Evaluating the Resilience of U-Net Deep Learning Architecture to Adverserial Perturbations

Aya Mohamed Abdelrahman
Cyber Security Department
The University of Alexandria
Alexandria, Egypt
cds.Ayamohamed80245@alexu.edu

Aya Abdelmoneim Mohamed
Artificial Intelligence Department
The University of Alexandria
Alexandria, Egypt
cds.ayaabdelmonem00444@alexu.edu.eg

Malak Mahmoud Aref
Artificial Intelligence Department
The University of Alexandria
Alexandria, Egypt
cds.malakmahmoud45867@alexu.edu.eg

Nouran Mohamed Hemdan
Artificial Intelligence Department
The University of Alexandria
Alexandria, Egypt
cds.NouranMohamed1349@alexu.edu.eg

Nouran Abdelsalam Mohamed
Cyber Security Department
The University of Alexandria
Alexandria, Egypt
cds.NouranAbdelsalam77349@alexu.edu.eg

Ranwah Gamal Asala
Cyber Security Department
The University of Alexandria
Alexandria, Egypt
cds.ranwahgamal2022@alexu.edu.eg

***Abstract*—Breast cancer is a significant global health issue, with early detection being crucial for successful treatment outcomes. In this paper, we present a methodology for breast cancer detection using ultrasound scan images combined with the U-Net architecture for semantic segmentation. The dataset consists of 780 ultrasound images categorized into normal, benign, and malignant classes. We preprocess the data, build and train a U-Net model for segmentation, and evaluate its performance. Additionally, we explore the robustness of the model against adversarial attacks such as Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD). Our experiments demonstrate the effectiveness of the proposed approach in accurately segmenting breast ultrasound images and detecting cancerous regions. The model achieves promising results in terms of evaluation metrics, suggesting its potential for clinical applications in breast cancer diagnosis.**

I. Introduction

Breast cancer is one of the most prevalent forms of cancer among women worldwide, with early detection playing a crucial role in improving patient outcomes and reducing mortality rates [1]. Mammography has traditionally been the primary screening modality for breast cancer. However, ultrasound imaging has emerged as a valuable adjunctive tool, particularly for evaluating dense breast tissue and characterizing suspicious lesions [2]. In recent years, there has been growing interest in leveraging machine learning techniques to enhance the diagnostic capabilities of ultrasound images for breast cancer detection.

The U-Net architecture, initially developed for biomedical image segmentation, has shown remarkable success in various medical imaging tasks, including breast cancer detection [3]. By enabling precise delineation of tumor boundaries, U-Net facilitates accurate classification and segmentation of breast ultrasound images, thereby aiding clinicians in diagnosis and treatment planning.

In this paper, we propose a methodology for breast cancer detection using ultrasound scan images and U-Net architecture for semantic segmentation. Our objectives are twofold: first, to develop a robust segmentation model capable of accurately delineating cancerous regions from ultrasound images, and second, to evaluate the model's performance and assess its robustness against adversarial attacks.

The remainder of this paper is organized as follows: Section II provides an overview of the dataset used in our study and the preprocessing steps applied to prepare the data for modeling. Section III describes the U-Net architecture and its adaptation for breast cancer detection. Section IV presents the experimental setup, including model training, evaluation metrics, and adversarial attack techniques applied to the model. Section V discusses the results of our experiments and analyzes the performance of the proposed methodology. Finally, Section VI concludes the paper with a summary of key findings and directions for future research.

II. Dataset

A. Description

The dataset used in this study is the Breast Ultrasound Images Dataset (BUSI), sourced from a public repository. Breast cancer is one of the most common causes of death among women worldwide. Early detection helps in reducing the number of early deaths. The data reviews the medical images of breast cancer using ultrasound scan. Breast Ultrasound Dataset is categorized into three classes : normal, benign, and malignant images. Breast ultrasound images can produce great results in classification, detection, and segmentation of breast cancer when combined with machine learning.

The data collected at baseline include breast ultrasound images among women in ages between 25 and 75 years old. This data was collected in 2018. The number of patients is 600 female patients. The dataset consists of 780 images with an average image size of 500*500 pixels. The images are in PNG

format. The ground truth images are presented with original images. The images are categorized into three classes, which are normal, benign, and malignant.

B. Data Preprocessing

The preprocessing steps involved normalization of pixel values to the range [0, 1] and resizing all images to 128x128 pixels. Additionally, augmentation techniques such as rotation, flipping, and zooming were applied to increase dataset diversity.

C. Samples

Sample images from the dataset, depicting normal, benign and malignant lesions along with their ground truth masks, are illustrated in the following Figures. The dataset was partitioned into training and test sets with proportions of 90% 10% respectively.

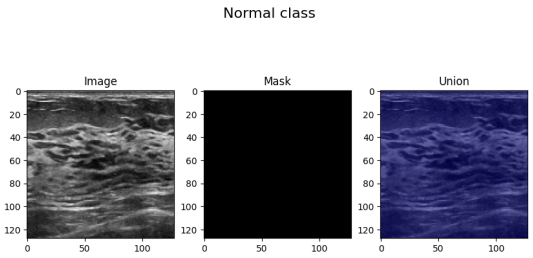

Fig. 1.

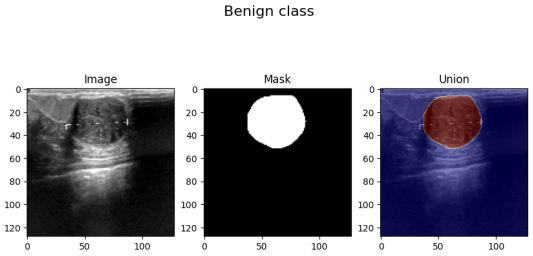

Fig. 2.

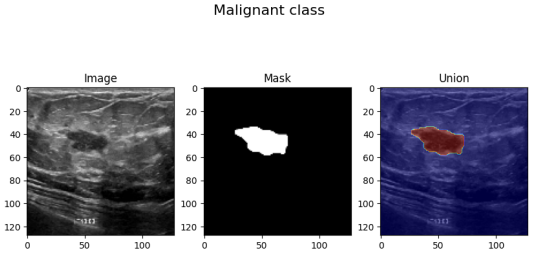

Fig. 3.

III. Model

In this section, we describe the architecture of the proposed model used for breast ultrasound image segmentation. The model architecture is based on the U-Net convolutional neural network, which has shown effectiveness in various image segmentation tasks.

The U-Net architecture consists of an encoder-decoder structure with skip connections. The encoder extracts hierarchical features from the input image, while the decoder generates the segmentation mask. Skip connections help preserve spatial information and enable precise localization of features.

The model is composed of the following components:

A. Convolutional Blocks

The convolutional block consists of two consecutive convolutional layers with ReLU activation functions and same padding. These layers learn spatial features from the input image and increase the depth of the network. The kernel size is set to 3×3 with a He normal initializer.

B. Encoder Block

The encoder block comprises a convolutional block followed by max-pooling. The convolutional block extracts features, while max-pooling reduces the spatial dimensions of the feature maps. This process helps the model capture hierarchical features at different scales.

C. Decoder Block

The decoder block consists of transposed convolutional layers to upsample the feature maps followed by concatenation with skip connections from the corresponding encoder block. This architecture enables the decoder to recover spatial information lost during encoding.

D. Final Layer

The final layer is a 1×1 convolutional layer with a sigmoid activation function. It produces the binary segmentation mask indicating the presence of breast lesions.

The model is trained using the Adam optimizer with binary cross-entropy loss. During training, the weights are initialized using the He normal initializer.

The overall architecture of the model is depicted in Figure 4.

The model is implemented using the TensorFlow framework and trained on a dataset consisting of benign and malignant breast ultrasound images with corresponding masks.

IV. Adversarial Attacks

Three commonly used adversarial image generation methods—Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD)—were employed to create adversarial images on our dataset. Each method aims to maximize the DL model's classification error while minimizing the perceptual difference between the adversarial image and the original image. All adversarial image generation methods are bounded by a predefined perturbation

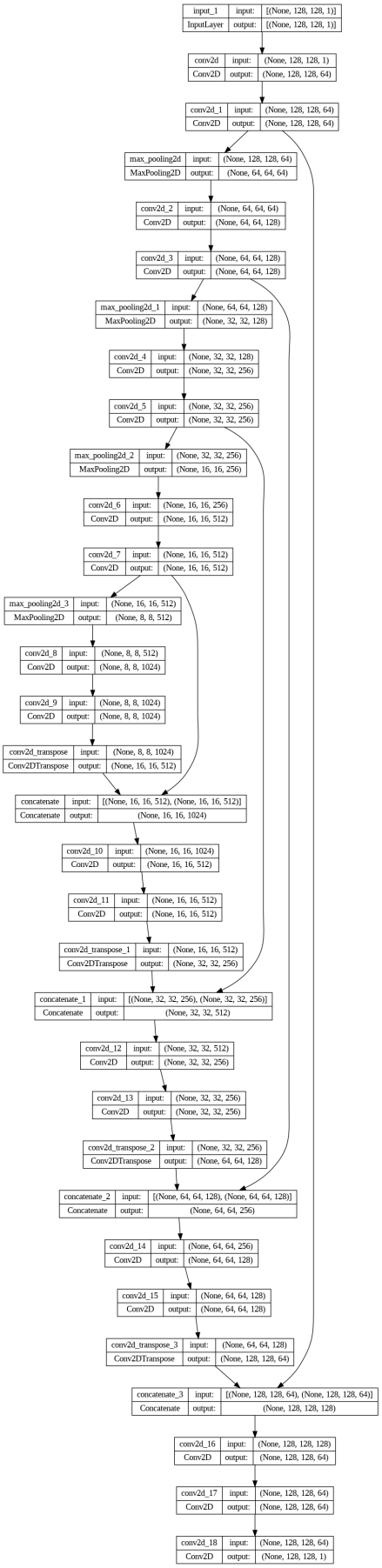

Fig. 4. Architecture of the proposed U-Net model for breast ultrasound image segmentation.

size ϵ, representing the maximum change to pixel values in an image. Vulnerability to adversarial images was evaluated by comparing changes in model performance compared to the baseline (without any adversarial images) under various perturbation sizes.

A. Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) is a white-box attack technique that perturbs input data based on the sign of the gradient of the loss function with respect to the input. It aims to maximize the loss to generate adversarial examples.

The single-step FGSM attack perturbs the original example by a fixed amount along the direction (sign) of the gradient of the adversarial loss. Given an input image x, perturbation size ϵ, loss function J, and target label y, the adversarial image x_{adv} can be computed as:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(y, x)) \tag{1}$$

where ϵ is a small constant representing the magnitude of the perturbation.

B. Basic Iterative Method (BIM)

The Basic Iterative Method (BIM) is an iterative variant of FGSM. It performs multiple small perturbations to the input image while ensuring that the resulting adversarial examples remain within an ϵ-ball around the original image. BIM iteratively updates the perturbed image based on the gradient of the loss function.

$$\mathbf{x}^t = \text{Clip}_{\mathbf{x},\epsilon}\left\{\mathbf{x}^{t-1} + \alpha \,\text{sign}\left(\nabla_{\mathbf{x}} J(\mathbf{x}^t, \mathbf{y})\right)\right\} \tag{2}$$

C. Projected Gradient Descent (PGD)

Projected Gradient Descent (PGD) is a more robust variant of iterative attacks that performs multiple steps of gradient descent with a small step size. PGD explicitly projects the perturbed image back into the ϵ-ball after each step to ensure it remains close to the original image.

$$\mathbf{x}^t = \prod_{\epsilon}\left(\mathbf{x}^{t-1} + \alpha \,\text{sign}\left(\nabla_{\mathbf{x}} J(\mathbf{x}^t, \mathbf{y})\right)\right) \tag{3}$$

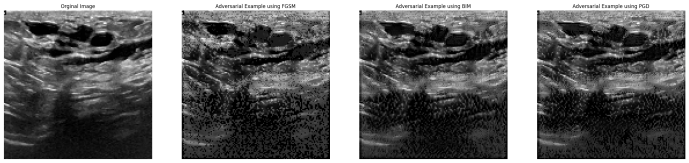

Fig. 5. Examples of clean images and their adversarial counterparts generated using FGSM, PGD, and BIM attack methods.

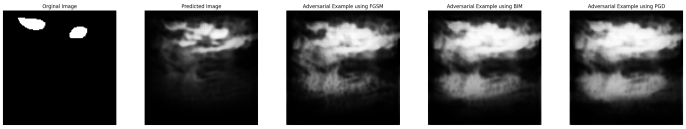

Fig. 6. Examples of the mask images and their adversarial counterparts generated using FGSM, PGD, and BIM attack methods.

V. Results and Achievements

A. Evaluation Metrics

The performance of the model was evaluated using various metrics including accuracy, loss, Mean Intersection-Over-Union (IoU), precision, and recall. These metrics provide insights into the model's performance both on clean and adversarial examples generated using FGSM, BIM, and PGD.

TABLE I
Model Performance Metrics

Attack Method	Clean Data	FGSM	BIM	PGD
Accuracy (%)	93.77	84.65	82.11	80.80
Loss	0.1720	0.3570	0.4105	0.4334
Mean IoU (%)	66.66	48.83	46.34	45.28
Precision (%)	43.8	25.6	23.9	23.6
Recall (%)	79.0	21.3	16.4	14.8

B. Robustness Analysis

The model's accuracy decreased and loss increased on adversarial examples compared to clean data. PGD generated the most challenging adversarial examples, resulting in the lowest accuracy and highest loss. The model's performance degraded as the perturbation magnitude increased, indicating vulnerability to adversarial attacks.

C. Insights and Conclusions

Our study demonstrates the vulnerability of the proposed U-Net model to adversarial attacks. The results highlight the need for enhancing the model's robustness through techniques like adversarial training or input preprocessing. Future research could focus on developing more resilient architectures or adaptive defenses to mitigate the impact of adversarial attacks.

References

[1] Marina Z. Joel, BS, Sachin Umrao, Enoch Chang, Rachel Choi,Daniel X. Yang, James S. Duncan, Antonio Omuro, Roy Herbst, Harlan M. Krumholz, Sanjay Aneja, "Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology. JCO Clin Cancer Inform." 2022 Feb;6:e2100170.

[2] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[3] Jin Xu; Zhendong Cai; Wei Shen, "Using FGSM Targeted Attack to Improve the Transferability of Adversarial Example"

[4] Hokuto Hirano and Kazuhiro Takemoto, "Simple Iterative Method for Generating Targeted Universal Adversarial Perturbations"

[5] Yingpeng Deng; Lina J. Karam, Universal Adversarial Attack Via Enhanced Projected Gradient Descent"