International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Adversarial Attacks on Medical Image Diagnosis Models And its Mitigation Techniques

Santhoshini Sahu[1], R. Lakshmi Prasanna[2], S. Neelima[3], B. Sai Siddhardha[4], Ch. Kavya[5], B. Dileep[6], G. Samarthu[7]

[1]Assistant Professor GMR Institute of Technology, Vizianagaram, India santhoshini.s@gmrit.edu.in
[2,3,4,5,6,7] Computer Science and Engineering GMR Institute of Technology, Vizianagaram, India
[2]21345A0505@gmrit.edu.in , [3] 21345A0502@gmrit.edu.in, [4] 20341A0530@gmrit.edu.in, [5] 20341A0565@gmrit.edu.in, [6] 20341A0550@gmrit.edu.in,
[7] 20341A0560@gmrit.edu.in

DOI: https://doi.org/10.55248/gengpi.5.0324.0774

ABSTRACT—

In recent years, deep learning (DL) models have become integral to numerous sectors, revolutionizing our daily lives and workflows. Particularly in healthcare, DL have brought about a paradigm shift in medical diagnosis through innovative image analysis capabilities. These computational tools offer exceptional precision and speed, significantly enhancing diagnostic accuracy and facilitating early disease detection. However, the widespread adoption and reliance on these models have opened the door to new forms of vulnerabilities, notably adversarial attacks. In the context of medical image diagnosis, adversarial attacks pose an alarming threat. They can manipulate diagnostic models into misinterpreting imaging data, leading to false positives or negatives. Such errors can result in misdiagnosis, delayed treatment, or unnecessary interventions, impacting patient safety and the overall quality of healthcare. This project delves into the landscape of adversarial attacks in the context of medical image diagnosis. This study looks at popular adversarial attack strategies. A Convolutional Neural Network (CNN) EfficientNet B0 Model trained to categorise Alzheimer's brain MRI images is subjected to the Vertical Perturbation attack, Fast Gradient sign Method (FGSM), and Square attack. Following that, it goes into one of the most common adversarial defence approaches, adversarial training. The performance of the model that has been trained on adversarial instances is next tested against the previously described attacks, and recommendations to improve the neural network's robustness are therefore supplied based on the experiment findings.

Keywords—Adversarial attacks, Medical image diagnosis, vulnerabilities, Adversarial training, resilience.

I. INTRODUCTION

Deep learning, a branch of artificial intelligence, is proving to be a valuable tool in transforming medical diagnosis and healthcare. By efficiently analyzing intricate patterns in vast datasets, deep learning has driven significant progress in medical imaging analysis, disease detection, and personalized medicine. Specifically, in medical imaging, deep learning algorithms have exhibited outstanding precision in interpreting MRI scans, X-rays, and CT scans, facilitating early disease identification and enhancing patient outcomes. Additionally, these models can analyze a wide range of patient data to support in disease diagnosis and prognosis, ultimately leading to more precise and timely diagnoses and treatment decisions. While deep learning models hold immense potential in medical diagnosis, they are vulnerable to manipulation through adversarial attacks. These attacks can have serious consequences, potentially causing misdiagnosis and harming patient well-being. Adversarial attacks involve attempting to deceive a model into making incorrect predictions by providing it with carefully crafted inputs known as adversarial examples. These examples are modified versions of legitimate data that are indistinguishable to humans but lead the model to misclassify them with great certainty. In the realm of medical imaging analysis, even slight modifications to medical images, such as adding imperceptible noise or making minor, targeted alterations, can cause deep learning algorithms to misinterpret the data, possibly resulting in misdiagnosis or incorrect treatment suggestions. Similarly, when it comes to disease diagnosis and personalized medicine, adversaries could tamper with patient data to trick the models into making inaccurate predictions or diagnoses.

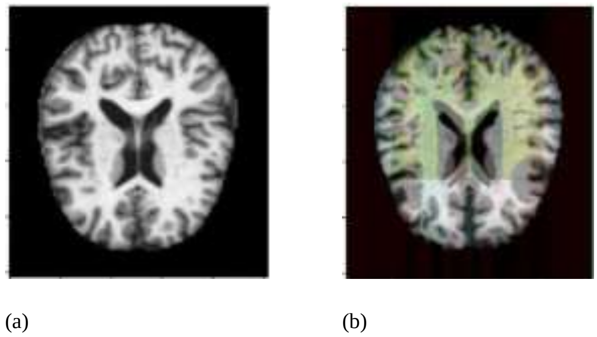

Fig. 1. Sample brain MRI images: (a) original MRI (b) Pertubuted MRI

To combat this threat, two key approaches are crucial: adversarial training and robust mechanisms against adversarial attacks.

In this paper, the analysis of vertical perturbation attack, square attack and FGSM attack, the approaches for adversarial attacks are carried out on a CNN model trained to classify brain MRI of Alzheimer's disease . The model is trained on the Alzheimer's diseases dataset. After that, adversarial training is used as a defence tactic.

II. LITERATURE SURVEY

Sneha Shukla et al. [1] proposed MedIS, a dynamic loss selection based adversarial attack intended primarily for use against medical image segmentation models. By offering instructions on how to swap out non-differentiable layers in these models for differentiable approximations, they solve the problem they present. They show through experimental results that their attack achieves a higher ASR (Attack Success Rate), outperforming previous attacks. They also state that their future work will be focused on creating more resilient deep learning models for medical images and investigating adversarial defense techniques. The research also emphasizes the authors' plan to investigate adversarial defense techniques, which is a proactive move to improve the security and dependability of medical image segmentation models against changing adversarial threats.

Tsai et al. [2] conducted a one-pixel attack on various medical image datasets, such as COVID-19, Chest, derma and pneumonia to generate adversarial images that can fool a trained model. It also performed a multi-pixel attack on the COVID-19 dataset to explore the impact of the number of perturbed pixels. A differential evolution (DE) algorithm is used in the study to find the optimal pixel location and intensity value to modify the original image. It also devised a robust evaluation method for the multi-label Chest dataset to assess the effectiveness of the attack. The paper reported the success rate, average confidence level, and class transformation ratio of the attack for each dataset. It found that the one-pixel attack can achieve high success rates (above 90%) for most datasets, and the multi-pixel attack can further increase the success rate and confidence level.

Lal et al. [3] introduced a framework that provides a defensive model against the adversarial speckle-noise attack which is a type of granular patterning that can be seen in radar coherent images, the adversarial training, and a feature fusion strategy, which preserves the classification with correct labelling. The paper proposes a method to improve the accuracy and robustness of deep neural networks for diabetic retinopathy detection, by using adversarial training and feature fusion techniques. The Authors generates adversarial images using different attack methods, such as FGSM, SN, and DF, and uses them to train the target model with a diversity adversarial training approach. This reduces the average attack success rates and increases the model's resilience against unknown adversarial attacks.

Koga et al. [4] have proposed This paper proposed a method for generating black-box universal adversarial perturbations (UAPs) for deep neural networks (DNNs) used in medical image classification. The paper aims to demonstrate that UAPs can be easily generated using a relatively small dataset under black-box conditions, and that they pose a serious security threat to DNN-based medical imaging systems. The paper extends the simple black-box attack (SimBA) method, which is a hill-climbing approach that estimates the loss gradient from the model outputs, to craft UAPs using a small dataset of images. The algorithm starts with no perturbations and iteratively updates the UAP using a set of search directions. The performance of the UAPs is evaluated for both nontargeted and targeted attacks. The approach evaluates the performance of the black-box UAPs on three medical image datasets (skin lesion, OCT, and chest X-ray) and shows that they achieve high fooling rates (40% to 80%) against various DNN models, regardless of the norm type of the UAPs. The findings show that black-box universal adversarial perturbations (UAPs) can be easily generated using a relatively small dataset.

Li et al. [5] introduced the topic of adversarial attack and defense for classification methods to the statistical community, and to encourage more statisticians to work on this important and exciting field. The study reviews the generation and guarding of adversarial examples, which are inputs that are intentionally perturbed to fool a classifier. It also discusses the challenges and open problems in this area, such as the lack of theoretical understanding, the trade-off between robustness and accuracy, and the evaluation of adversarial robustness. It also presents some numerical experiments to illustrate the effectiveness and limitations of different methods, using MNIST, CIFAR-10, and ImageNet datasets.

kwon et al. [6] proposed AdvU-Net, a method for generating adversarial examples that target the U-Net model, which is a deep neural network used for image segmentation in the biomedical field. Fast gradient sign method (FGSM) is used to create adversarial examples that induce mis-segmentation by the U-Net model and also analyzes the performance of adversarial examples in terms of pixel error, image distortion, epsilon value and adversarial noise. The paper analyzes the performance of the adversarial examples on the ISBI 2012 dataset, which consists of electron microscopy images of the Drosophila nervous system. Results shows that the adversarial examples generated by AdvU-Net can significantly degrade the segmentation performance of the U-Net model, with pixel errors increasing from 0.15 to 3.54 or higher as the epsilon value (which controls the amount of noise added) increases from 0.1 to 0.9. Using the proposed method, it was possible to generate adversarial example for image segmentation model used in medical field and the vulnerability of U-Net model to adversarial example is confirmed.

Bortsova et al. [7] introduced the adversarial attack vulnerability of deep learning systems used in medical image analysis (MedIA). It concentrates on situations in which adversarial examples are created using surrogate models in situations where attackers have restricted access to the target model. The study examines how various factors, including model architecture variations, weight initialization, and variations in development data, affect the effectiveness of these attacks. The paper provides insights into three medical domains and excels in examining factors that are often overlooked but affect the vulnerability of deep learning MedIA systems.

Shan et al. [8], presented a defense that is enabled by trapdoors. Analytical evidence shows that trapdoors affect adversarial attacks and cause attack inputs to resemble trapdoors. Experimental results demonstrate that trapdoor- protected models detect cutting-edge attacks with high accuracy while having no effect on standard classification in a variety of areas. Additionally, robustness against adaptive attacks is shown. This paper outlines the development, analysis, and assessment of a trapdoor-enabled defense against adversarial examples, introducing the concept of "trapdoors" as honeypots. The work demonstrates analytical proofs, empirical robustness, key properties, countermeasure efficacy, and vulnerability exploration, presenting DNN honeypots as a promising defense paradigm deserving increased research attention. This paper advocates using honeypots to bolster DNN defense against adversarial examples. Incorporating trapdoors into models proves effective in detecting state-of-the-art attacks, ensuring high detection rates while negligibly impacting normal input classification. Robustness against adaptive attacks is confirmed.

Sorin et al. [9] have presented image classification methods adversarial attacks could be dangerous, thus stronger cybersecurity precautions and moral standards are needed to ensure that deep learning technology is used safely in medicine. Adversarial assaults pose a serious threat to healthcare, particularly radiology, as they modify input data to trick models. These attacks go after algorithms and even try to fool doctors into making inaccurate diagnoses. They are divided into white-box and black-box attacks, targeted and untargeted attacks, and misclassification hazards particular to each. Focus on the classification of chest X-rays highlighted the susceptibility of deep learning models by demonstrating successful attacks that led to incorrect diagnoses. Targeted algorithms covered a wide range of applications and revealed differing resistance levels in distinct imaging modalities and networks. Adversarial attacks on natural language processing (NLP) algorithms confront difficulties because text alterations can be easily detected, however many assaults have focused on picture classification.

Puttagunta et al. [10] introduced in-depth overview of the numerous adversarial attack strategies and defense methods. The attack models (FGSM, BIM, PGD) have been applied to various deep learning models and datasets, including ResNet, AlexNet, LeNet, MNIST, CIFAR-10, LSUN, and ImageNet, for the purpose of evaluating the robustness of deep learning models in medical image classification, segmentation. The results suggest that the specific attributes of medical images and deep learning models call for customized defense strategies to ensure the robustness and reliability of these systems. Implementation of customized defense strategies in medical deep learning systems can lead to improved privacy protection, diagnostic accuracy, trust, and ethical considerations, ultimately benefiting the customers

III. MEDICAL IMAGE CLASSIFICATION

Medical imaging is now a critical tool in present-day healthcare, playing a significant role in diagnosis, treatment, and patient care. Techniques such as X-rays, CT scans, MRIs, and ultrasounds help visualize internal organs and structures, aiding in the early detection of diseases like cancer, heart disease, and bone fractures. This enables quick intervention and potentially enhances treatment outcomes. Through detailed images of internal structures, medical imaging assists in distinguishing between different disease conditions, leading to more precise diagnoses compared to relying solely on physical exams or symptoms.

A. Model

An EfficientNet B0 convolutional neural network (CNN) structure is used to analyze brain MRI pictures to detect Alzheimer's disease (AD). The efficiency of the B0 structure allows for strong performance while using resources effectively, making it suitable for medical applications. By extensively training on labeled MRI data, the model has been taught to recognize subtle, disease-specific patterns in brain composition, which could potentially assist in early and precise diagnosis of AD.

B. Dataset

The Alzheimer's disease dataset contains 1200 subjects' MRI scans with three different classes on classifying between Alzheimer's, Parkinson's and healthy control subjects.

IV. METHODOLOGY

The research was carried out using a CNN model that had been trained to classify Alzheimer's disease brain MRI images. The purpose was to understand the impact of adversarial attacks on deep learning models and the effectiveness of defense strategies.

Initially, we tested the performance of our CNN model against adversarial attacks such as vertical perturbation, square, and FGSM attacks. This evaluation was done using an EfficientNet B0 model that had not been exposed to adversarial examples during training. Following this, we implemented adversarial training in various ways and compared the model's performance when trained using different techniques against adversarial examples from vertical perturbation, square, and FGSM methods.

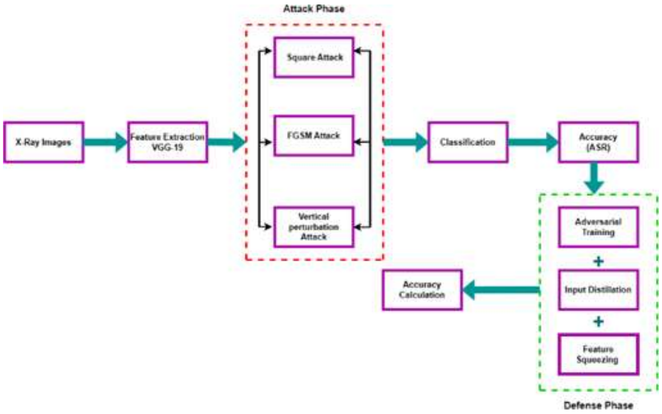

Fig. 2. Architecture of proposed System

A. Vertical Perturbation attack

A vertical perturbation attack is a type of security threat targeting machine learning models, particularly those used in classification tasks. Unlike traditional attacks that manipulate input data with horizontal shifts or alterations, vertical perturbation attacks focus on modifying the confidence scores or output probabilities assigned to individual classes by the model.

In a vertical perturbation attack, the attacker aims to manipulate the decision boundaries of the model by perturbing the feature values of instances belonging to specific classes. This perturbation is designed to cause misclassification or bias the model's predictions towards certain outcomes, often in a targeted manner.

B. Square Attack

The Square Attack technique involves systematically introducing localized, square-shaped alterations to the image. These subtle changes, often undetectable to the human eye, adjust the decision boundaries that the deep learning model has learned, pushing it towards incorrect classifications. Unlike other attack methods that rely on gradient information, the Square Attack does not depend on gradients, making it resistant to gradient masking defenses. This method utilizes an optimization process that aims to achieve two goals simultaneously: increasing the model's prediction error and minimizing the size of the perturbations.

C. FGSM Attack

The Fast Gradient Sign Method (FGSM) is a well- known and direct adversarial technique employed to deceive AI models into incorrectly classifying inputs. This method involves introducing precisely designed noise into an image, generating an adversarial instance that closely resembles the original but induces the AI to provide an inaccurate prediction.

D. Adversarial Training

Adversarial training exposes the model to both clean data and intentionally generated adversarial examples during training. By "seeing" these deceptive examples, the model learns to recognize and resist similar attacks in the future.

It works as below steps:

Generate adversarial examples

Train with both clean and adversarial data Repeat and refine

E. VGG-19

The VGG neural network architecture is widely utilized in deep applications for its feature extraction capabilities. Comprising 19 layers of both convolutional and fully connected layers, VGG19 is a robust tool for extracting detailed features from images. Its depth enables it to identify intricate patterns in visual data, enhancing image recognition and classification accuracy. In this study, to improve image analysis accuracy VGG19 for feature extraction is utilized.

V. RESULTS AND DISCUSSION

In this paper, the performance of the proposed model is calculated by taking the accuracies of the data before and after the attack and along with that accuracy after applying the defensive algorithms are taken into consideration. Attacks performed are Vertical perturbation, square attack and FGSM attack. The accuracy of the model are calculated after the attacks are performed.

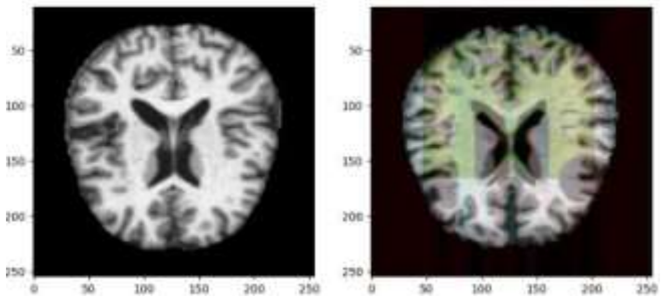

Fig.3. Addition of vertical perturbation

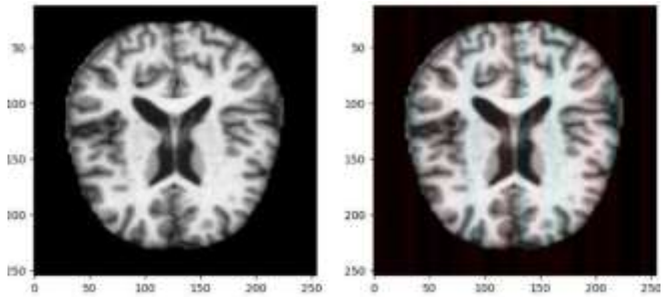

Fig.4.Image before and after square attack

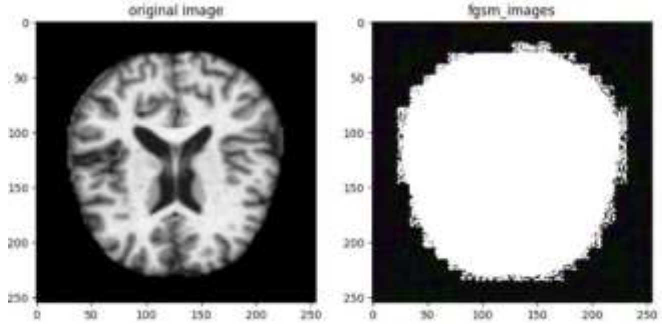

Fig.5. Image before and after FGSM attack

Attack performed	Accuracy Before	Accuracy after
Vertical perturbation	96%	56%
Square Attack	96%	38%
FGSM Attack	96%	6%

Table.1. Representing Accuracy of the model before and after the attack

VI. CONCLUSION

The experiment underscores the susceptibility of medical image diagnosis models to adversarial attacks, with notable reductions in accuracy observed: 56% accuracy with Vertical Perturbation attack, 38% with square attack, and only 6% with FGSM attack, compared to the baseline accuracy of 96%. These findings emphasize the critical importance of fortifying AI systems in healthcare. While defensive techniques like adversarial training, cycle GAN, and distillation show promise in enhancing model resilience, further research and development are essential to mitigate the risks posed by adversarial attacks and ensure the reliability and safety of deep learning-based medical diagnostics. This underscores the ongoing necessity for robust defense mechanisms in healthcare AI, warranting continuous exploration and innovation in this field.

REFERENCES

[1] Shukla, S., Gupta, A. K., & Gupta, P. (2023). Exploring the feasibility of adversarial attacks on medical image segmentation. Multimedia Tools and Applications, 1-24.

[2] J. Jung, H. Moon, G. Yu and H. Hwang, "Generative Perturbation Network for Universal Adversarial Attacks on Brain-Computer Interfaces" in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 11, pp. 5622-5633, Nov. 2023, doi: 10.1109/JBHI.2023.3303494.

[3] Muoka, G. W., Yi, D., Ukwuoma, C. C., Mutale, A., Ejiyi, C. J., Mzee, A. K., ... & Al-antari, M. A. (2023). A Comprehensive Review and Analysis of Deep Learning- Based Medical Image Adversarial Attack and Defense. Mathematics, 11(20), 4272.

[4] Zbrzezny, A. M., & Grzybowski, A. E. (2023). Deceptive Tricks in Artificial Intelligence: Adversarial Attacks in Ophthalmology. Journal of Clinical Medicine, 12(9), 3266.

[5] Nallamolu, S., & Padmanabhuni, S. (2023). A Privacy Preserving Generative Adversarial Network for Image Data. In ITM Web of Conferences (Vol. 53). EDP Sciences.

[6] Vera Sorin, Shelly Soffer, Benjamin S. Glicksberg, Yiftach Barash, Eli Konen, Eyal Klang, Adversarial attacks in radiology – A systematic review, European Journal of Radiology, Volume 167, 2023, 111085, ISSN 0720-048X.

[7] Puttagunta, M. K., Ravi, S., & Nelson Kennedy Babu, C. (2023). Adversarial examples: attacks and defences on medical deep learning systems. Multimedia Tools and Applications, 1-37.

[8] Hu, J., Wen, J., & Fang, M. (2023). A survey on adversarial attack and defense of deep learning models for medical image recognition. Metaverse, 4(1), 17.

[9] Fontanella, A., Antoniou, A., Li, W., Wardlaw, J., Mair, G., Trucco, E., & Storkey, A. (2023). ACAT: Adversarial Counterfactual Attention for Classification and Detection in Medical Imaging. arXiv preprint arXiv:2303.15421.

[10] Dong, J., Chen, J., Xie, X., Lai, J., & Chen, H. (2023). Adversarial Attack and Defense for Medical Image Analysis: Methods and Applications. arXiv preprint arXiv:2303.14133.

[11] Sorin, V., Soffer, S., Glicksberg, B. S., Barash, Y., Konen, E., & Klang, E. (2023). Adversarial attacks in radiology–A systematic review. European Journal of Radiology, 111085.

[12] Croce, F., Singh, N. D., & Hein, M. (2023). Robust Semantic Segmentation: Strong Adversarial Attacks and Fast Training of Robust Models. arXiv preprint arXiv:2306.12941.

[13] Yao, Q., He, Z., Li, Y., Lin, Y., Ma, K., Zheng, Y., & Zhou, S. K. (2023). Adversarial Medical Image with Hierarchical Feature Hiding. *IEEE Transactions on Medical Imaging*.

[14] Letafati, M., Behroozi, H., Khalaj, B. H., & Jorswieck, E. A. (2023). Learning-Based Secret Key Generation in Relay Channels Under Adversarial Attacks. IEEE Open Journal of Vehicular Technology.

[15] Pervin, M. T., Tao, L., & Huq, A. (2023). Adversarial attack driven data augmentation for medical images. International Journal of Electrical and Computer Engineering (IJECE), 13(6), 6285-6292.