IRAQI Academic Scientific Journals
العراقية المجلات الأكاديمية العلمية
University of Kut Journal
ISSN (E): 2616 - 7808 II ISSN (P): 2414 - 7419
js.uokut.edu.iq
k.u.c.j.sci@alkutcollege.edu.iq
Vol. 11 , Issue 1 , June 2026

Enhancement Robustness of Breast Cancer Models Against Adversarial Attacks

Hayder Al-Dujaili [1]

Abstract

Adversarial attacks pos a critical threat to the reliability of deep learning models, especially in medical imaging, where small pixel-level perturbations can make severe diagnostic misclassifications. This research discusses the vulnerability of a breast cancer histopathology classification model to adversarial attacks and proposes a robust defense framework. Using the BreaKHis 400X dataset, a DenseNet121-based transfer learning model achieved 87.16% accuracy on clean data. But under a Fast Gradient Sign Method (FGSM) attack with ε = 0.05, accuracy decreased to 19.45%, with an attack success rate of 80.55%. To handle these issues, a highly accurate defense model was developed, integrating adversarial training with a denoising module and adversarial detection. The proposed model significantly improved robustness that raising sensitivity to 96.75% and specificity to 90.91% on clean data, while dramatically reducing false negatives. The results show that combining adversarial training with targeted preprocessing can effectively enhance model resilience by offering a practical pathway toward more secure and reliable deep learning systems in clinical settings.

Keywords: Adversarial Attacks, Breast Cancer, Deep Learning, FGSM

Affiliation of Author

[1] Department of Computer Science, Faculty of Arts, Sciences and Technology in Lebanon, Lebanon, Beirut,

[1]Hayder.sami6000@gmail.com

[1] **Corresponding Author**

Paper Info.

Published: Jun. 2026

تعزيز موثوقية نماذج سرطان الثدي في مواجهة الهجمات المعادية

حيدر الدجيلي [1]

المستخلص

تُشكّل الهجمات الخصومية (Adversarial Attacks) تهديدًا بالغ الخطورة على موثوقية نماذج التعلّم العميق، ولا سيما في مجال التصوير الطبي، حيث يمكن لاضطرابات طفيفة على مستوى البكسل أن تؤدي إلى أخطاء تشخيصية جسيمة. يناقش هذا البحث قابلية نموذج تصنيف أنسجة سرطان الثدي للتأثر بالهجمات الخصومية، ويقترح إطارًا دفاعيًا قويًا للتصدي لها. وباستخدام مجموعة بيانات BreaKHis 400X، حقّق نموذج تعلّم بالنقل قائم على DenseNet121 دقّة بلغت 87.16% على البيانات النظيفة. إلا أنه عند التعرّض لهجوم طريقة إشارة التدرّج السريع (FGSM) بقيمة ε=0.05، انخفضت الدقّة إلى 19.45%، مع معدل نجاح للهجوم قدره 80.55%. ولمعالجة هذه التحديات، تم تطوير نموذج دفاعي عالي الدقّة يدمج بين التدريب الخصومي ووحدة لإزالة الضجيج وآلية لاكتشاف الهجمات الخصومية. وقد حسّن النموذج المقترح المتانة بشكل ملحوظ، إذ رفع الحساسية إلى 96.75% والخصوصية إلى 90.91% على البيانات النظيفة، مع تقليل كبير في حالات السلبية الكاذبة. وتُظهر النتائج أن الجمع بين التدريب الخصومي والمعالجة المسبقة الموجّهة يمكن أن يعزّز بفاعلية صمود النماذج، ويوفّر مسارًا عمليًا نحو أنظمة تعلّم عميق أكثر أمانًا وموثوقية في البيئات السريرية.

الكلمات المفتاحية: الهجمات الخصومية، سرطان الثدي، التعلم الآلي، هجوم طريقة إشارة التدرج السريع

انتساب الباحثين

[1] قسم علوم الحاسب، كلية الآداب والعلوم والتكنولوجيا في لبنان، لبنان، بيروت

[1]Hayder.sami6000@gmail.com

[1] **المؤلف المراسل**

معلومات البحث

تأريخ النشر : حزيران 2026

Introduction

Nowadays, there has been rapid development in technology, especially in artificial intelligence. Deep learning, which is a part of artificial intelligence, has brought about a major transformation in several fields, particularly in the medical domain.

In oncology, especially breast cancer, which is one of the most common cancers among women, deep

learning techniques based on convolutional neural networks have been widely used. These models have high capability of extracting patterns from histological and mammographic images, have been used for early detection and accurate tumor classification [1][2]. This early detection enables the patient to receive the appropriate treatment at the right time and reduces the risk that this cancer poses to a woman's life.

Despite this achievement in the field of early tumor detection using deep learning, adversarial attacks have become one of the most important challenges threatening the reliability of these systems. These attacks are defined as slight pixel-level changes in the image that cause the model to misclassify, leading to misdiagnosis that affects the patient's health and life, compromises clinical decision-making, and increases the rate of false alarms.

Based on the above, this research focused on the varying impact of these attacks on the classification accuracy of the deep learning model, where the FGSM attack was used, followed by building an improved defense model capable of resisting these attacks.

The proposed methodology relied on two scenarios: the first scenario analyzed the effect of the FGSM attack on the accuracy of a deep learning model based on convolutional neural networks, where this attack led to the model's accuracy decreasing to less than 20 percent. The second scenario improved the deep learning model by combining adversarial training with the addition of a set of preprocessing operations to enhance the model's performance. The defense model showed a noticeable improvement in accuracy on clean data and increased accuracy under attack compared to the original model. In summary, the evaluation results showed that this model forms an effective framework for enhancing the reliability of deep learning systems against adversarial attacks.

Related Works

According to the importance of a generative adversarial network (GAN) in generating unseen adversarial attack patterns, the authors integrated this algorithm with the target classifier into the attack generation framework. It allows to learn perturbations that closely follow the distribution of real medical images. The method was implemented on four major medical imaging datasets, especially breast histology. They achieved high attack success rates of 81.36%, 95.23%, 77.75%, and 51.67%, respectively. These results show that even highly accurate deep learning diagnostic models remain vulnerable to subtle adversarial manipulations [3].

To appear the impact of adversarial attacks' effectiveness on deep learning models, this study applied the Fast Gradient Sign Method (FGSM) adversarial attack to CNNs trained on both brain tumor MRI images and the MNIST dataset. Initially, without attack, the model achieved 67% accuracy on brain tumor classification, but under FGSM attack with $\varepsilon = 0.3$, the tumor sample was misclassified with 853% confidence, appearing extremely susceptible to perturbations. For MNIST, model confidence and accuracy decreased progressively as ε increased, with one correctly classified digit (87% confidence) being misclassified with 83% confidence after perturbation. The results clearly show how adversarial noise can drastically distort predictions in both medical and simple image datasets by emphasizing the critical need for strong defensive strategies in clinical deep learning applications [4].

A similar study in the same domain investigates

how adversarial perturbations can compromise both classification accuracy and interpretability in breast cancer classification-based deep learning systems. First, the authors train a ResNet-50–based Multi-Task Learning (MTL) model to classify breast nodules as benign or malignant while also generating Grad-CAM importance maps for visual explanation. They then apply small imperceptible adversarial perturbations, which are scaled between 0 and 1 compared to the 0–255 intensity range of ultrasound images. They evaluate how these attacks affect both model predictions and the corresponding explanation maps. Their proposed methodology includes analyzing two scenarios: are normal scenario and a perturbation that causes the model to misclassify the image with high confidence despite minimal visible change. The findings show that even when the MTL model achieves high performance (99.09% accuracy), extremely small adversarial noise can either distort the importance maps or completely change the diagnostic prediction, exposing a critical vulnerability in both CNN inference and interpretability tools used in clinical decision support [5][6].

According to show importance of handling adversarial attacks that have bad effectiveness on the model's performance. The authors used three diagnostic imaging modalities: CT scans for lung nodule classification, mammograms for breast lesion detection, and MRI scans for identifying brain metastases. The models were exposed to small pixel-level adversarial perturbations only 0.004 in magnitude (normalized 0–1) to measure the resulting drop in classification performance. The findings show that all three models have misclassification under adversarial attacks, with accuracy collapsing to 25.6% (CT), 23.9% (mammogram), and 6.4% (MRI). However, applying iterative adversarial training substantially improved robustness by raising accuracies to 67.7%, 63.4%, and 87.2%, respectively [7].

Based on building defense methods against adversarial attacks, this paper proposed a novel method that enhances adversarial training by simultaneously learning from clean and adversarial perturbed images while using a feature-correlation objective to suppress spurious features and strengthen clinically relevant ones. They used two real-world mammography datasets include 9,548 samples, and the researchers built diagnostic models and conducted extensive evaluations. This method outperformed several similar studies' baselines and demonstrated strong generalizability across datasets by improving model resilience against adversarial attacks without sacrificing accuracy on standard data [8].

Another defense strategy to improve model robustness. The authors demonstrate that a transfer-learning-based model trained to classify breast pathology images achieves high accuracy (98.72%) on clean data but suffers dramatic drops in performance (to 10.99%) under adversarial perturbations, especially FGSM. To mitigate this, they introduce a defense by superimposing Gaussian noise on the input images during training and retraining the model, which preserves high accuracy on clean data (98.08%) while substantially reducing the impact of adversarial attacks, with accuracy under attack improving to 27.47%, representing a 16.48% gain compared to the original model. The findings reveal that while deep learning systems can achieve excellent diagnostic performance where using simple preprocessing defenses like noise augmentation can meaningfully enhance security and reliability [9].

Based on defense systems against adversarial attacks, this study proposed a semi-supervised computer-aided diagnosis (CAD) system for mammographic breast mass classification using Virtual Adversarial Training (VAT) to leverage information from unlabeled images. The authors combine supervised loss for labeled data with a virtual adversarial loss applied to unlabeled data, which generates small perturbations to improve model robustness and generalizability. They used two CNN architectures, a large and a small CNN, which were trained on a dataset of 1,024 breast mass images with varying proportions of labeled data (20%, 40%, 80%). The VAT-based models outperformed standard CNNs when 40% and 80% of data were labeled by achieving classification accuracies of 0.740 ± 0.015 and 0.760 ± 0.015, indicating that unlabeled data can effectively enhance performance [10].

A new method is proposed to handle a new adversarial attack Prompt2Perturb (P2P), which is a language-guided adversarial attack method that generates realistic adversarial breast ultrasound images by optimizing learnable text prompts within a diffusion model's text encoder. Unlike previous studies of retraining the diffusion model or relying on fixed-norm pixel perturbations, P2P directly updates the text embeddings, enabling efficient creation of subtle, imperceptible perturbations that still mislead classifiers. The proposed solution reduces computation by modifying only the early steps of the reverse diffusion process that preserves structural details and ensures high image quality. The authors evaluated their method based on three breast ultrasound datasets. P2P produces adversarial images that are more natural in appearance and harder to detect, achieving superior performance in FID, LPIPS, and attack success rate compared to state-of-the-art techniques. The results show that P2P is highly effective, clinically realistic, and especially valuable in data-limited medical imaging environments by enabling strong attacks without the need for domain-specific pre-trained models [11].

Another similar study proposed an enhancement on the ResNet-50 model to improve adversarial robustness on small breast ultrasound datasets. The researchers iteratively generate adversarial examples during training to encourage smoother decision boundaries, while the DM layer by replacing the first max-pooling layer in ResNet-50, removes unstable features and helps the model learn perturbation-resistant representations. The approach is evaluated on a breast ultrasound dataset of 1,190 images and tested against three common attacks: FGSM, PGD, and CW. The result demonstrates that this method achieves state-of-the-art robustness on all adversarial attacks, improving F1-scores over RST and MI-RST baselines by up to 10.93%, while the DM layer alone boosts RST's adversarial F1-scores has values of +36.66% for PGD and +58.51% for CW. Additionally, SimCLR pretraining further enhances MIRST's robustness, raising F1-scores by 18.06% (FGSM), 5.46% (PGD), and 2.84% (CW). Overall, the method consistently strengthens defense performance without sacrificing generalization, demonstrating its suitability for building more secure and reliable breast ultrasound diagnostic models [12].

Most existing studies focus on either demonstrating the vulnerability of deep learning models to adversarial attacks or proposing individual defense mechanisms such as adversarial training, noise augmentation, or robust feature learning, as shown in **Table (1)**. Although a high number of solutions are proposed still work report

high attack success rates or improved robustness, many of them are limited to single defense strategies, specific attack types, or small-scale datasets. So, this paper attempts adversarial detection and preprocessing-based defenses by leaving a gap in developing end-to-end robust systems suitable for real clinical deployment.

Table (1): Comparison between Related Works.

Ref	Study Focus	Method	Dataset	Key Findings	Importance
[3]	Attack	GAN-based adversarial attack	Breast histology + 3 medical datasets	Attack success rates: 81.36%, 95.23%, 77.75%, 51.67%	Shows GANs can generate realistic unseen adversarial patterns by exposing vulnerabilities
[4]	Attack	FGSM on CNN	MNIST	Accuracy dropped from 67% to misclassification, high confidence errors	Demonstrates high sensitivity of CNNs to simple gradient-based attacks
[5][6]	Attack + Explain ability	ResNet-50 MTL + Grad-CAM	Breast ultrasound	Tiny perturbations distort predictions and explanation maps, even at 99.09% accuracy	Reveals the vulnerability of both prediction and interpretability tools
[7]	Attack + Defense	Pixel-level attacks + Adversarial training	CT, Mammogram, MRI	Accuracy dropped to 25.6%, 23.9%, 6.4%, improved to 67–87% after defense	Appears necessity of adversarial training
[8]	Defense	Enhanced adversarial training + feature correlation	Mammography (9,548 images)	Improved robustness without accuracy loss	Strengthens clinically relevant features, good generalization
[9]	Defense	Gaussian noise augmentation	Breast pathology images	Accuracy under attack improved 10.99% to 27.47%	Shows simple preprocessing can improve robustness
[10]	Defense (Semi-supervised)	Virtual Adversarial Training (VAT)	Mammography	Accuracy up to 0.760 ± 0.015 using unlabeled data	Demonstrates the benefit of unlabeled data in robustness

[11]	Advanced Attack	Prompt2Perturb (Diffusion + text prompts)	Breast ultrasound datasets	Higher attack success, better FID & LPIPS	Introduces language-guided, realistic medical attacks
[12]	Defense	ResNet-50 DM layer adversarial training	Breast ultrasound	F1-score improved up to +58.51% (CW attack)	Strong state-of-the-art robustness with no generalization loss

Methodology

The proposed system includes pipelines to achieve the required enhancement, as shown in Figure (1).

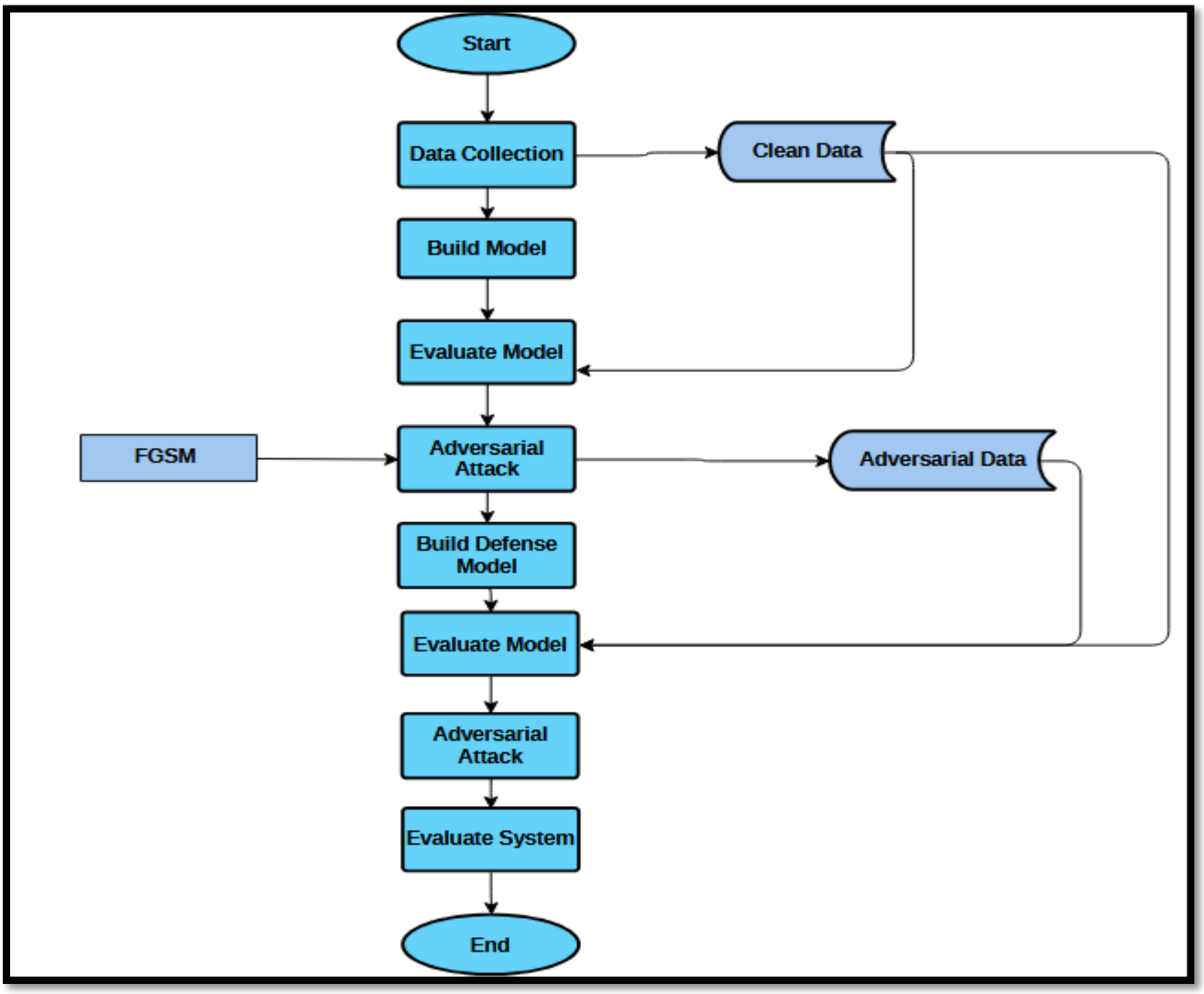

Figure (1): Methodology Pipelines.

1. Data Collection

The system was implemented on BreaKHis 400X breast histopathology dataset, obtained from Kaggle and organized into a standardized directory structure containing separate training and testing splits. The dataset includes two diagnostic classes are benign and malignant, as shown **Figure (2)**. A total of 1693 microscopic images were identified, distributed into 1148 training images (371 benign and 777 malignant) and 545 testing images (176

benign and 369 malignant). This distribution reflects the natural imbalance present in clinical breast tissue samples, particularly the predominance of malignant cases. The dataset's well-structured organization and high-resolution imaging make it suitable for evaluating deep learning models in histopathological tumor classification.

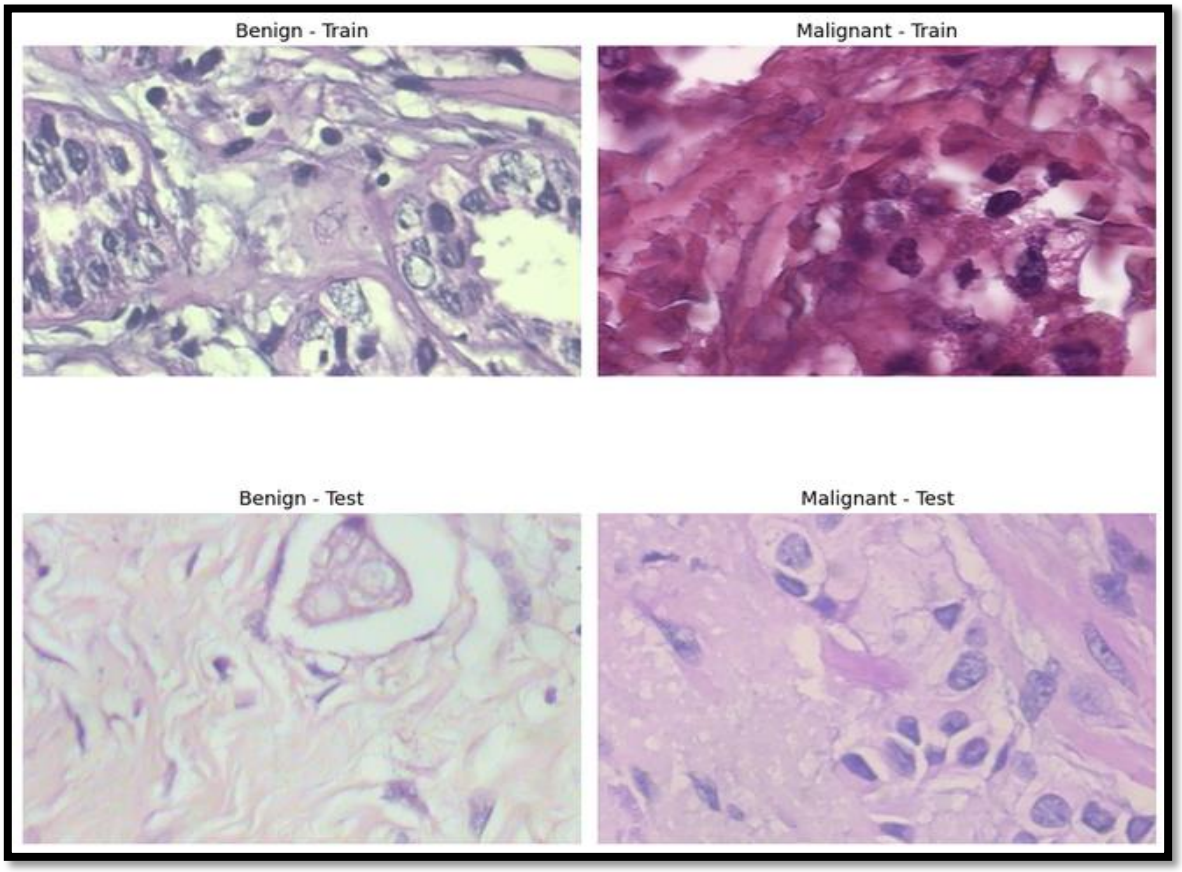

Figure (2): Samples of the Clean Dataset

2. Build Model

Build DenseNet121, which includes a convolutional neural network originally trained on ImageNet. DenseNet121 is designed with dense feature connectivity, allowing each layer to receive information from all preceding layers that enhances feature reuse, improves gradient flow, and enables stronger representation learning with fewer parameters. In this study, the pretrained backbone was frozen to preserve its high-level visual features, while a new classification head was added to adapt the model specifically for binary breast cancer recognition. This combination provides a strong balance between computational efficiency and classification accuracy.

Transfer learning is used to accelerate training and improve accuracy by reusing the pretrained visual features of DenseNet121 and fine-tuning only the task-specific layers. As shown in Table (2).

Table (2): Base Model Hyperparameter

Category	Description
Model	DenseNet121 (Transfer Learning)
Total Parameters	7,249,410

Trainable Parameters	295,554
Non-Trainable Parameters	6,953,856
Loss Function	CrossEntropyLoss
Optimizer	Adam (LR = 0.001)
Learning Rate Scheduler	ReduceLROnPlateau (patience = 5)
Epochs	200
Batch Size	32
Training Images	1003
Validation Images	145
Testing Images	545

3. Evaluation Process

The defense and base model was evaluated by using some classification metrics such as F1 score, recall, precision, and confusion matrix.

4. Implement the FGSM Attack

This study implemented FGSM, which is one of the most widely used single-step adversarial attacks designed to mislead deep learning classifiers by introducing imperceptible perturbations to input images. FGSM works by computing the gradient of the loss function with respect to the input and then shifting the image in the direction that maximally increases the model's prediction error.

$$\boldsymbol{x_{\text{adv}} = x + \epsilon \cdot \text{sign}\left(\nabla_x J(\theta, x, y)\right)} \qquad (1)$$

The attack hypermeter is shown as **Equation (1).**

1- x_{adv} : Adversarial image by adding a small perturbation.
2- x : Original image.
3- $J(\theta, x, y)$: The sign of the gradient of the loss.
4- ϵ : Controls the perturbation magnitude. This attack exploits local linearity in deep networks by creating adversarial inputs that appear visually unchanged to humans yet reliably fool the model. In this study assigned 0.05 for ϵ.

5. Adversarial and Clean Data

Before building a defense model must prepare new data, which includes adversarial and clean data. The dataset consists of 1,148 breast cancer images, split into 1,003 images for training and 145 for validation, while 545 images were reserved for testing. To enhance defense, the training set incorporated a mixture of clean samples and 30% adversarial examples generated using FGSM with $\varepsilon = 0.05$. This combination ensured that the defense model was exposed to realistic perturbations during training by allowing it to learn both denoising and adversarial pattern suppression. The dataset reflects a clinically meaningful distribution of benign and malignant cases, enabling reliable evaluation of classification performance under clean and adversarial conditions. Some samples are shown in Figure (3).

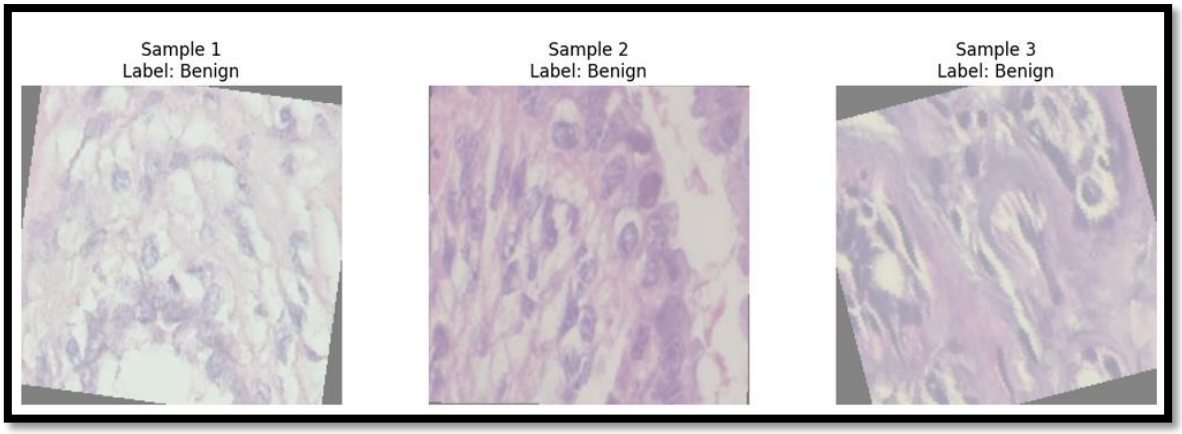

Figure (3): Samples of the Clean and Adversarial Dataset

6. Build Defense Model

The proposed defense model includes several integrated modules: a CNN-based denoiser, an adversarial detector, the main classifier and a regularization component. The model was trained on clean and adversarial data. The model hyperparameters are shown in Table (3).

Table (3): Defesnes Model Hyperparameter

Parameter	Value
Total parameters	7,798,817
Trainable parameters	7,789,281
Batch size	32
Learning rate	0.0005
Optimizer	AdamW
Scheduler	Cosine Annealing Warm Restarts
Epochs	40
Adversarial ratio	30% of training data
FGSM epsilon	0.05

Results

This section summarizes findings that were obtained from the implementation of the proposed methodology.

The classification report illustrates the performance of the DenseNet121 model across the two diagnostic classes are benign and malignant. Metrics such as precision, recall, and F1-score have moderate values reflect the models' ability to detect malignant patterns with high sensitivity while maintaining strong precision for benign predictions. In addition to, recall has high value for malignant tumors is particularly important in the clinical context, as it minimizes the likelihood of missing cancerous cases. The strong F1-scores across both classes indicate a reliable balance between false positives and false negatives by confirming that the model generalizes well to unseen breast histopathology images, as shown in Figure (4).

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
Benign	0.8081	0.7898	0.7989	176
Malignant	0.9008	0.9106	0.9057	369
accuracy			0.8716	545
macro avg	0.8545	0.8502	0.8523	545
weighted avg	0.8709	0.8716	0.8712	545

Figure (4): Classification Report for Base Model

Another important metric is the confusion matrix, that provides a detailed examination of model behavior by showing how many images from each class were correctly or incorrectly classified. In breast cancer domain analysis, this study emphasizes minimizing false negatives, because incorrectly labeling a malignant sample as benign can delay urgent medical treatment. This metric results show that the model achieves a high number of true positives and true negatives with relatively few misclassifications that indicate to robust performance. The low false-negative rate reflects the effectiveness of DenseNet121 in capturing fine-grained tissue abnormalities typical of malignant tumors. Such outcomes reinforce the model's suitability for decision-support roles in medical imaging. As shown in Figure (5).

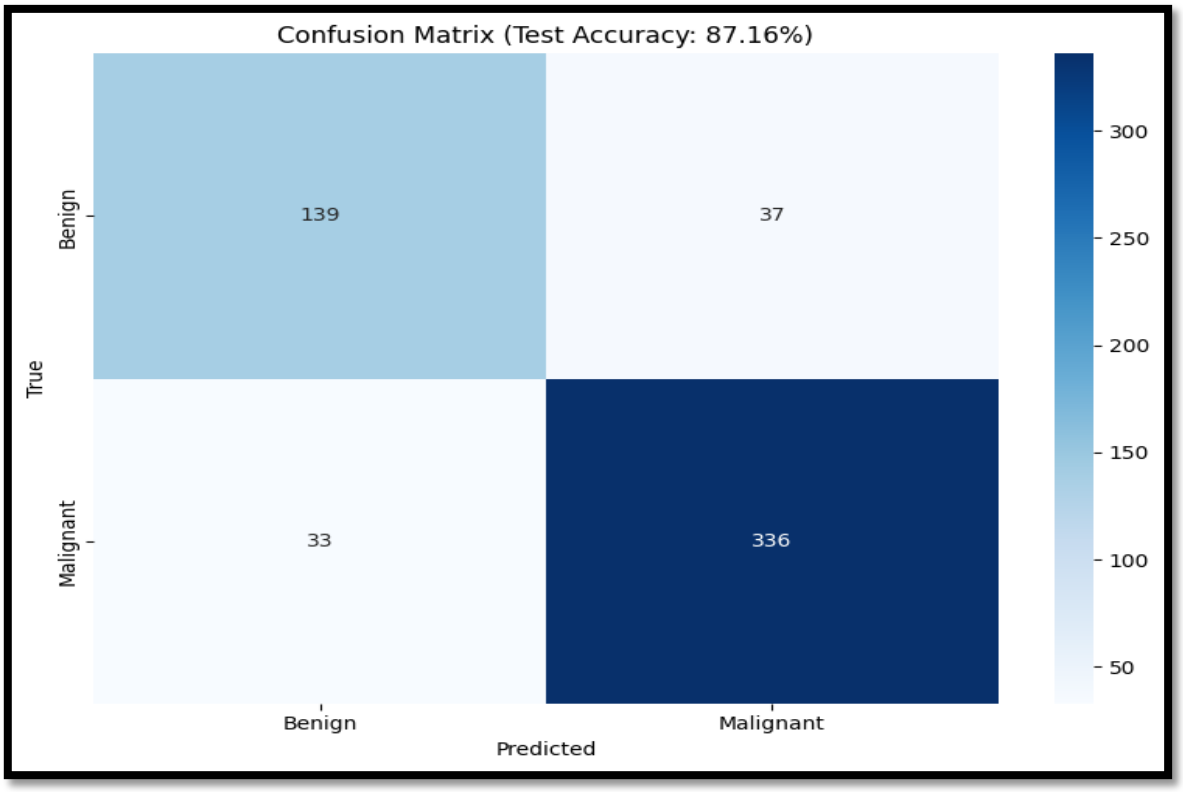

Figure (5): Confusion Matrix for Base Model

In summary, the transfer learning model demonstrated strong performance in the classification of breast cancer histopathology images. Its high accuracy, strong evaluation metrics and low misclassification rates highlight its effectiveness in identifying malignant and benign tissue patterns.

The examples of attack samples illustrate how small gradient-based noise, although subtle, can shift the predicted breast cancer class from the true category into an incorrect one, as shown in Figure (6). These samples demonstrate the deceptive nature of FGSM, where the adversarial images retain their structural and clinical characteristics, yet the model interprets them as belonging to a different diagnostic category.

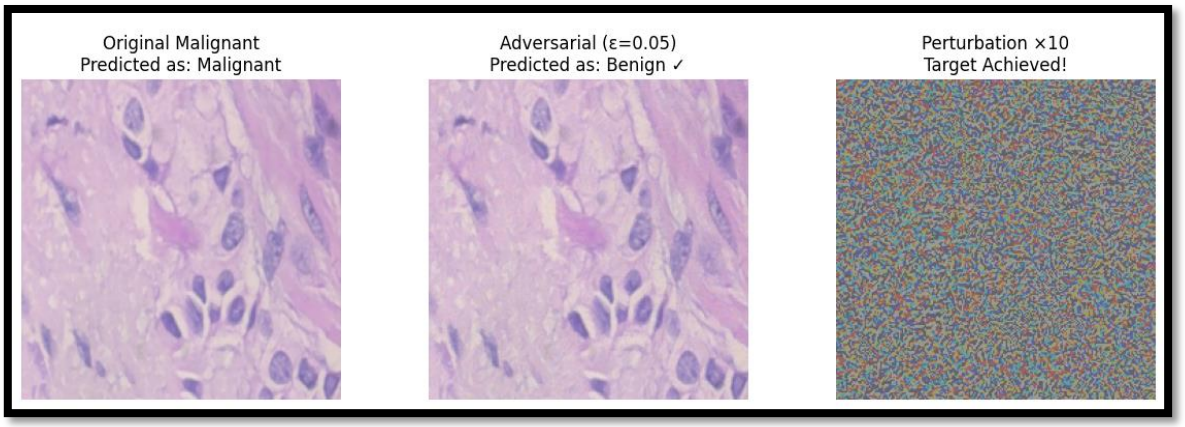

Figure (6): Samples of FGSM Attack

The experiments result show the effectiveness of FGSM on the breast cancer classification model. Before applying the attack, the model achieved an accuracy of 87.16%, suggesting strong baseline performance. But after implemented FGSM accuracy dropped drastically to 19.45%, representing a reduction of 67.7 percentage points. In the other hand, based on attack success rate of 80.55%, showing that a large majority of the input images were successfully manipulated, as shown in Figure (7). Such a severe decline highlights the model's vulnerability to adversarial perturbations and emphasizes the need for robust training strategies.

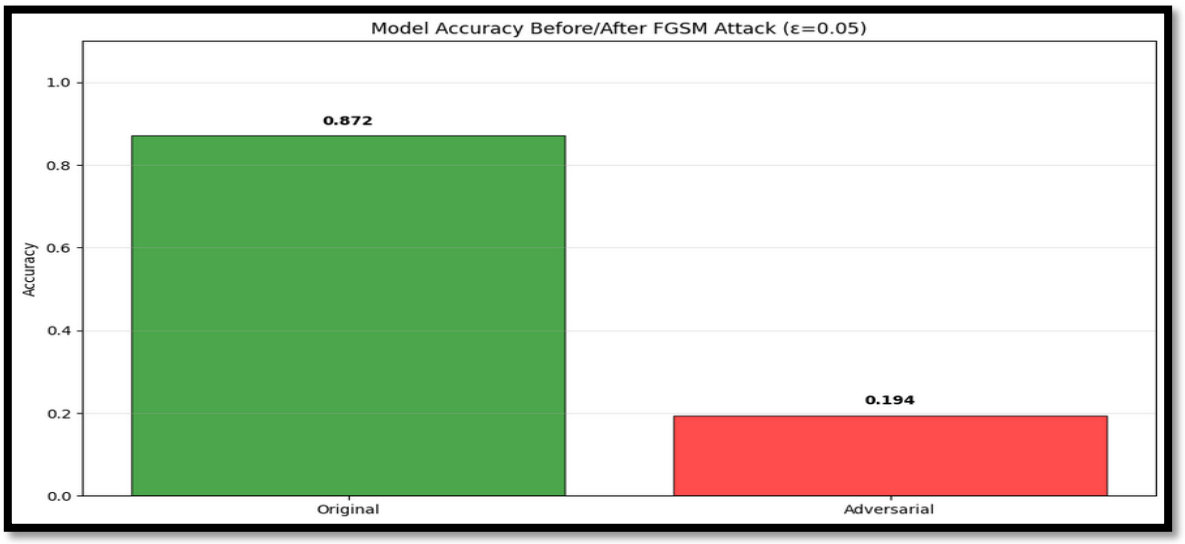

Figure (7): Base Model Accuracy Before and After FGSM Attack

The comparison between the baseline and defense model based on the confusion matrix shows the original model achieved a sensitivity of 88.35% and a specificity of 84.66%, as reflected in its confusion matrix benign correctly predicted = 149, malignant = 326. After implementing the defense architecture, the clean-data confusion matrix shows significantly enhanced performance with sensitivity increasing to 96.75% and specificity to 90.91%. False negatives were reduced dramatically from 43 to 12 and precision improved to 95.71%. These results demonstrate that the defense model provides a more reliable diagnostic output, particularly for malignant cases where minimizing missed detections is critical as shown in Figure (8).

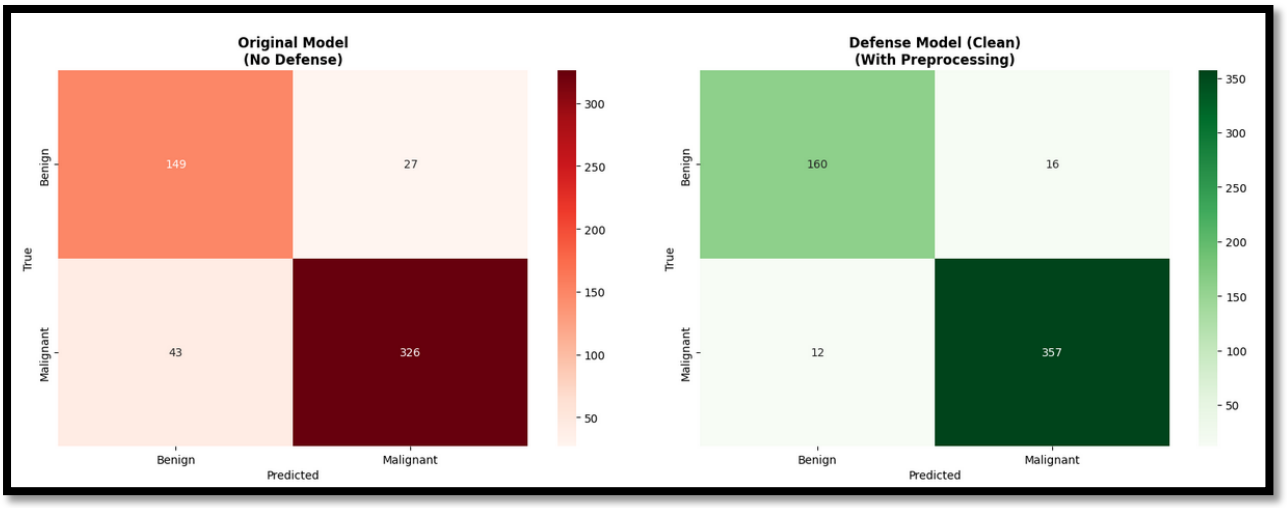

Figure (8): Comparison between Base and Defense Model Based on Confusion Matrix

The comparison between base and defenses model as shown in **Figure (9).** The comparison clearly shows the effectiveness of the proposed defense model in enhancing robustness against adversarial attacks. While the original model achieved a reasonable clean test accuracy of 87.16%, it suffered a severe performance degradation under adversarial conditions with accuracy dropping to only 19.45% and an attack success rate of 80.55% that indicate high vulnerability to FGSM perturbations. In contrast, the defense model not only improved clean data performance to 94.86% but also showed a substantial increase in adversarial accuracy to 74.31%, reflecting a gain of 54.86%. Most importantly, the attack success rate was reduced dramatically to 25.69% that represents a significant decrease of 54.86%. The comparison ensures that integrating adversarial training with targeted preprocessing and detection mechanisms can effectively mitigate adversarial threats while simultaneously improving overall model reliability by making the proposed approach well-suited for deployment in safety-critical clinical diagnostic systems.

.

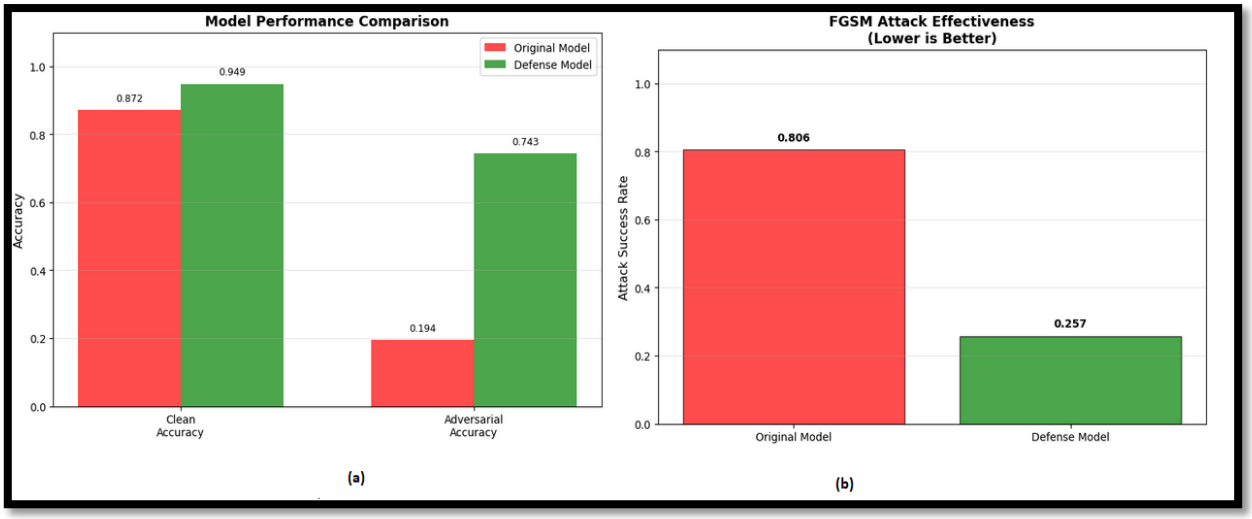

Figure (9): (a) Comparison between Base and Defense Model Based on Accuracy, (b) Comparison between Base and Defense Model Based on FGSM Attack Effectiveness

Conclusion

This study shows the profound vulnerability of deep learning-based diagnostic systems to adversarial attacks, even when they exhibit high performance on clean medical images. The FGSM attack successfully degraded the classification accuracy of a DenseNet121 model from 87.16% to 19.45%, underscoring the ease with which adversarial perturbations can mislead state-of-the-art networks. In response, a defense framework was proposed that combines adversarial examples during training alongside a dedicated denoising and detection module. The defense model improved robustness under attack by reducing false negatives and increasing both sensitivity and specificity. Future work should search different attack strategies such as PGD to add more complexity and real-time defense mechanisms to further strengthen the security.

References

[1] Laith, M., & Zuhair, H. (2025). Intelligent cyber attacks detection approaches in Internet of Medical Things environment: Current challenges and future solutions. *Alkut University College Journal*, 2025, 873–890.

[2] Akar, R. A., Al-Salhi, A. A., Safi, T. B., Ashour, N. A., Sherif, K. H., Mez'al, N. F., Mansour, M. H., & Najem, F. A. (2025). Applications of artificial intelligence in analyzing and identifying beneficial bacteria and their role in human health and the environment. *Alkut University College Journal*, 2025, 1038–1045.

[3] Kanca, E., Gulsoy, T., Ayas, S., & Kablan, E. B. (2024). Generating robust adversarial images in medical image classification using GANs. In *Proceedings of the 2024 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1–5). IEEE..

[4] Jhajharia, K., & Shrivastava, Y. K. (2024). Adversarial attacks and defenses on deep learning models. In *Proceedings of the 3rd International Conference for Advancement in Technology (ICONAT)* (pp. 1–5). IEEE.

[5] Rasaee, H., & Rivaz, H. (2021). Explainable artificial intelligence and susceptibility to adversarial attacks: A case study in

classification of breast ultrasound images. In *Proceedings of the IEEE International Ultrasonics Symposium (IUS)*. IEEE.

[6] Joel, M. Z., et al. (2023). Comparing detection schemes for adversarial images: Applications to medical imaging. *Cancers*, 15(4), Article 1123.

[7] Sun, S., Xian, M., Vakanski, A., & Ghanem, H. (2022). MIRST-DM: Multi-instance robust self-training with drop-max layer for robust classification of breast cancer. *arXiv*.

[8] Hao, D., Arefan, D., Zuley, M., Berg, W., & Wu, S. (2024). Adversarially robust feature learning for breast cancer diagnosis. *arXiv*. https://arxiv.org/abs/2402.08768.

[9] Li, Y., et al. (2023). Adversarial attack and defense in breast cancer deep learning systems. *Bioengineering*, 10(3), Article 325.

[10] Chen, X., Wang, X., Zhang, K., et al. (2022). Virtual adversarial training for semi-supervised breast mass classification. *arXiv*.

[11] Medghalchi, Y., Heidari, M., Allard, C., Sigal, L., & Hacihaliloglu, I. (2024). *Prompt2Perturb (P2P): Text-guided diffusion-based adversarial attacks on breast ultrasound images*. University of British Columbia.

[12] Sun, S., Xian, M., Vakanski, A., & Ghanem, H. (2022). MIRST-DM: Multi-instance robust self-training with drop-max layer for robust classification of breast cancer. *arXiv*.