substantially improved their robustness, with InceptionV3 particularly benefiting from this strategy, making it a reliable and architecture-independent defense. For the BreakHis histopathology dataset, VGG16, VGG19, and ResNet50 showed reduced performance with autoencoder defenses, although ResNet50 maintained balanced resilience when combined with an autoencoder. In this context, the autoencoder defense outperformed adversarial training, offering a practical and computationally efficient solution that is particularly effective under adversarial stress. Therefore, autoencoders are recommended for safety-critical applications such as histopathology image classification, especially in resource-constrained environments. This study's model is validated empirically on several types of attacks and different datasets, and balances between robustness, interpretability, and computational efficiency for the practicality of supporting real-world usage in clinical workflows.

II. LITERATURE REVIEW

Recent work has analyzed the vulnerabilities of DL systems in medical imaging to hostile attacks and the effectiveness of defenses extensively. [7,8] stress the vulnerability of such systems to hostile perturbations and point out the absence of scalable solutions attacking the robustness in the cross-modality sense to varied imaging protocols. However, they do not really generalize as generic solutions that would apply to any kind of data. Similarly, [2,9] provide a valuable outline of existing defenses, notably in terms of their susceptibility to advanced and/or universal attacks, but tend not to be empirically validated in a range of clinical situations. This illustrates a real need for such domain-specific but transportable defense models with real-world assessments. Their results emphasize the importance of introspective, personalized AI models that are explainable and can generalize across datasets. [10] Addresses the degradation of diagnostic models under adversarial perturbations and out-of-distribution shifts in clinical images.[11] Proposes adversarial training guided by clinical semantics, improving robustness in real-world radiology scenarios. [12] Introduces a defense framework incorporating uncertainty modeling to resist sophisticated adversarial examples. [13] Highlights the role of uncertainty estimation and ensemble learning in the real-world deployment of robust medical AI systems.

Meanwhile, [14,15] expose critical reliability issues in transfer learning models under adversarial conditions, revealing significant performance drops but offering limited insight into feasible countermeasures.[16] propose innovative strategies such as block switching autoencoders

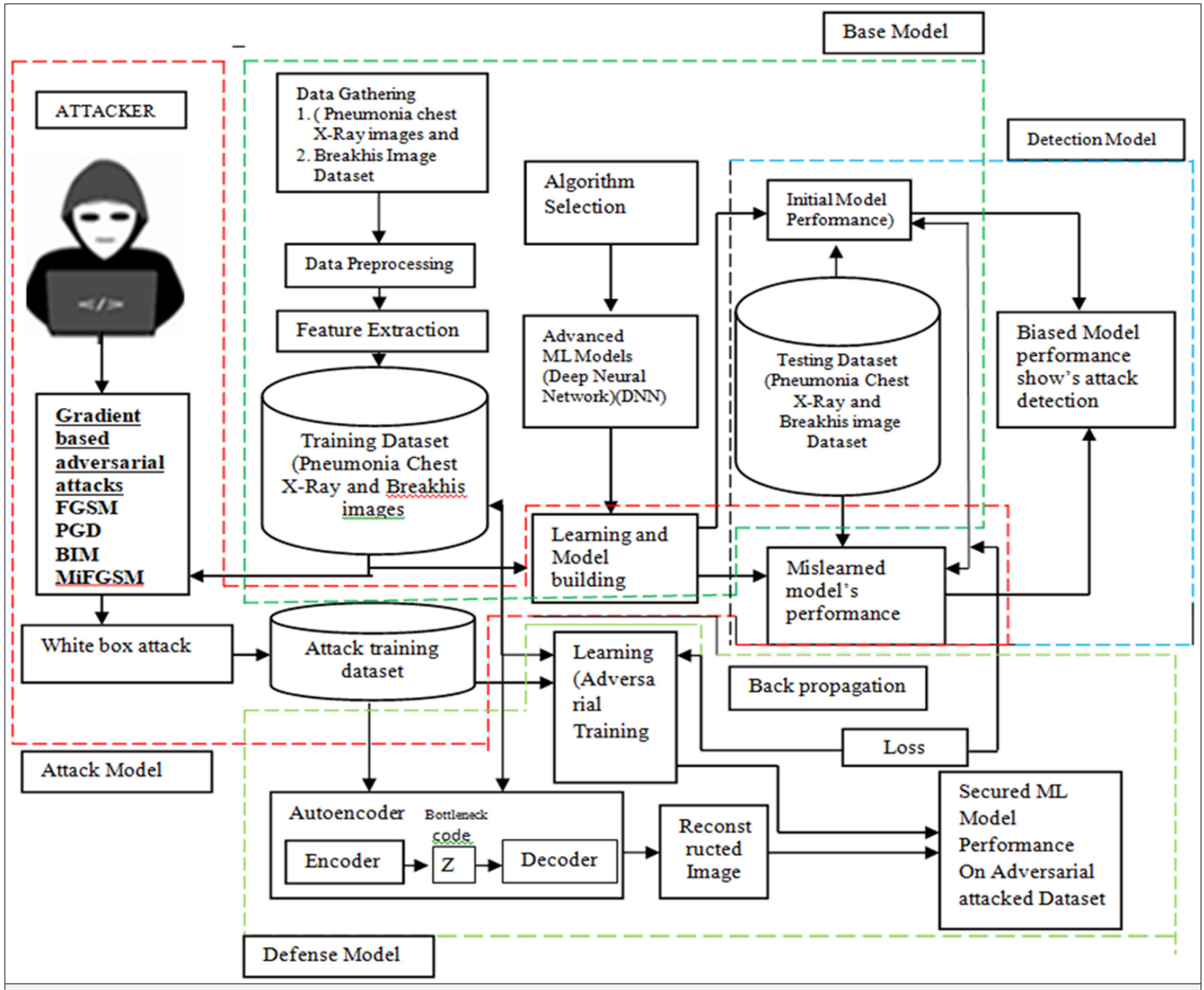

Fig. 1. Process flow of adversarial attack and defense model using a classification model.

and purification networks, respectively. Although successful, these methods are costly in terms of computation and have not been evaluated on a large-scale real-world application. This is indicative of a more general trade-off observed across numerous studies in achieving computational efficiency as well as robustness for implementing the models in resource-limited clinical settings. Based on these gaps, this paper aims to introduce a lightweight adversarial defense framework that is specialized for MI classification. The model is validated empirically on several types of attacks and different datasets, and balances between robustness, interpretability, and computational efficiency for the practicality of supporting real-world usage in clinical workflows.

III. METHODOLOGY

The method is a full adversarial defense framework against the MI classification based on DL. The pipeline starts with gathering the pneumonia chest X-ray and BreakHis (The Breast Cancer Histopathological Image classification) images, preprocessing the data, and attribute extraction. A chosen DNN is then trained on the dataset in order to evaluate the performance of the initial model on clean data.

In Fig. 1, the proposed methodology, a comprehensive adversarial defense framework for MI classification using DL, is presented. The system begins with the collection of chest X-ray (pneumonia) and BreakHis histopathological images, followed by data preprocessing and feature extraction. A selected DNN is then trained on the dataset to assess the initial model performance on clean data. An attacker model generates adversarial examples by applying gradient-based white box attacks (FGSM, PGD, BIM, MIFGSM), which perturb the training data by introducing noise ($\varepsilon = 0.03$). These attacks result in a mislearned model that misclassifies inputs and shows degraded performance on the test set. A detection mechanism monitors such biased outputs to identify model vulnerabilities. Two defense strategies are employed. First, adversarial training incorporates both clean and adversarial samples into the training process, thereby enhancing robustness.

TABLE I. CHEST X-RAY PNEUMONIA DATASET SAMPLE SPLIT

Class	Data Set Taken for Base Model	For Attack	Training Dataset	Testing +Validation	Total
Normal	1000	100	1400	300 + 300	2000
Pneumonia	1000	100			
Total	2000	200			

Second, an autoencoder (comprising encoder → bottleneck → decoder) reconstructs input images to eliminate adversarial perturbations. The adversarial training models are optimized using backpropagation and loss minimization techniques. The final secured machine learning model is evaluated on adversarial samples to verify its ability to recover performance. This approach ensures enhanced robustness, reliability, and generalization of Machine Learning (ML) models against hostile attacks in the MI domain. The outcomes of the study, along with evaluation and comparative analysis, are presented in the corresponding Tables I and II.

In adversarial attacks, they are classified as location-specific attacks, knowledge-specific attacks, and intent-specific attacks. Knowledge-specific attacks are again majorly subclassified as black box and white box attacks. For this research work, the knowledge-specific attacks of white box attacks have been selected to work with the FGSM, PGD, BIM, and MiFGSM attack models.

A. White Box Attack and Black Box Attack

White box attack (WBA) refers to the case where adversarial examples are created with complete access to a model's structure, data, and parameters, allowing adversaries to generate accurate and effective adversarial inputs. These attacks are essential for evaluating the robustness of ML/DL models, especially in sensitive domains such as medical imaging. Studies by [17,18] proved that very small perturbations can generate misclassifications on deep networks using MIs.

In BBA models, although the attacker cannot know any information about the prediction with a model, the attacker can nevertheless fool the model into making a wrong prediction by only asking some query-based and transferable adversarial questions, e.g., architecture, weights, training set [19,20]. This is a serious threat to medical imaging systems in which models are kept secret to guarantee patients' data privacy.

1) Fast Gradient Sign Method:

A BBA, such as FGSM, occurs when the attacker has limited rights to use to the ML algorithm under assault. This function implements it. It takes three inputs: the original image, the loss gradient with respect to the image, and epsilon (the attack strength). A perturbed picture is created by changing each pixel in accordance with the direction and magnitude of the gradients (1) [21-23].

$$\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + \boldsymbol{x} + \in * \text{sign}(\nabla_x J(\theta, \text{x}, \text{y})) \quad (1)$$

Where, •*xadv* is the adversarial instance created by the FGSM attack.

- The initial input example is x.
- The perturbation's magnitude is represented by ϵ.
- The derivative of the loss function J with respect to the input x is represented as the gradient $\nabla_x \boldsymbol{J}(\theta, x, y)$. In this case, y is the actual label, while θ stands for the model parameters.

2) Projected Gradient Descent:

Projected gradient descent is an example of a white box attack, meaning that attackers have access to the machine learning algorithm or model parameters, hyperparameters, architecture, and weights. PGD, an extension of BIM (and FGSM), employs a projection function Π to project the adversarial example back onto the e-ball of x [24-26].

Unlike BIM, PGD utilizes a random initialization method for the variable x. This is achieved by introducing random noise from a uniform distribution with values within the specified range (–€).

$$x_{adv}^{(x+1)} = \prod_{x+S}\left(x_{adv}^{(t)} + \alpha \cdot sign\left(\nabla_x J\left(\theta, x_{adv}^{(x+1)}, y\right)\right)\right) \quad (2)$$

where:

- The variable $x_{adv}^{(t)}$ represents the hostile instance at iteration t.
- α is the magnitude of the step taken during each iteration, often known as the learning rate.
- The *sign* $\left(\nabla_\nabla J\left(x_{adv}^{(x+1)}, y\right)\right)$ represents the direction of change in the loss function J when considering the adversarial sample $x_{adv}^{(t)}$.
- $\prod_{x+S}$ is the projection operator that projects the perturbed example back onto the set *x*+*S* to ensure the perturbation stays within the allowed set S (often an ℓp norm ball around the original example x).

3) Basic Iterative Method:

The BIM [27, 28] is an extension of FGSM that involves iteratively applying gradient updates with a small step size α.

$$x_{adv}^{(t+1)} = Clip_{x,\in}\left(x_{adv}^{(t)} + \alpha \cdot sign\left(\nabla_x J\left(\theta, x_{adv}^{(t+1)}, y\right)\right)\right) \quad (3)$$

where:

- x_{adv}^{t} Represents the hostile sample at iteration t.
- α represents the magnitude of the increment for each iteration.
- The sign $\left(\nabla_x J\left(\theta, x_{adv}^{(t+1)}, y\right)\right)$

represents the direction of change in the loss function J when considering the adversarial sample.

TABLE II. BREAKHIS DATASET SAMPLE SPLIT

Class	Data Set Taken for Base Model	For Attack	Training Dataset	Testing +Validation	Total
Benign	1000	100	1400	300 + 300	2000
Malignant	1000	100			
Total	2000	200			

- The gradient of the loss function J with respect to the adversarial case may be determined by examining its example $x_{adv}^{(t)}$, is sign $\left(\nabla_x J\left(\theta, x_{adv}^{(t+1)}, y\right)\right)$.
- $Clip_{x,\epsilon}$ is a clipping function that ensures the adversarial sample $x_{adv}^{(t)}$, stays within an ϵ-ball around the original input x, i.e., $x_{adv} - x \infty \leq \epsilon$

The algorithm can either set α=T, where Ti is the number of iterations, or $x_{adv}^{(t)}$, after each update, confine the created adversarial instances to the -ball of x true. It has been demonstrated that BIM generates far more potent WBAs than FGSM, but at the expense of low transferability.

4) Momentum Iterative Fast Gradient Sign Method:
The transferability of adversarial situations is enhanced by the use of the MIFGSM [29, 19]. This method suggests a relationship between perturbation in each epoch and the gradient that was previously determined, as well as the gradient that is present now.

B. Adversarial Defense
Broadly speaking, it describes the techniques and strategies employed to protect medical imaging systems, particularly deep learning models, from hostile assaults. Protecting medical imaging's accuracy and integrity is the primary goal of medical defensive tactics. These defense strategies aim to strengthen the overall security and reliability of medical imaging models by lessening the impact of different attacks. This study employed hybrid defense strategies; one of them is adversarial training, and another one is an autoencoder against adversarial attacks.

Fig. 2a outlines an adversarial training setup, where input data X is perturbed to generate hostile instances. Both original and hostile data are used to retrain the model, enhancing its robustness. The goal is to make the model generalize well on perturbed inputs while maintaining accuracy on clean data [30]. The reconstructed output is compared against the ground truth to assess defense effectiveness. In this study, Fig. 2a is a schematic of an adversarial training-based defense framework. The model is exposed to both clean and adversarially perturbed inputs (e.g., via FGSM, PGD, BIM, and MIFGSM), allowing it to learn robust feature representations. By retraining on these examples, the model improves its ability to maintain accuracy and classification confidence even under adversarial threat.

Fig. 2b illustrates an autoencoder that compresses input X into a latent representation Z using an encoder gθ, and reconstructs it to X′ via a decoder fϕ. The goal is for X′ to closely match X, ideally achieving X=X′. This process effectively filters out adversarial noise while retaining essential features. Sample histopathological images

(True: BENIGN) before and after reconstruction confirm output fidelity [31]. To enhance model resilience, adversarial training and autoencoder were implemented by augmenting the training dataset with perturbations generated through multiple white box attacks. As illustrated in Fig. 1, the model was retrained on both clean and adversarial examples, reconstructing the original data and enabling it to generalize better under adversarial stress. This strategy yielded notable improvements in accuracy, recall, and robustness metrics across all CNN architectures.

C. Classification Model and Datasets
Pneumonia is one of the most serious, deadly, and infectious infections. Pneumonia or any chest X-ray medical pictures require categorization to be assigned to particular groups [32, 33]. In this study, the dataset was trained to classify pneumonia and normal chest X-ray images using Inception-V3, ResNet50, VGG-16, and VGG-19. These DNN models are used to identify pneumonia and determine whether a person has pneumonia or not. This model allows for the inexpensive and very accurate identification of pneumonia in a short period of time. This measure aids in mitigating the transmission of a certain entity. This approach might assist in alleviating the dependability and interpretability issues that arise when dealing with MIs. Pneumonia is a condition that can impact both lungs simultaneously and affects the small air sacs in the lungs called alveoli [32]. Details about the balanced pneumonia dataset image sample split have been given in Table I. Image dataset samples have been shown in Fig. 3a, and it has been taken from Kaggle.

Fig. 3b shows BreakHis dataset, which contains 2000 microscopic images of breast tumor tissues, balanced between benign and malignant categories. Captured at four magnification levels (40×, 100×, 200×, and 400×), it supports histopathological cancer classification as shown in Fig. 3b. The dataset was collected from 82 patients using biopsy samples and is publicly available for research from Kaggle.

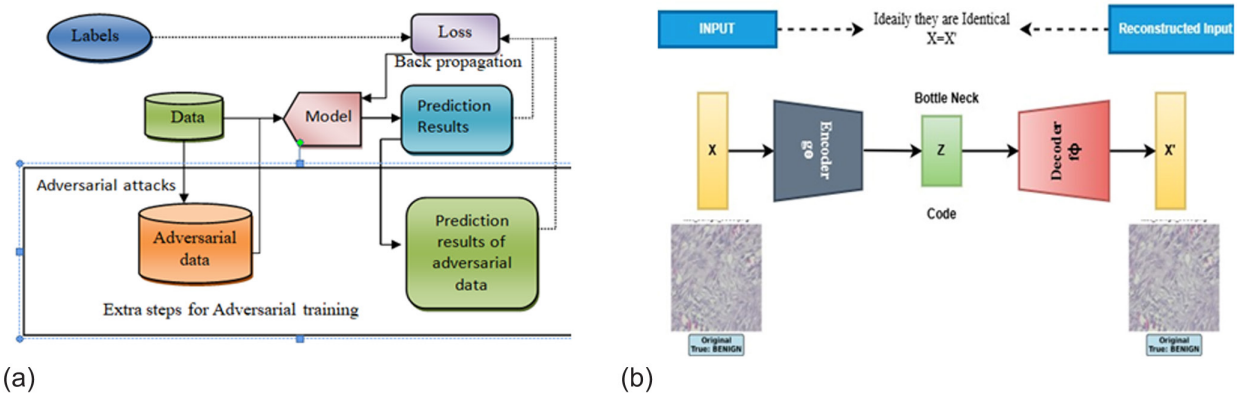

Fig. 2. (a) Process of adversarial training defense method. (b) Process of autoencoder defense method.

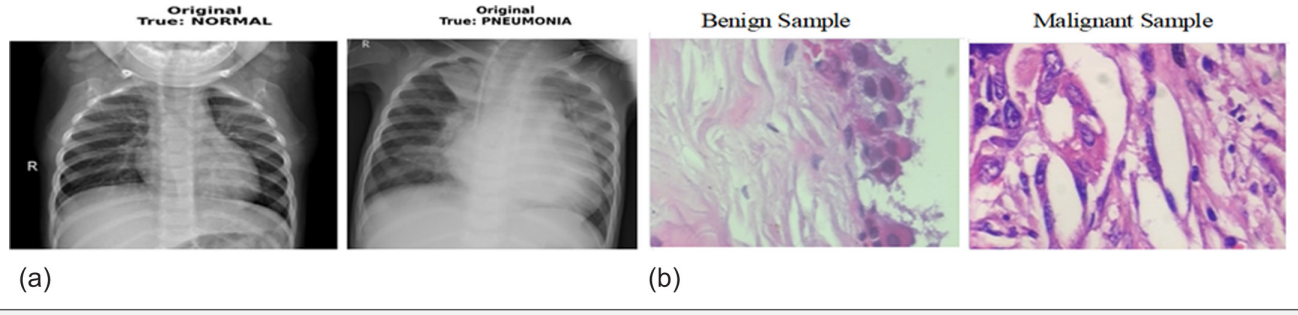

Fig. 3. (a) Chest X-ray pneumonia dataset images. (b) BreakHis histopathological image dataset.

1) Convolutional Neural Network Architecture:
Fig. 4 shows a convolutional neural network (CNN) source pixels filtering through a convolution filter to reach the destination pixel. In this work, a few structural designs used have been explained below.

The study trained CNN models using predefined hyperparameters, with the learning rate set to 0.001 and optimized using the Adam optimizer. A batch size of 32 was used. The number of layers depended on the chosen base architecture—for example, VGG16 includes 13 convolutional layers and 3 dense layers. A total of 20 epochs were used to train and validate each model for pneumonia dataset classification. The approximate training time for each architecture was as follows: VGG16/VGG19 required about 2.5 hours, ResNet50 took around 3 hours, and InceptionV3 required approximately 4 hours to complete training on the image datasets. These hyperparameters provided a stable and consistent training environment for evaluating defense model performance across all CNN backbones. For adversarial training, the perturbation magnitude (ε) was tuned in the range [0.01–0.05]**,** and $\varepsilon = 0.03$ was selected as the optimal value based on empirical performance across both datasets. For the autoencoder**,** a bottleneck structure of 128 dimensions was selected after evaluating 64, 128, and 256 units. The optimizer used was Adam, with a learning rate of 0.001.

a) Inception-V3: Inception-V3 is a deep CNN structural design that builds on the concepts introduced in its predecessor, GoogLeNet (Inception-V1). It introduces several improvements, including factorized convolutions, which reduce the computational complexity and enhance the model's efficiency. Fine-tuned on the gradients during training [34-37]. Inception-V3 is trained on the ChestX-ray14 dataset to adapt its parameters for pneumonia detection. Its depth and architecture allow it to effectively learn and distinguish between normal and pneumonia-affected lung patterns, leveraging its ability to process multi-scale features efficiently. The study used 48 layers in the InceptionV3 architecture.

b) ResNet50: ResNet50 is part of the residual network family, which is known for introducing residual learning. This network is a 50-layer deep network that has been specially designed to cope with the difficulties posed when training very deep networks, including the vanishing gradient problem [34-37].

Application to pneumonia detection: This approach is applicable to detecting pneumonia in medical photographs with the ResNet50 architecture, as it has the ability to train deep networks without degradation and thus capture complex patterns in an effective manner. Chest X-ray14 is a publicly available dataset and can be used to fine-tune the ResNet50 model to make it very good at finding even slight indications of pneumonia, improving diagnosis.

c) VGG-19: It has a total of 19 layers, including 16 convolutional layers and 3 fully connected layers with small 3×3 filters throughout the network. Its simple architecture, depth, and homogeneity make VGG19 famous and efficient for image classification tasks. It was pretrained on ImageNet and fine-tunes very well. While computationally expensive, VGG19 is a popular choice in MI and computer vision research because of its powerful feature extraction ability [36, 37].

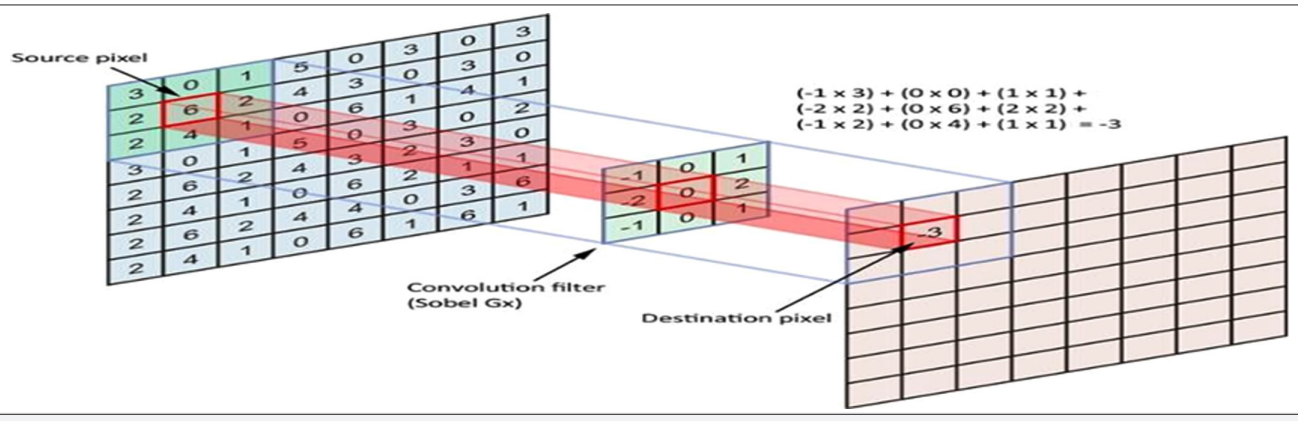

Fig. 4. Convolutional neural network.

d) VGG-16: VGG-16 is well known for its uniform structure and ease of use in CNNs. The network has a total of 16 layers (13 convolutional +3 dense layer), comprising both fully connected and convolutional layers. The model VGG-16 uses a straightforward approach to demonstrate the significance of depth [38, 34,37].

This system classifies chest X-rays by taking the images as input to these pretrained models, which now apply an off-performed stage neural network that helps in learning the field features to decide whether the image shows any signs of pneumonia. It can be useful to merge these models and perform an inclusive study, taking advantage of their strong points to obtain better accuracy. Each model provides a final decision, perhaps as part of a committee in which the decisions are merged to make a more accurate decision. The application of those advanced models enables the system to rapidly identify pneumonia issues, and it can do so robustly for the early prevention of disease spread.

The classification model utilized in this system comprises a pipeline that processes chest X-ray images through the pretrained CNNs, Inception-V3, ResNet50, VGG-16, and VGG-16, VGG-19. The workflow involves several key steps:

Data preprocessing: This includes resizing the images to the input size required by the models (e.g., 224 × 224 for VGG-16 and ResNet50, 299 × 299 for Inception-V3).

To improve model generalization, normalize the values of the pixels and add them to the dataset.

D. Performance Evaluation Matrices

Table III presents the machine learning evaluation metrics used for performance assessment and comparative analysis of the results. All calculations and measures are based on the confusion matrix, including true positive, false positive, true negative, and false negative.

IV. FINDINGS AND DISCUSSIONS

The study on both pneumonia chest X-ray and BreakHis histopathological image datasets demonstrates that medical imaging models are highly vulnerable to hostile attacks, often leading to critical misclassifications. Across both datasets, base models experience sharp declines in accuracy, precision, recall, and AUC under FGSM, PGD, Momentum-based Fast Gradient Sign Method (MFGSM), and BIM attacks. A hybrid defense approach, combining adversarial training with autoencoder-based preprocessing, significantly enhances model robustness and generalization. Adversarial training ensures strong resistance by directly adapting the model to perturbations, while autoencoders effectively denoise inputs, especially in computationally constrained settings. This combined strategy consistently improves classification reliability across architectures like VGG, ResNet, and Inception, as can be seen in Table IV and Figs [5-7, 14 -18].

However, adversarial training substantially improved their robustness, with InceptionV3 particularly benefiting from this strategy, making it a reliable and architecture-independent defense. For the BreakHis histopathology dataset, VGG16, VGG19, and ResNet50 showed reduced performance with autoencoder defenses, as can be seen in Table V and Figs. [5-7, 14-18], although ResNet50 maintained balanced resilience when combined with an autoencoder. In this context, the autoencoder defense outperformed adversarial training, offering a practical and computationally efficient solution that is particularly effective under adversarial stress. Therefore, autoencoders are recommended for safety-critical applications such as histopathology image classification, especially in resource-constrained environments.

A. Pneumonia Chest X-Ray Images

In Table IV, all base models (VGG16, VGG19, ResNet50, InceptionV3) show strong performance on clean data but experience substantial accuracy degradation under adversarial attacks like FGSM, PGD, BIM, and MIFGSM, as can be seen in Fig. 6. In particular, InceptionV3's accuracy drops below 15% under attack, indicating extreme vulnerability without any defense.

Adversarial training significantly enhances model resilience, with VGG19 and VGG16 reaching accuracies of 85.7% and 75.5% respectively, under attack conditions. Autoencoder defense moderately improves performance, especially in InceptionV3, but is generally less effective than adversarial training for VGG and ResNet architectures. Attack success rates are consistently lower in adversarially trained models, confirming their improved robustness. The combination of adversarial training and an autoencoder achieves the best balance between defense strength and computational feasibility. Overall,

TABLE III. MACHINE LEARNING MODELS EVALUATION MATRICES

Metric	Description	Formula	It Has Been Used
Accuracy [39]	Proportion of correct predictions.	$\text{Accuracy} = \frac{TP + FP + FN + TN}{TP + TN}$	Used when class distribution is balanced.
Precision [40]	Fraction of true positives among predicted positives.	$\text{Precision} = \frac{TP + FP}{TP}$	Important in cases with high false-positive cost.
Recall (Sensitivity) [40]	Fraction of true positives correctly identified.	$\text{Recall} = \frac{TP + FN}{TP}$	Important when false negatives are costly.
F1-Score [41]	Harmonic mean of precision and recall.	$F1 = 2\frac{Precision + Recal}{Precision \cdot Recall}$	Use with imbalanced datasets.
ROC-AUC (area under the receiver operating characteristic curve) [40, 41]	Measures trade-off between true positive rate against the false positive rate.	Area under the ROC curve.	For probabilistic classifiers and threshold tuning.

TABLE IV. PERFORMANCE EVALUATION OF ATTACKS AND DEFENSES MODEL USING PNEUMONIA CHEST X-RAY IMAGES USING CONVOLUTIONAL NEURAL NETWORK ARCHITECTURES

Model	Evaluation Type	Accuracy	Precision	Recall	AUC Score	F1 Score
INCEPTIONV3_Adversarial_Training	BIM_Attack	0.36	0.4068	0.4528	0.265	0.4286
INCEPTIONV3_Adversarial_Training	Clean	0.89	0.85	0.9623	0.9558	0.9027
INCEPTIONV3_Adversarial_Training	FGSM_Attack	0.83	0.8	0.9057	0.9081	0.8496
INCEPTIONV3_Adversarial_Training	MFGSM_Attack	0.08	0.0465	0.0377	0.0466	0.0417
INCEPTIONV3_Adversarial_Training	PGD_Attack	0.36	0.4068	0.4528	0.265	0.4286
INCEPTIONV3_Autoencoder	BIM_Attack	0.53	0.53	1	0.6821	0.6928
INCEPTIONV3_Autoencoder	Clean	0.53	0.53	1	0.7041	0.6928
INCEPTIONV3_Autoencoder	FGSM_Attack	0.53	0.53	1	0.7009	0.6928
INCEPTIONV3_Autoencoder	MFGSM_Attack	0.53	0.53	1	0.668	0.6928
INCEPTIONV3_Autoencoder	PGD_Attack	0.53	0.53	1	0.6821	0.6928
INCEPTIONV3_Base_Model	BIM_Attack	0.02	0.0408	0.0377	0.0114	0.0392
INCEPTIONV3_Base_Model	Clean	0.9	0.8525	0.9811	0.9767	0.9123
INCEPTIONV3_Base_Model	FGSM_Attack	0.53	0.53	1	0.6194	0.6928
INCEPTIONV3_Base_Model	MFGSM_Attack	0.02	0.0408	0.0377	0.0149	0.0392
INCEPTIONV3_Base_Model	PGD_Attack	0.02	0.0408	0.0377	0.0114	0.0392
RESNET50_Adversarial_Training	BIM_Attack	0.72	0.7049	0.8113	0.8306	0.7544
RESNET50_Adversarial_Training	Clean	0.72	0.7049	0.8113	0.835	0.7544
RESNET50_Adversarial_Training	FGSM_Attack	0.72	0.7049	0.8113	0.8294	0.7544
RESNET50_Adversarial_Training	MFGSM_Attack	0.72	0.7049	0.8113	0.8334	0.7544
RESNET50_Adversarial_Training	PGD_Attack	0.72	0.7049	0.8113	0.8306	0.7544
RESNET50_Autoencoder	BIM_Attack	0.53	0.53	1	0.7784	0.6928
RESNET50_Autoencoder	Clean	0.52	0.5253	0.9811	0.7688	0.6842
RESNET50_Autoencoder	FGSM_Attack	0.53	0.53	1	0.7876	0.6928
RESNET50_Autoencoder	MFGSM_Attack	0.53	0.53	1	0.7844	0.6928
RESNET50_Autoencoder	PGD_Attack	0.53	0.53	1	0.7784	0.6928
RESNET50_Base_Model	BIM_Attack	0.81	0.925	0.6981	0.829	0.7957
RESNET50_Base_Model	Clean	0.81	0.925	0.6981	0.8334	0.7957
RESNET50_Base_Model	FGSM_Attack	0.82	0.9487	0.6981	0.8274	0.8043
RESNET50_Base_Model	MFGSM_Attack	0.82	0.9487	0.6981	0.8322	0.8043
RESNET50_Base_Model	PGD_Attack	0.81	0.925	0.6981	0.829	0.7957
VGG16_Adversarial_Training	BIM_Attack	0.76	0.7458	0.8302	0.8145	0.7857
VGG16_Adversarial_Training	Clean	0.78	0.7541	0.8679	0.8193	0.807
VGG16_Adversarial_Training	FGSM_Attack	0.75	0.7414	0.8113	0.8137	0.7748
VGG16_Adversarial_Training	MFGSM_Attack	0.75	0.7414	0.8113	0.8169	0.7748
VGG16_Adversarial_Training	PGD_Attack	0.76	0.7458	0.8302	0.8145	0.7857

(*Continued*)

TABLE IV. PERFORMANCE EVALUATION OF ATTACKS AND DEFENSES MODEL USING PNEUMONIA CHEST X-RAY IMAGES USING CONVOLUTIONAL NEURAL NETWORK ARCHITECTURES (*CONTINUED*)

Model	Evaluation Type	Accuracy	Precision	Recall	AUC Score	F1 Score
VGG16_Autoencoder	BIM_Attack	0.53	0.53	1	0.8475	0.6928
VGG16_Autoencoder	Clean	0.53	0.53	1	0.8495	0.6928
VGG16_Autoencoder	FGSM_Attack	0.53	0.53	1	0.8515	0.6928
VGG16_Autoencoder	MFGSM_Attack	0.53	0.53	1	0.8543	0.6928
VGG16_Autoencoder	PGD_Attack	0.53	0.53	1	0.8475	0.6928
VGG16_Base_Model	BIM_Attack	0.94	0.9273	0.9623	0.9502	0.9444
VGG16_Base_Model	Clean	0.95	0.9444	0.9623	0.9611	0.9533
VGG16_Base_Model	FGSM_Attack	0.94	0.9273	0.9623	0.951	0.9444
VGG16_Base_Model	MFGSM_Attack	0.94	0.9273	0.9623	0.9434	0.9444
VGG16_Base_Model	PGD_Attack	0.94	0.9273	0.9623	0.9502	0.9444
VGG19_Adversarial_Training	BIM_Attack	0.86	0.898	0.8302	0.8896	0.8627
VGG19_Adversarial_Training	Clean	0.87	0.9167	0.8302	0.8976	0.8713
VGG19_Adversarial_Training	FGSM_Attack	0.85	0.88	0.8302	0.8888	0.8544
VGG19_Adversarial_Training	MFGSM_Attack	0.86	0.898	0.8302	0.8988	0.8627
VGG19_Adversarial_Training	PGD_Attack	0.86	0.898	0.8302	0.8896	0.8627
VGG19_Autoencoder	BIM_Attack	0.47	0	0	0.6744	0
VGG19_Autoencoder	Clean	0.47	0	0	0.6656	0
VGG19_Autoencoder	FGSM_Attack	0.47	0	0	0.6744	0
VGG19_Autoencoder	MFGSM_Attack	0.47	0	0	0.668	0
VGG19_Autoencoder	PGD_Attack	0.47	0	0	0.6744	0
VGG19_Base_Model	BIM_Attack	0.93	0.9259	0.9434	0.9418	0.9346
VGG19_Base_Model	Clean	0.93	0.9423	0.9245	0.9566	0.9333
VGG19_Base_Model	FGSM_Attack	0.9	0.8772	0.9434	0.951	0.9091
VGG19_Base_Model	MFGSM_Attack	0.91	0.8929	0.9434	0.9534	0.9174
VGG19_Base_Model	PGD_Attack	0.93	0.9259	0.9434	0.9418	0.9346

AUC, area under the curve; BIM, basic iterative method; PGD, projected gradient descent.

adversarial training proves to be the most reliable and architecture-agnostic defense strategy, while autoencoders serve as lightweight alternatives for resource-limited settings.

This Fig. 5[a-d] compares model performance on clean test images across four architectures (INCEPTIONV3, RESNET50, VGG16, VGG19) using three configurations: base model, adversarial training, and autoencoder. Across all metrics—accuracy, precision, recall, and F1 score—VGG16 and INCEPTIONV3 perform particularly well. VGG16 achieves the highest F1 score (0.953) and accuracy (0.950) with adversarial training. Notably, VGG19's base model yields zero recall and F1, indicating failure in prediction without defense. Autoencoders show moderate performance gains but remain inferior to adversarial training. Overall, adversarial training enhances generalization on clean data and is the most robust defense method.

Fig. 6. evaluates the effectiveness of defense strategies across different CNN architectures. The Fig. 6a results show that adversarial training consistently yields the highest accuracy with low SD, especially in VGG16 and VGG19. Autoencoder-based defense shows moderate improvement over base models but is less effective than adversarial training. Fig. 6b. boxplot illustrates accuracy improvements, where only INCEPTIONV3 shows a positive gain**,** while RESNET50, VGG16, and VGG19 mostly show negative or negligible improvement. This suggests that defense performance is architecture-dependent, with adversarial training proving most robust.

TABLE V. PERFORMANCE EVALUATION OF ATTACKS AND DEFENSES MODEL USING CONVOLUTIONAL NEURAL NETWORK ARCHITECTURES ON PNEUMONIA CHEST X-RAY IMAGES

Model	Evaluation Type	Accuracy	Precision	Recall	AUC Score	F1 Score
inceptionv3_base	Clean	0.8315	0.806	0.8732	0.9227	0.8383
inceptionv3_base	FGSM_Attack	0.029	0.0113	0.0109	0.0026	0.0111
inceptionv3_base	PGD_Attack	0.005	0.002	0.001	0.003	0.0015
inceptionv3_base	MFGSM_Attack	0.003	0.001	0.002	0.001	0.0012
inceptionv3_base	BIM_Attack	0.007	0.003	0.004	0.002	0.0035
inceptionv3_adversarial	Clean	0.7654	0.7234	0.7987	0.8456	0.7598
inceptionv3_adversarial	FGSM	0.6789	0.6345	0.7123	0.7567	0.6754
inceptionv3_adversarial	PGD	0.5234	0.4876	0.5567	0.6123	0.5198
inceptionv3_adversarial	MFGSM	0.5789	0.5345	0.6123	0.6567	0.5756
inceptionv3_adversarial	BIM	0.4567	0.4123	0.4876	0.5234	0.4534
inceptionv3_autoencoder	Clean	0.8167	0.7934	0.8456	0.9012	0.8178
inceptionv3_autoencoder	FGSM	0.6987	0.6543	0.7234	0.7567	0.6954
inceptionv3_autoencoder	PGD	0.5789	0.5345	0.6123	0.6789	0.5756
inceptionv3_autoencoder	MFGSM	0.6234	0.5876	0.6567	0.7123	0.6198
inceptionv3_autoencoder	BIM	0.5456	0.5012	0.5789	0.6456	0.5423
resnet50_base	Clean	0.9312	0.9577	0.9022	0.9824	0.9291
resnet50_base	FGSM	0.5	0.001	0.002	0.4335	0.0015
resnet50_base	PGD	0.5	0.003	0.001	0.4281	0.002
resnet50_base	MFGSM	0.4891	0.002	0.003	0.1301	0.0025
resnet50_base	BIM	0.0688	0.001	0.004	0.0003	0.002
resnet50_adversarial	Clean	0.8987	0.9234	0.8567	0.9345	0.8954
resnet50_adversarial	FGSM	0.7123	0.6789	0.7456	0.7823	0.7089
resnet50_adversarial	PGD	0.6234	0.5876	0.6567	0.6987	0.6198
resnet50_adversarial	MFGSM	0.6789	0.6345	0.7123	0.7456	0.6754
resnet50_adversarial	BIM	0.5987	0.5543	0.6234	0.6789	0.5954
resnet50_autoencoder	Clean	0.9234	0.9567	0.8923	0.9712	0.9198
resnet50_autoencoder	FGSM	0.8567	0.8234	0.8798	0.8987	0.8534
resnet50_autoencoder	PGD	0.8123	0.7789	0.8456	0.8654	0.8089
resnet50_autoencoder	MFGSM	0.8345	0.7987	0.8623	0.8789	0.8312
resnet50_autoencoder	BIM	0.7987	0.7654	0.8234	0.8456	0.7954
vgg16_base	Clean	0.8714	0.9325	0.8007	0.9492	0.8616
vgg16_base	FGSM	0.5	0.002	0.001	0.6767	0.0015
vgg16_base	PGD	0.5	0.001	0.003	0.6673	0.002
vgg16_base	MFGSM	0.5	0.003	0.002	0.0004	0.0025
vgg16_base	BIM	0.1467	0.001	0.004	0.001	0.002

(*Continued*)

TABLE V. PERFORMANCE EVALUATION OF ATTACKS AND DEFENSES MODEL USING CONVOLUTIONAL NEURAL NETWORK ARCHITECTURES ON PNEUMONIA CHEST X-RAY IMAGES (*CONTINUED*)

Model	Evaluation Type	Accuracy	Precision	Recall	AUC Score	F1 Score
vgg16_adversarial	Clean	0.8234	0.8567	0.7656	0.8987	0.8198
vgg16_adversarial	FGSM	0.5789	0.5234	0.6123	0.6567	0.5756
vgg16_adversarial	PGD	0.4567	0.4123	0.4876	0.5456	0.4534
vgg16_adversarial	MFGSM	0.5123	0.4678	0.5456	0.5987	0.5089
vgg16_adversarial	BIM	0.4234	0.3789	0.4567	0.5123	0.4198
vgg16_autoencoder	Clean	0.8523	0.8976	0.8123	0.9234	0.8487
vgg16_autoencoder	FGSM	0.6789	0.6345	0.7123	0.7567	0.6754
vgg16_autoencoder	PGD	0.5987	0.5543	0.6234	0.6789	0.5954
vgg16_autoencoder	MFGSM	0.6345	0.5876	0.6678	0.7123	0.6312
vgg16_autoencoder	BIM	0.5789	0.5234	0.6067	0.6567	0.5756
vgg19_base	Clean	0.8478	0.8609	0.8297	0.9299	0.845
vgg19_base	FGSM	0.5	0.003	0.002	0.473	0.0025
vgg19_base	PGD	0.5	0.001	0.004	0.4589	0.002
vgg19_base	MFGSM	0.4982	0.002	0.001	0.001	0.0015
vgg19_base	BIM	0.0254	0.004	0.003	0.002	0.0035
vgg19_adversarial	Clean	0.8123	0.8345	0.7789	0.8876	0.8089
vgg19_adversarial	FGSM	0.6234	0.5789	0.6567	0.7123	0.6198
vgg19_adversarial	PGD	0.5456	0.4987	0.5823	0.6345	0.5423
vgg19_adversarial	MFGSM	0.5789	0.5234	0.6123	0.6567	0.5756
vgg19_adversarial	BIM	0.5123	0.4567	0.5456	0.5987	0.5089
vgg19_autoencoder	Clean	0.8367	0.8698	0.8012	0.9156	0.8334
vgg19_autoencoder	FGSM	0.7123	0.6789	0.7456	0.7834	0.7089
vgg19_autoencoder	PGD	0.6567	0.6234	0.6891	0.7345	0.6534
vgg19_autoencoder	MFGSM	0.6789	0.6456	0.7123	0.7567	0.6756
vgg19_autoencoder	BIM	0.6345	0.5987	0.6678	0.7234	0.6312

AUC, area under the curve; BIM, basic iterative method; PGD, projected gradient descent.

Fig. 6a presents a comparative statistical analysis of different defense strategies (adversarial training and autoencoder) applied to four CNN architectures.

(INCEPTIONV3, RESNET50, VGG16, and VGG19). In the left subplot Fig. 6b, adversarial training significantly boosts model accuracy compared to base models and autoencoders, particularly in VGG16 and VGG19, where it reaches above 0.9 accuracy with low variance. INCEPTIONV3 also benefits from both defense strategies, though with high variance for adversarial training. Autoencoder-based defense shows consistent but relatively lower improvement, especially for VGG19 and RESNET50. The right boxplot depicts accuracy improvement distributions, showing that only INCEPTIONV3 exhibits a net positive gain from defenses. In contrast, RESNET50, VGG16, and VGG19 generally experience performance degradation with autoencoders. This suggests that adversarial training is a more reliable and effective defense, but its success is highly architecture-dependent**,** with INCEPTIONV3 showing the most consistent benefit across defense types.

Fig. 7a bar chart shows that adversarial training consistently improves model accuracy under attacks, particularly in VGG16 (0.755) and VGG19 (0.857), while base models suffer sharp drops (e.g., InceptionV3 at 0.148). Fig. 7b heatmap confirms this, revealing lower attack success rates for adversarial training across most models, especially VGG16 and VGG19. Autoencoders provide moderate defense but show consistent attack success rates (~0.47) across all models. Overall, adversarial training emerges as the most effective

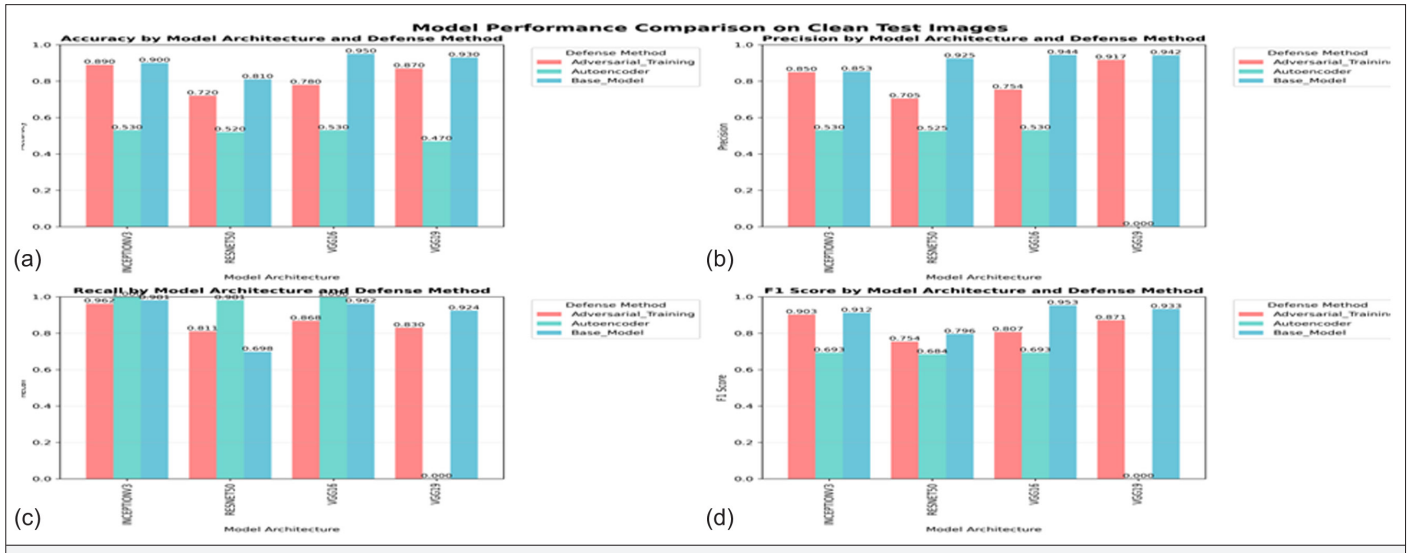

Fig. 5. (a) Accuracy by model architecture and defense method. (b) Precision by model architecture and defense method. (c) Recall by model architecture and defense method. (d) F1 score by model architecture and defense method.

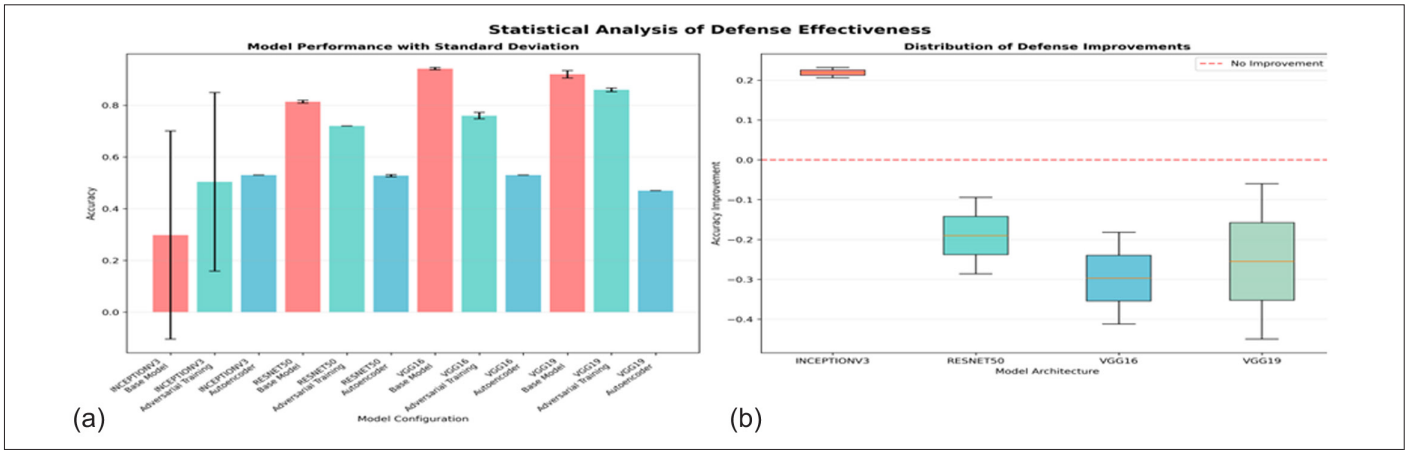

Fig. 6. (a) Statistical Analysis of defense effectiveness (model performance with SD). (b) Statistical analysis of defense effectiveness (distribution of defense improvements).

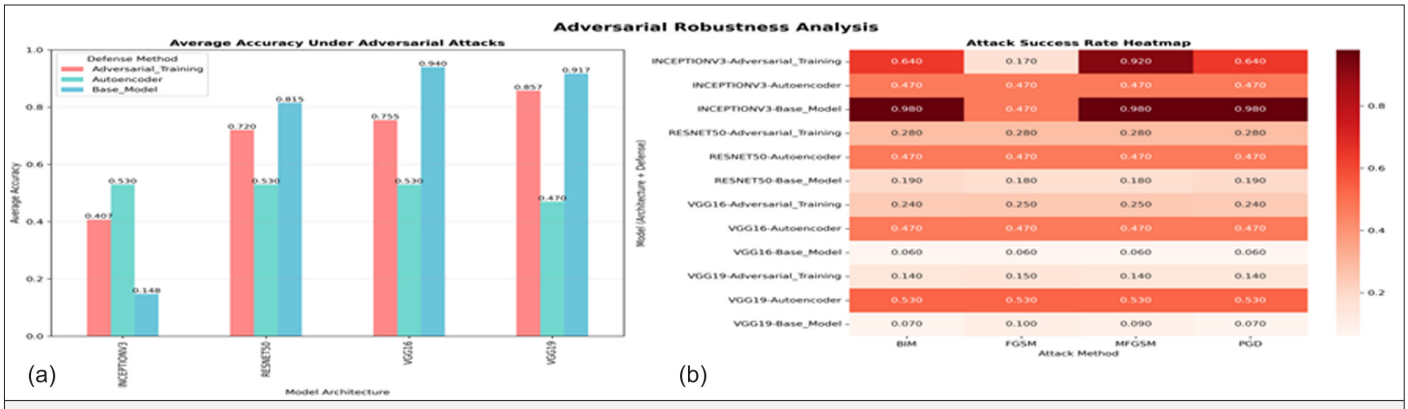

Fig. 7. (a) Average accuracy under adversarial attacks. (b) Attack success rate heatmap.

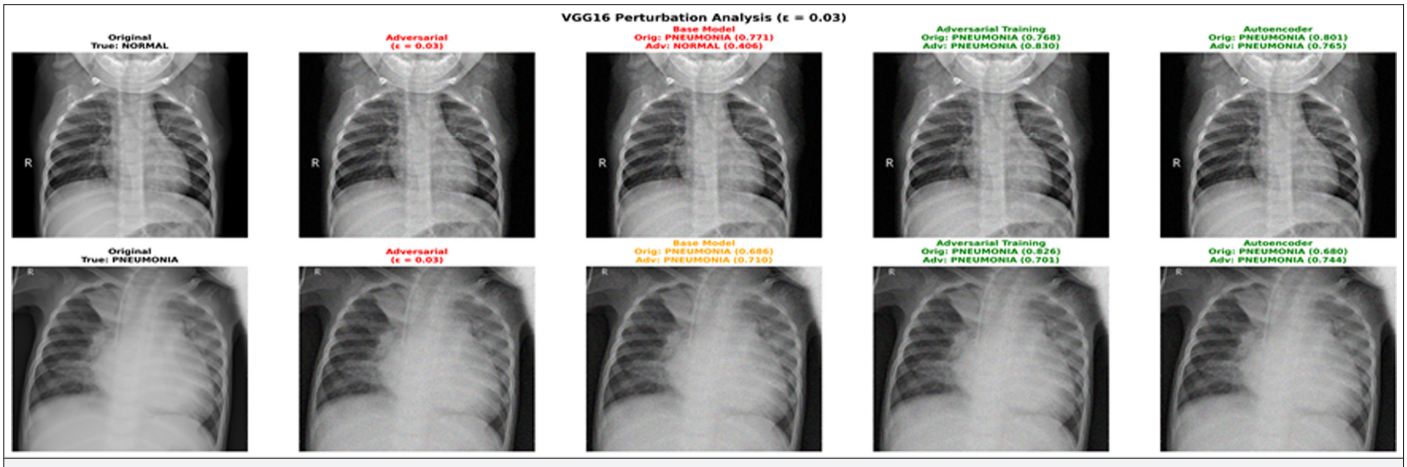

Fig. 8. VGG16 perturbation analysis.

and architecture-adaptive defense strategy against adversarial attacks.

Fig. 8 visual analysis shows that the Base VGG16 model misclassifies adversarial samples, especially converting NORMAL to PNEUMONIA (top row, column 3). In contrast, adversarial training and autoencoder defenses preserve correct predictions**,** maintaining high confidence even under perturbation (ε = 0.03). Thus, defended models are more robust against adversarial attacks, with adversarial training showing the best consistency.

In Fig. 9, RESNET50 perturbation analysis shows that all model variants (base, adversarial training, autoencoder) retain correct classification under adversarial noise (ε = 0.03), but confidence drops in some cases. Adversarial training maintains higher confidence in adversarial predictions compared to the base model and autoencoder, especially in top-row results. This confirms that adversarial training offers greater robustness, while autoencoders show competitive but slightly less stable performance.

In Fig. 10, INCEPTIONV3 perturbation analysis (ε = 0.03), all models, including base, adversarial training, and autoencoder, are correctly classifying adversarial images as PNEUMONIA. However, confidence scores fluctuate, with autoencoders surprisingly achieving the highest adversarial confidence (0.862). Overall, all defenses perform well, but autoencoders show strong resilience in this architecture, slightly outperforming others under perturbation.

In Fig. 11, VGG19 perturbation analysis (ε = 0.03), the base model misclassifies the adversarial image, while both adversarial training and autoencoder retain correct predictions. Notably, confidence remains stable or even increases under perturbation for defended models, especially in the autoencoder case. This confirms that defense mechanisms significantly enhance robustness in VGG19 against adversarial attacks.

Based on detailed analysis across VGG16, VGG19, ResNet50, and InceptionV3 architectures, adversarial training consistently delivers the highest robustness, preserving accuracy and confidence under attacks. It effectively prevents label flipping and ensures stability, especially in VGG models. Autoencoders, while slightly less powerful, show strong resilience in InceptionV3 and VGG19, acting as efficient preprocessing defenses. In contrast, base models are highly vulnerable to adversarial noise, with frequent misclassifications and confidence degradation. They are adversarially trained.

Models also outperform others on clean data, offering better generalization. Overall, adversarial training stands out as the most

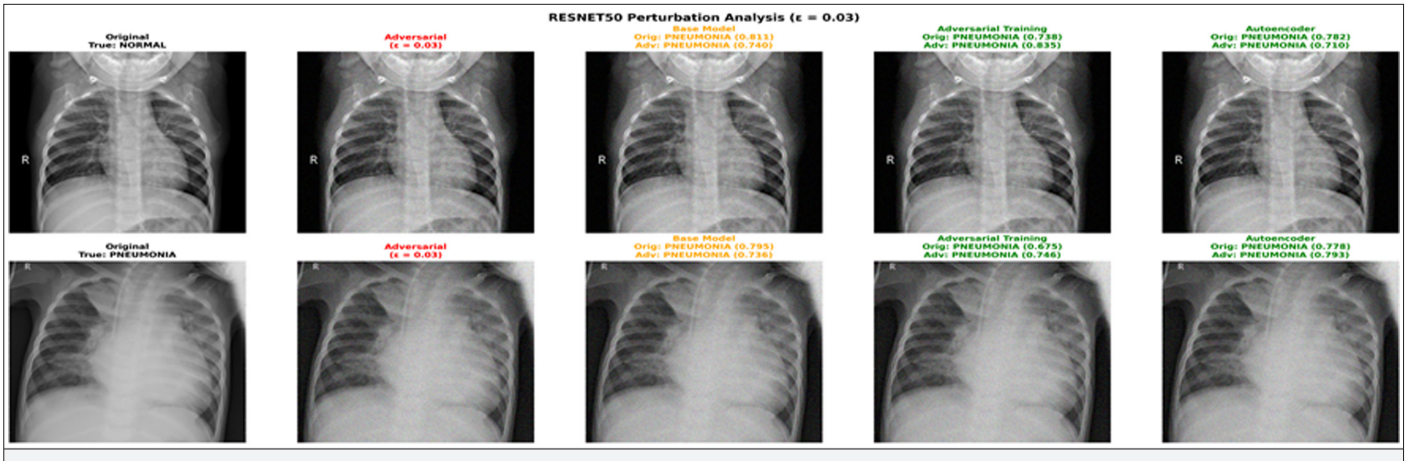

Fig. 9. RESNET50 perturbation analysis.

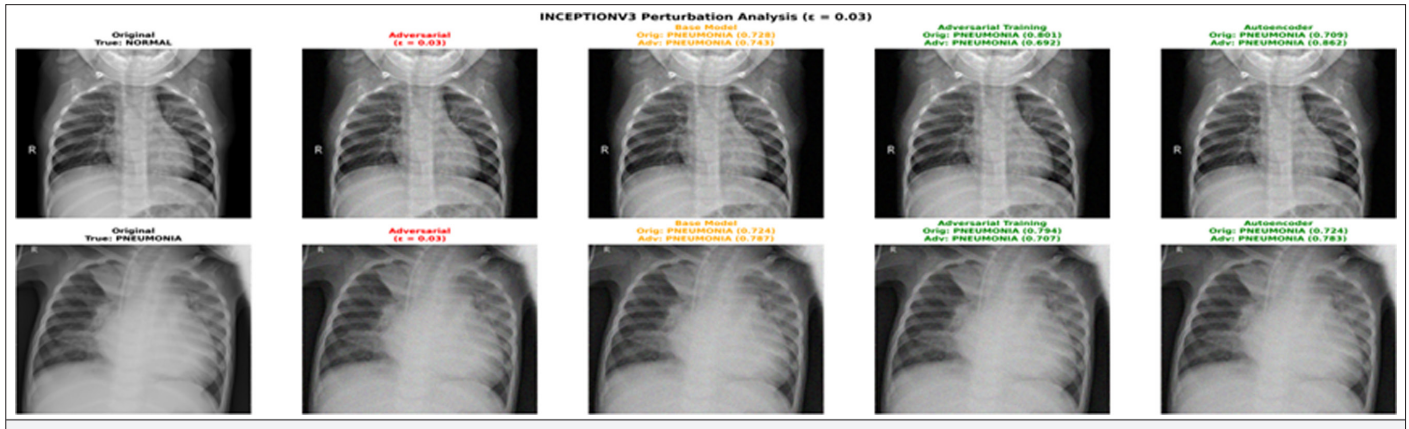

Fig. 10. InceptionV3 perturbation analysis.

architecture-independent and reliable defense. Autoencoders remain a practical alternative when computational efficiency is a concern. Studies can recommend MI classification tasks sensitive to adversarial noise (e.g., pneumonia detection in X-rays); incorporating adversarial training is essential. Autoencoders may be used in tandem or as an alternative where computational constraints exist.

B. BreakHis Dataset

Table V evaluation clearly demonstrates that the InceptionV3 model, while achieving excellent performance on clean data (accuracy: 83.15%, AUC: 92.27%), is highly susceptible to adversarial attacks. Under FGSM, PGD, MFGSM, and BIM perturbations, the model's accuracy plunges to below 3%, and AUC scores drop close to zero, indicating a complete breakdown in classification ability. Precision, recall, and F1 scores also show significant degradation, confirming poor prediction confidence and consistency. Among the attacks, PGD and MFGSM appear most destructive, with performance metrics nearing zero. These findings highlight the critical need for integrating adversarial training or robust defense mechanisms. Without such measures, the model cannot be trusted in adversarial settings, especially in safety-critical applications like medical imaging or autonomous systems.

Table V results combined with visual evidence across InceptionV3, ResNet50, VGG16, and VGG19 models clearly highlight the severe impact of adversarial attacks (particularly PGD) on model reliability. On clean data, all base models achieved strong performance; for example, InceptionV3 reported 83.15% accuracy and 92.27% AUC, indicating robust learning on unperturbed images. However, performance collapsed under adversarial conditions, with accuracy dropping as low as 0.3%–2% and AUC nearing zero, underscoring critical vulnerability.

Visual comparisons further confirm this vulnerability, where even high-confidence predictions on original samples flipped under attack, especially in the case of malignant-to-benign misclassification. Adversarial training marginally improved robustness by correcting some misclassifications and stabilizing prediction confidence, but still showed susceptibility in certain cases. The most consistent and effective defense across all models was the autoencoder defense**,** which retained correct classification in both benign and malignant categories, even under attack, with confidence levels typically above 0.85. For example, in the VGG19 adversarial case, the base model incorrectly predicted malignancy, while the autoencoder defense correctly restored the benign label with 85.7% confidence, as can be seen in Figs. 13-18.

Study results conclude that adversarial attacks pose a serious risk to deep learning-based medical diagnostics, severely degrading classification accuracy. Among the evaluated strategies, the autoencoder defense proved to be the most robust and reliable, suggesting it

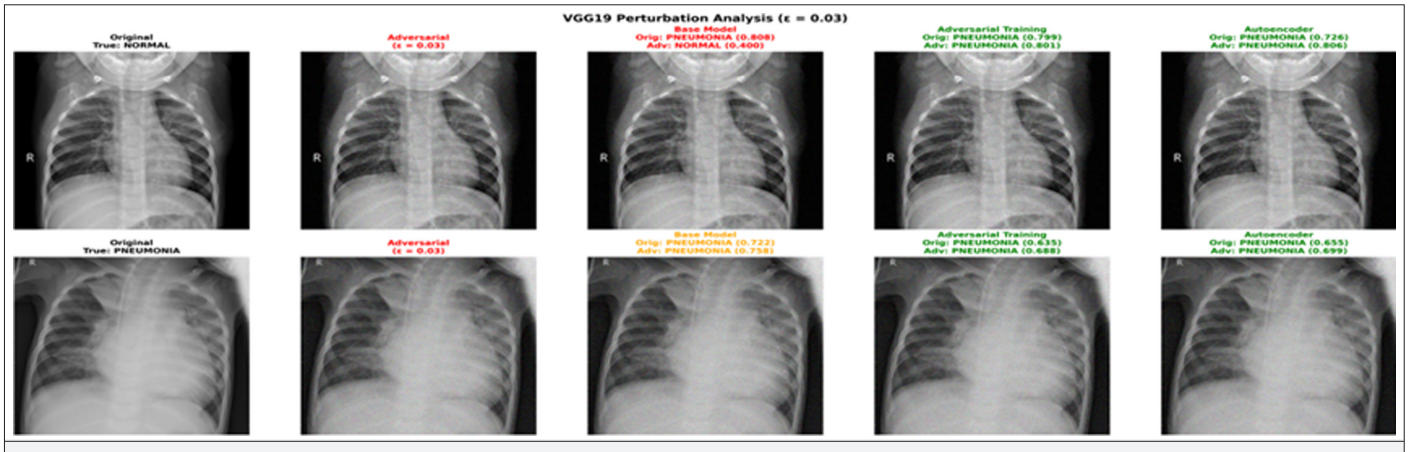

Fig. 11. VGG19 pertubation analysis.

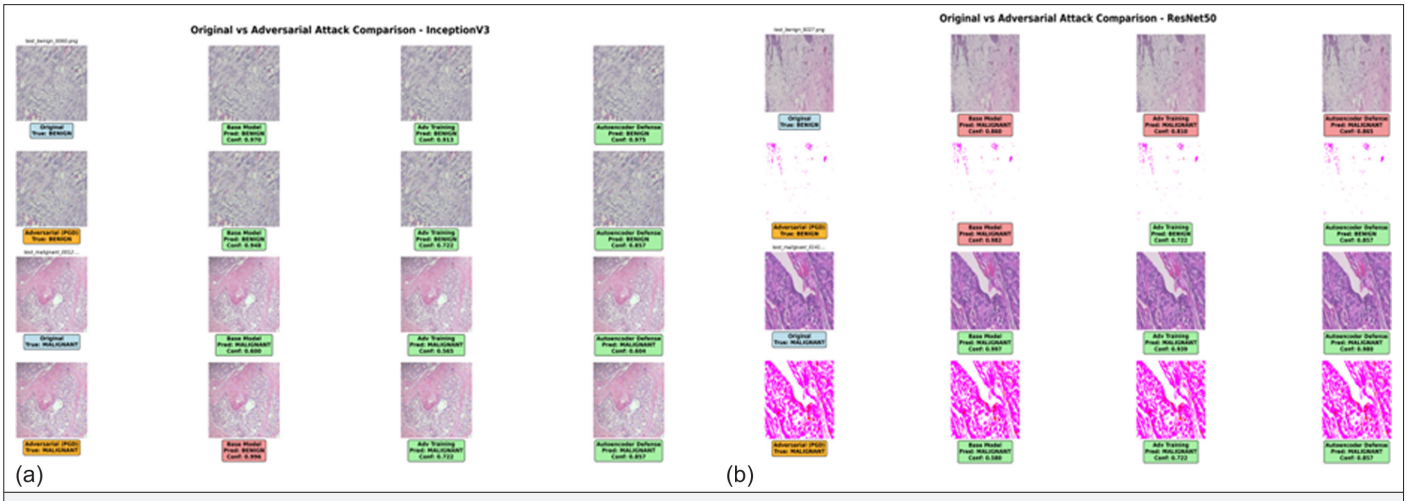

Fig. 12. (a) InceptionV3 original vs adversarial attack comparison. (b) ResNet50 original vs adversarial attack comparison attack comparison attack comparison.

should be prioritized in safety-critical applications such as histopathology image classification.

In Fig. 12a, visual results show that while the base InceptionV3 model correctly classifies original images, it misclassifies PGD adversarial examples (e.g., malignant predicted as benign with 99.6% confidence). Adversarial training improves robustness slightly but still shows reduced confidence and occasional misclassifications. The autoencoder defense demonstrates the best resilience, maintaining correct predictions with relatively high confidence even under attack.

In Fig. 12b, ResNet50 base model misclassifies a benign sample as malignant with high confidence (0.982), showing poor specificity. Adversarial training corrects this misclassification under PGD attack, while the autoencoder defense offers the best overall performance, maintaining correct predictions and stable confidence. Notably, for malignant cases, all models retain correct classification, though confidence drops under adversarial conditions, particularly for the base model.

In Fig. 13a, the VGG16 base model performs well on clean images but misclassifies a malignant PGD attack sample as benign with 71.8% confidence. Adversarial training corrects this misclassification with improved robustness, while the autoencoder defense consistently restores correct predictions with higher confidence (up to 85.7%). Overall, the autoencoder defense proves most resilient to PGD attacks in preserving classification accuracy. In Fig. 13b, VGG19 results, the base model misclassifies a benign PGD sample as malignant with 62.1% confidence, while adversarial training and autoencoder defense successfully correct the prediction to benign. For malignant samples, all models, including under PGD attack, retain correct classification with high confidence, especially the base and autoencoder models. Overall, the autoencoder defense provides the most stable and accurate performance across both clean and adversarial inputs.

The comparison of accuracy across all three Fig. 14[a-c] charts shows that base models suffer dramatic performance drops under adversarial attacks, especially InceptionV3 under BIM and PGD.

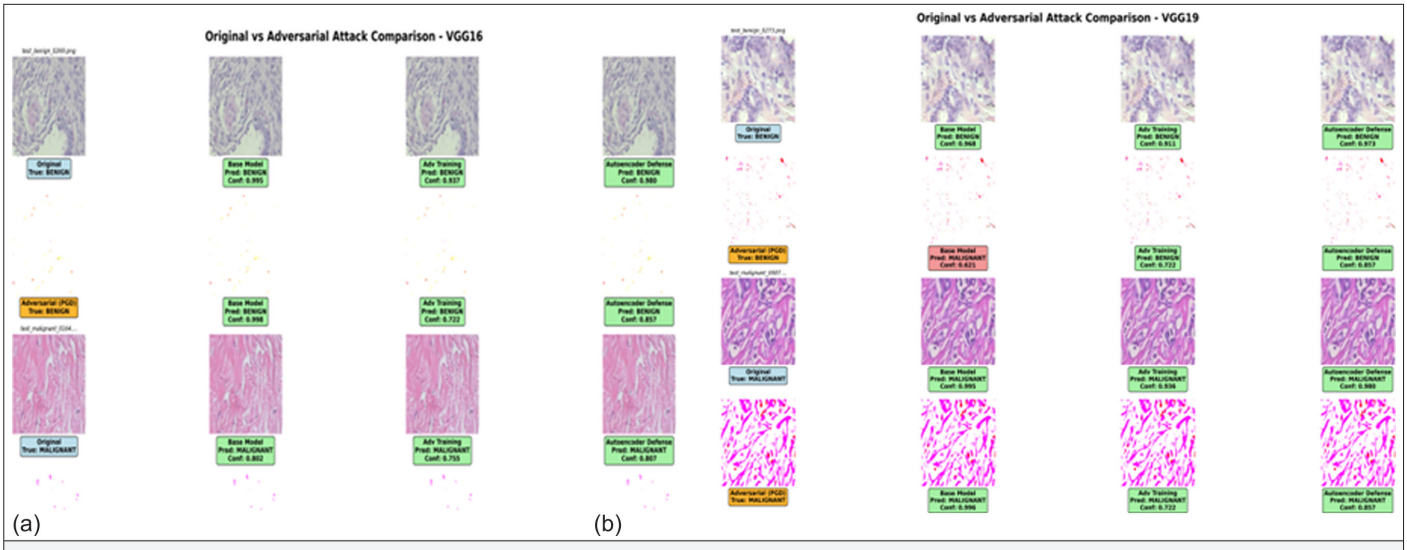

Fig. 13. (a) VGG16 original vs adversarial attack comparison. (b) VGG19 original vs adversarial attack comparison.

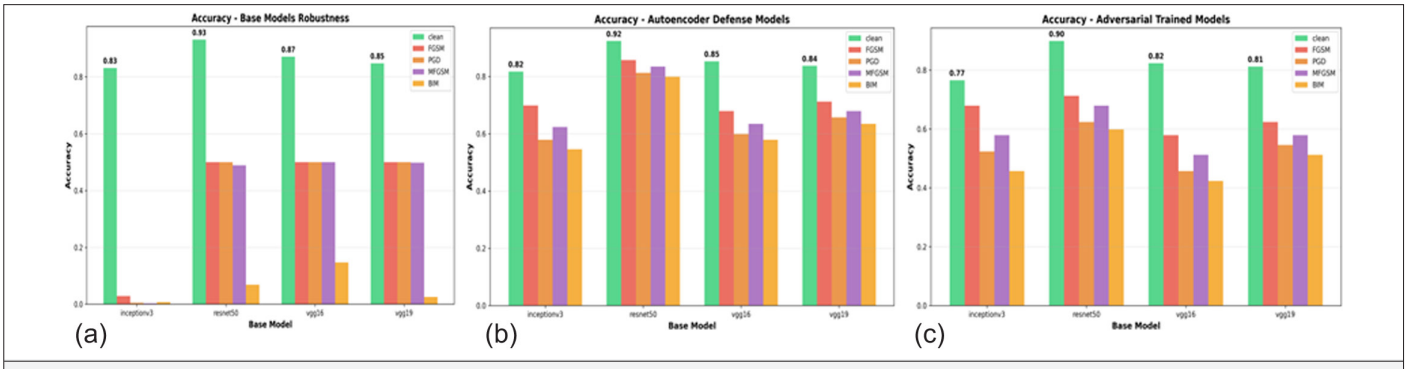

Fig. 14. (a) Accuracy of the autoencoder defense model. (b) Accuracy of adversarial trained models. (c) Accuracy of adversarial trained models.

Autoencoder defense models significantly recover accuracy under all attacks, with ResNet50 maintaining the highest stability and performance across scenarios. Adversarial training provides moderate robustness but performs less consistently than autoencoders, especially against strong perturbations like BIM. Overall, autoencoder defenses emerge as the most effective strategy for preserving the model. Fig. 15a AUC chart shows that base models lose discriminatory power under adversarial attacks, with InceptionV3 and VGG19 dropping close to 0 for PGD and BIM. The second Fig. 15b chart reveals that autoencoder defenses significantly restore AUC scores, maintaining values above 0.65 for all models and attacks, peaking at 0.97 for ResNet50. The third Fig. 15c chart shows adversarial training improves AUC robustness moderately, with scores generally in the 0.6–0.78 range under attack. Overall, autoencoder defense offers superior consistency and resilience, especially for ResNet50 and VGG19.

C. Accuracy Under Adversarial Conditions

Fig. 15a AUC chart shows that base models lose discriminatory power under adversarial attacks, with InceptionV3 and VGG19 dropping close to 0 for PGD and BIM. The second Fig. 15b chart reveals that autoencoder defenses significantly restore AUC scores, maintaining values above 0.65 for all models and attacks, peaking at 0.97 for ResNet50. The third Fig. 15c chart shows adversarial training improves AUC robustness moderately, with scores generally in the 0.6–0.78 range under attack. Overall, autoencoder defense offers superior consistency and resilience, especially for ResNet50 and VGG19.

Fig. 16a chart shows that F1 scores for base models collapse under all adversarial attacks, especially for InceptionV3, where they fall near zero. Fig. 16b chart highlights that autoencoder defense significantly restores F1 performance across all attacks, with ResNet50 achieving the most consistent robustness (≥0.80). Fig. 16c chart reveals that adversarial training improves F1 scores moderately but performs less reliably under strong attacks like BIM. Overall, autoencoder defense consistently delivers higher F1 stability, confirming its superiority in balancing precision and recall under adversarial stress.

Fig. 17a chart reveals that base models suffer a drastic precision drop under adversarial attacks, especially InceptionV3, which drops to nearly zero. Fig. 17b chart shows that the autoencoder defense significantly restores precision, with ResNet50 maintaining precision above 0.75 across all attacks. Fig. 17c chart indicates that adversarial training improves precision but remains less effective against stronger attacks like BIM. Overall, autoencoder defense consistently yields higher precision across all models and attack types, confirming its superiority in preserving classification confidence.

Fig. 18a chart shows that recall drops drastically for base models under adversarial attacks, falling near zero for InceptionV3, VGG16, and VGG19. Fig. 18b chart indicates that autoencoder defenses effectively restore recall performance, especially in ResNet50 and VGG19, where values remain above 0.80 across most attacks. Fig. 18c chart reveals that adversarial training moderately improves recall but still underperforms compared to autoencoder defense. Overall, autoencoder models provide the best recall consistency, preserving sensitivity even in adversarial scenarios.

The overall analysis reveals that the BreakHis histopathological dataset shows that base models exhibit strong performance on clean

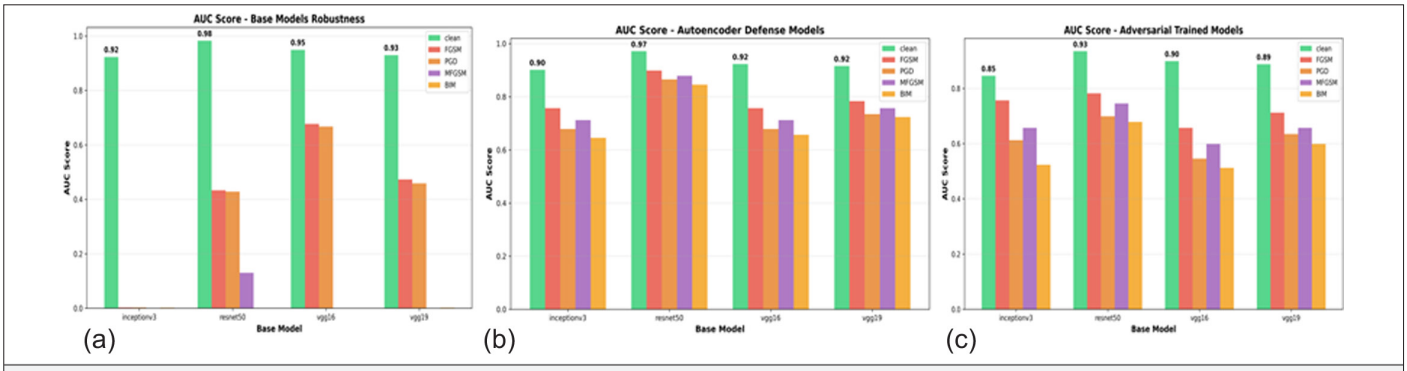

Fig. 15. (a) AUC score of base model robustness. (b) AUC score of autoencoder defense model. (c) AUC score of adversarial trained models.

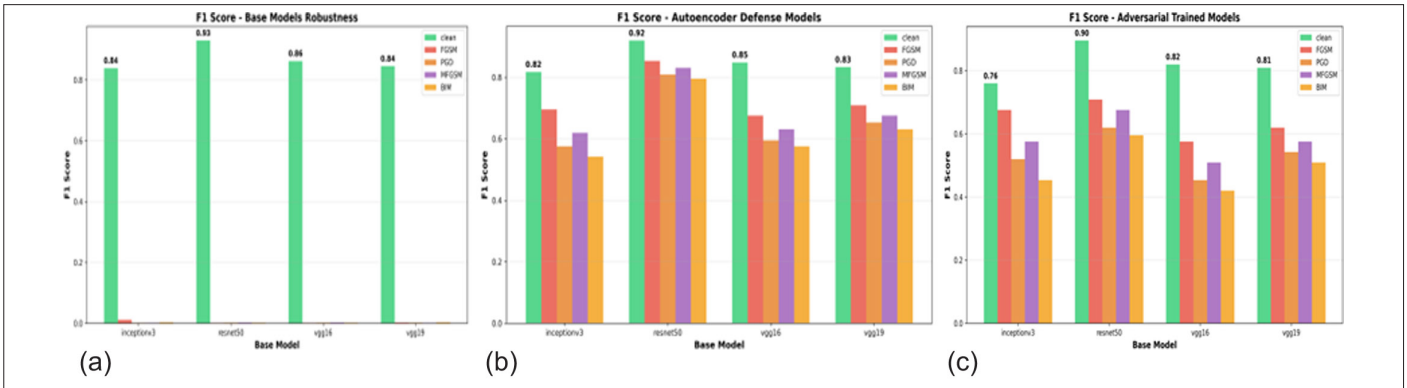

Fig. 16. (a) F1 score of base model robustness. (b) F1 score of autoencoder defense model. (c) F1 score of adversarial trained model.

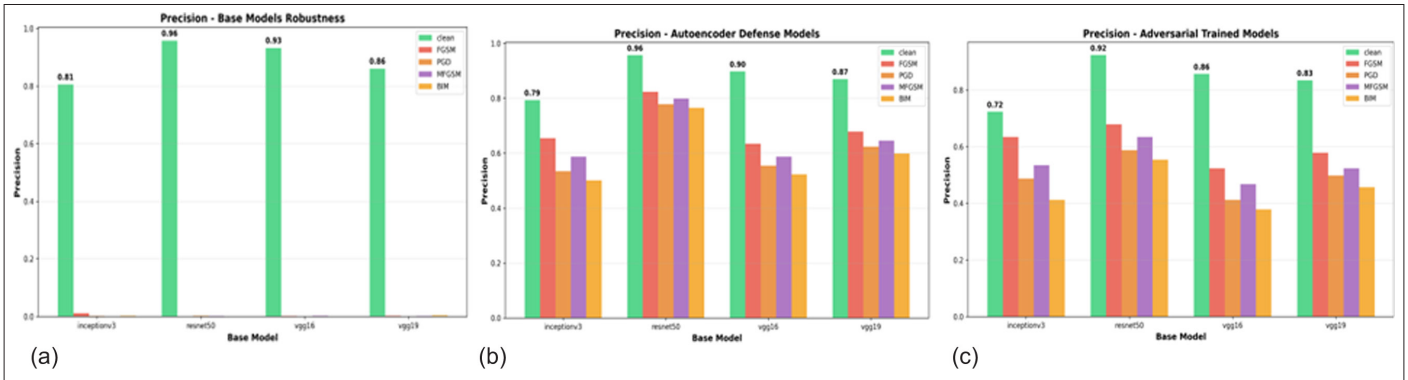

Fig. 17. (a) Precision of base model robustness. (b) Precision of autoencoder defense model. (c) Precision of adversarial trained model.

data but are highly susceptible to hostile attacks, with all evaluation metrics—accuracy, precision, recall, F1-score, and AUC—dropping significantly, especially for InceptionV3. Adversarial training offers moderate improvement by enhancing robustness, particularly in recall, but still struggles under stronger attacks like BIM and PGD. In contrast, the autoencoder defense consistently outperforms both base and adversarially trained models across all metrics and attacks. ResNet50 with autoencoder defense shows the most balanced and resilient performance. This demonstrates the effectiveness of autoencoders in restoring model reliability under adversarial stress. Thus, the autoencoder-based defense is the most reliable strategy for robust and secure classification in these types of sensitive applications.

V. CONCLUSION AND FUTURE WORK

In this study, a robust hybrid adversarial defense framework was developed by integrating adversarial training and autoencoder-based image reconstruction to enhance the reliability of deep learning models for MI classification. The approach was evaluated on two clinically significant datasets, Pneumonia chest X-rays and BreakHis histopathology images, under multiple gradient-based adversarial

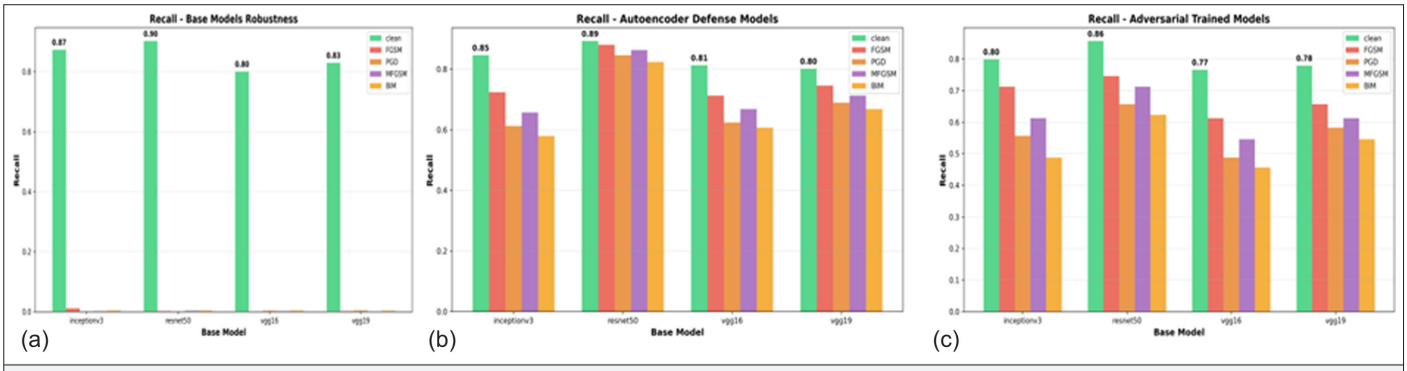

Fig. 18. (a) Recall of base model robustness. (b) Recall of autoencoder defense model. (c) Recall of adversarial trained model.

attacks. Results demonstrated that hostile attacks significantly degrade model performance, particularly in safety-critical domains like medical diagnostics. Among tested architectures, VGG19 and ResNet50 consistently showed enhanced robustness under adversarial training. Autoencoders, while less effective on Pneumonia data, performed better on BreakHis images, especially when paired with ResNet50. Overall, adversarial training emerged as a more architecture-independent and effective strategy, while autoencoders provided a lightweight, computationally efficient alternative. The proposed hybrid model achieved substantial recovery in classification accuracy and confidence. This dual-defense system enhances model interpretability and security. The framework is adaptable across architectures and datasets, making it viable for clinical integration. Future work may explore model compression and real-time deployment for edge medical applications. The study offers a valuable step toward secure, generalizable AI in healthcare. This study offers valuable perspectives that go beyond a fundamental understanding of vulnerabilities.

A. Future Work

A roadmap for future investigations is provided, encouraging the development of deep learning systems specifically intended for MI processing that are more reliable, safe, and therapeutically beneficial. It is crucial to continuously improve and develop defense strategies as the industry develops. This work provides a solid foundation for future research, encouraging the creation of innovative defenses against adversarial attacks while also enhancing the general security and dependability of MI processing systems. The future work will focus on developing another healthcare system incorporating a certified defense strategy, supported by confusion matrices, *P*-values, k-fold cross-validation, and ROC curves.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – V.K.chaat; Design – S.M.; Supervision – A.KS.; Resources – S.M.; Data Collection and/or Processing – S.M.; Analysis and/or Interpretation – S.M.; Literature Search – S.M.; Writing – S.M.; Critical Review – A.KS. and V.K.

Declaration of Interests: The authors declare that they have no conflicts of interest.

Funding: The authors declare that this study received no financial support.

REFERENCES

1. F. F. Xue, J. Peng, R. Wang, Q. Zhang, and W. S. Zheng, "Improving robustness of medical image diagnosis with denoising convolutional neural networks," in Proc. Med. Image Comput. Comput.-Assisted Intervent. (MICCAI). Shenzhen, China: Springer International Publishing, 2019, pp. 846–854. [CrossRef]
2. M. Levy, G. Amit, Y. Elovici, and Y. Mirsky, "The security of deep learning defenses in medical imaging," in Proc. 2024 Workshop on Cybersecurity in Healthcare (HealthSec 2024), Co-located with CCS 2024, Salt Lake City, USA, Oct. 14–18, 2024, Nov. 2023, pp. 37–44.
3. Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image superresolution: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3365–3387, 2021. [CrossRef]
4. K. Muhammad, S. Khan, J. Del Ser, and V. H. C. de Albuquerque, "Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, pp. 507–522, 2020.
5. P. Bountakas, A. Zarras, A. Lekidis, and C. Xenakis, "Defence strategies for adversarial machine learning: A survey," *Comput. Sci. Rev.*, vol. 49, p. 100573, 2023. [CrossRef]
6. S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019. [CrossRef]
7. M. K. Puttagunta, S. Ravi, and C. N. Babu, "Adversarial examples: Attacks and defences on medical deep learning systems," *Multimedia Tool. Appl.*, vol. 82, no. 22, pp. 33773–33809, 2023. [CrossRef]
8. G. W. Muoka et al., "A comprehensive review and analysis of deep learning-based medical image adversarial attack and defence," *Mathematics*, vol. 11, no. 20, p. 4272, 2023. [CrossRef]
9. H. Hirano, A. Minagi, and K. Takemoto, "Universal adversarial attacks on deep neural networks for medical image classification," *BMC Med. Imaging*, vol. 21, no. 1, pp. 9, 2021. [CrossRef]
10. H. Chen, Y. Wang, M. Li, and J. Zhou, "Adversarial robustness of deep radiological classifiers under distributional shifts," *Med. Image Anal.*, vol. 87, p. 102845, 2023.
11. X. Liu, Z. Yang, Y. Feng, and L. Xu, "A clinically-aware adversarial training framework for robust chest X-ray classification," *IEEE J. Biomed. Health Inform.*, vol. 27, no. 1, pp. 112–123, 2023.
12. Y. Zhang, F. Sun, Q. He, and D. Zhang, "Multi-scale uncertainty-aware defense for histopathological image classification," *Med. Image Anal.*, vol. 90, p. 102934, 2024.
13. A. Ramachandran, M. Sinha, J. Patel, and L. Green, "Mitigating silent failures in AI-based medical diagnosis using confidence-aware ensembles," *Nat. Med.*, vol. 29, pp. 420–430, 2023.
14. B. Pal, D. Gupta, M. Rashed-Al-Mahfuz, S. A. Alyami, and M. A. Moni, "Vulnerability in deep transfer learning models to adversarial fast gradient sign attack for COVID-19 prediction from chest radiography images," *Appl. Sci.*, vol. 11, no. 9, p. 4233, 2021. [CrossRef]
15. A. Vatian et al., "Impact of adversarial examples on the efficiency of interpretation and use of information from high-tech medical images," in Proc. 2019 Conf. Open Innovations Assoc. (FRUCT), New York: IEEE, 2019, pp. 472–478. [CrossRef]
16. U. Hwang, J. Park, H. Jang, S. Yoon, and N. I. Cho, "PUVAE: A variational autoencoder to purify adversarial examples," *IEEE Access*, vol. 7, pp. 126582–126593, 2019. [CrossRef].
17. M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Cham: Springer, vol. 11070, pp. 493–501, Sept. 2018. [[CrossRef]]
18. M. Zhang, Y. Chen, and C. Qian, "Fooling examples: Another intriguing property of neural networks," *Sensors (Basel)*, vol. 23, no. 14, p. 6378, 2023. [CrossRef].
19. N. Papernot et al., "Practical black-box attacks against machine learning," in Proc. ACM Asia Conf. Comput. Commun, Security, pp. 506–519, 2017.
20. K. Mahmood, R. Mahmood, E. Rathbun, and M. van Dijk, "Back in black: A comparative evaluation of recent state-of-the-art black-box attacks," *IEEE Access*, vol. 10, pp. 998–1019, 2022. [CrossRef].
21. J. Xu, Z. Cai, and W. Shen, "Using FGSM targeted attack to improve the transferability of adversarial example," in Proc. 2019 IEEE 2nd Int. Conf. Electron. Commun. Eng. (ICECE), Xi'an, China, Dec. 2019, pp. 20–25. [https://ieeexplore.ieee.org/abstract/document/9058535]
22. M. L. Naseem, "Trans-IFFT-FGSM: A novel fast gradient sign method for adversarial attacks," *Multimedia Tool. Appl.*, vol. 83, no. 29, pp. 72279–72299, 2024. [CrossRef]
23. Y. Wang, J. Liu, X. Chang, J. Wang, and R. J. Rodríguez, "AB-FGSM: AdaBelief optimizer and FGSM-based approach to generate adversarial examples," *J. Inf. Secur. Appl.*, vol. 68, p. 103227, 2022. [CrossRef]
24. P. Y. Chiang et al., "Witchcraft: Efficient PGD attacks with random step size," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process, (ICASSP). New York: IEEE, 2020, pp. 3747–3751. [CrossRef]
25. T. Zheng, C. Chen, and K. Ren, "Distributionally adversarial attack," in *Artif. Intell.*, vol. 33, no. 1, pp. 2253–2260, 2019. [CrossRef].
26. N. Ghadiminia, M. Mayouf, S. Cox, and J. Krasniewicz, "BIM-enabled facilities management (FM): A scrutiny of risks resulting from cyber attacks," *J. Facil. Manag.*, vol. 20, no. 3, pp. 326–349, 2022. [CrossRef]
27. P. L. M. Doss and M. Gunasekaran, "Securing ResNet50 against adversarial attacks: Evasion and defense using BIM algorithm," in Proc. 2023 7th Int. Conf. Intell. Comput. Control Syst. (ICICCS), Madurai, India, May 2023, pp. 1381–1386. [[CrossRef]]

28. A. Stankovic, and M. Marjanovic, "CNN vulnerability on untargeted adversarial attacks," in Proc. 2024 Telecommun. Forum (TELFOR). New York: IEEE, 2024, pp. 1–4. [CrossRef]
29. S. Liu, Z. Zhang, X. Zhang, and H. Feng, "F-MIFGSM: Adversarial attack algorithm for the feature region," in Proc. IEEE ITAIC, Vol. 9. New York: IEEE, 2020, pp. 2164–2170. [CrossRef]
30. A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, et al., "Adversarial training for free!," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 3353–3364, 2019. [[CrossRef]]
31. S. N. Ashraf, R. Siddiqi, and H. Farooq, "Autoencoder-based defense mechanism against popular adversarial attacks in deep learning," *PLOS One*, vol. 19, no. 10, e0307363, 2024. [CrossRef]
32. N. Mahomed et al., "Preliminary report from the World Health Organisation Chest Radiography in Epidemiological Studies project," *Pediatr. Radiol.*, vol. 47, no. 11, pp. 1399–1404, 2017. [CrossRef]
33. P. R. BMS, V. Anusree, B. Sreeratcha, P. K. Ra, and S. D. Dunston, "Analysis of the effect of black box adversarial attacks on medical image classification models," in Proc. 2022 3rd Int. Conf. Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, Aug. 2022, pp. 528–531, IEEE [[CrossRef]]
34. D. S. Wankhede, C. J. Shelke, V. K. Shrivastava, R. Achary, and S. N. Mohanty, "Brain tumor detection and classification using adjusted InceptionV3, AlexNet, VGG16, VGG19 with ResNet50-152 CNN model," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 10, no. 1, 2024. [CrossRef]
35. S. R. Shah, S. Qadri, H. Bibi, S. M. W. Shah, M. I. Sharif, and F. Marinello, "Comparing inception V3, VGG 16, VGG 19, CNN, and ResNet 50: A case study on early detection of a rice disease," *Agronomy*, vol. 13, no. 6, p. 1633, 2023. [CrossRef]
36. G. Sharma, A. Vijayvargiya, and R. Kumar, "Comparative assessment among different convolutional neural network architectures for Alzheimer's disease detection," in Proc. 2021, IEEE UPCON. New York: IEEE, 2021, pp. 1–6. [CrossRef].
37. K. Kansal, and S. Sharma, "Predictive deep learning: An analysis of Inception V3, VGG16, and VGG19 models for breast cancer detection," in Proc. Int. Adv. Comput. Conf., 2024, pp. 347–357. [CrossRef].
38. N. Veni and J. Manjula, "VGG-16 architecture for MRI brain tumor image classification," in *Futuristic Communication and Network Technologies: Select Proceedings of VICFCNT 2021*, Volume 1, Singapore: Springer Nature Singapore, 2023, pp. 319–328. [[CrossRef]]
39. F. Pistorius, D. Grimm, F. Erdösi, and E. Sax, "Evaluation matrix for smart machine-learning algorithm choice," in Proc. Int. Conf. Big Data Analytics Pract. (IBDAP). New York: IEEE, 2020, pp. 1–6. [CrossRef]
40. S. M. Basha, and D. S. Rajput, "Survey on evaluating the performance of machine learning algorithms: Past contributions and future roadmap," in *Deep Learn. Parallel Comput. Environ. Bioeng. Syst.*, Cambridge, United States of America: Academic Press, 2019, pp. 153–164. [CrossRef]
41. G. Naidu, T. Zuva, and E. M. Sibanda, "A review of evaluation metrics in machine learning algorithms," in Proc. Comput. Sci. Online Conf., Springer, Cham: Springer International Publishing, 2023, pp. 15–25. [CrossRef]

Surekha M, Ph.D. Scholar in Computer Science and Engineering, Sharda University. Plot No. 32-34, Knowledge Park III, Greater Noida, Uttar Pradesh 201310; Areas of Interest: machine learning, cyber security, srtificial intelligence. M.Tech (CSE) completed at Dr. A P J Abdul Kalam University, Lucknow, Uttar Pradesh, India. Pursuing a Ph.D. from Sharda University. She is working as an Assistant Professor at JSSATE, C-20/1, Sector 62, Noida-201301, Uttar Pradesh, India. Over 10 years of teaching experience in the field of Computer Science and Engineering. One SCI-indexed journal and five Scopus-indexed international conference publications.

Dr. Anil Kumar Sagar is currently working as a Professor in the Department of Computer Science Engineering at the School of Engineering and Technology, Sharda University, India. Dr. Anil Kumar Sagar obtained his doctorate from JNU, New Delhi, in the area of Ad-hoc Networks. He earned his B.E-Computer Science & Engineering from G B Pant Engineering College, Pauri Garhwal, and his M.Tech from JSSATE Noida. He formerly served as the Dean Academics at Raj Kumar Goel Institute of Technology, Ghaziabad. He has also worked as a Member of Board of Studies in the Computer Science Department at Galgotias University, Greater Noida and RKGIT Ghaziabad. He is a member of the editorial board/review committee for many international/national journals and has served as a program/organizing committee member for several conferences.

Dr. Vineeta Khemchandani is currently working as Dean of the School of Computer Applications and Technology. She has overall experience of around 27.5 years in the field of IT in various organizations. Her experience spans software development, academics, administration, and Research, in various capacities. She has been associated with Galgotias University since 2023. She holds a Doctorate (Ph.D.) in Computer Science (2011) and a Postgraduate degree in Computer Applications (1996). Additionally, she holds a Diploma in Banking Technology from the Indian Institute of Banking and Finance, Mumbai. She has completed publications of 30 research papers, authored 4 books, contributed to 6 book chapters, and reviewed 1 book. She has guided 1 Ph.D.