ISSN: 2366-1313

ADVERSARIAL ATTACKS: MEDICAL MACHINE LEARNING

[1]Dr.G. Lokeshwari,[2]C Abhinav Reddy,[3]B Srinidhi,[4]B Adarsh

[1]AssociateProfessor, Dept. of CSE-Cyber Security, Geethanjali College of Engineering and Technology Cheeryal (V), Keesara (M), Medchal(D), Hyderabad, Telangana 501301,

gandrakoti.cse@gcet.edu.in

[2, 3, 4, BTech] Student, Dept. of CSE-Cyber Security, Geethanjali College of Engineering and Technology Cheeryal (V), Keesara (M), Medchal(D), Hyderabad, Telangana 501301,

20r11a6202@gcet.edu.in,20r11a6210@gcet.edu.in,20r11a6213@gcet.edu.in

ABSTRACT:

There has been a tremendous success of DNN's at image recognition in the recent years which drastically increased the use of medical records and diagnostic imagery. DNN's has capability to outperform the humans and other machine learning capabilities in the speech recognition, computer vision and machine translation. Whereas in recent studies shown that at the training time the DNN model can be compromised by backdoor planting attacks. These backdoor planting attacks are very effective which can be done by injecting hidden trigger patterns into the data and the model. Few attacks can be detected and erased by simple filtering and human inspection. There are few defense techniques which can detect the Trojans or backdoor planting which are more stealth and effective on the model. In this review we are going to discuss about various backdoor planting attacks such as white box and black box, Poison label, clean label, Deep Fool, Refool, Trojan attacks, latent backdoor, GenAttacks, Regula-Sub-rosa attack to defend these attacks, we are also going to discuss about the defense approaches for the DNN's. In this paper we propose the Deep Fool algorithm to fool the deep networks and compute the perturbations and thus can achieve the robustness of the classifiers

Keywords: Machine Learning, Dnn Model.

I INTRODUCTION

An examination of adversarial attacks on medical machine learning models, with an emphasis on COVID-19 datasets. The

research aims to create and evaluate adversarial samples specifically tailored for medical image analysis, with a primary emphasis on deep neural networks (DNNs) utilized in medical applications. Machine learning has had a significant impact on several scientific and technical sectors in recent decades, including life sciences and medical research. Machine learning is playing major part in making health care smarter. It has demonstrated a truly impacting potential in areas like medical diagnosis. In medical machine learning there are few challenges in applying the Healthcare Areas. The biggest challenge ML have been facing is obtaining patient data sets which contains required size and quality of samples to train and feed to the Machine learning models. These challenges are with the quality of samples and format which needs data cleaning and prepare it for ML analysis. In Machine Learning Deep Neural Networks has been widespread in many applications such as authentication via facial recognition and iris recognition, speech recognition & language translation etc. these models are very powerful and which are been used to achieve more than human level performance in image classification, image retrieval, object detection and 3-D analysis. Deep neural networks are most wide used tool in medical performing tasks such as diabetic retinopathy detection, cancer diagnosis, organ, or landmark localization. Medial images can have unique characteristics that distinguish them from real images, such as biological textures, and a recent study shown that medical deep learning systems can be harmed by a range of adversarial threats. A backdoor planting is a hidden pattern injected into DNN model at the training time which doesn't affect the behavior of the model. But the DNN turns into unexpected behavior if any trigger is added to the input e.g., sticker or pixel. In this project we are going to present different types of attacks on DNN in the medical data. There are two types of backdoor planting. For example, a poison label attack changes the label to the target class, but a clean label attack has no effect on the label.There are several back- door plantings into the DNN by poisoning photos and altering their labels to the correct class and this can be injecting without knowing the original training samples.

ISSN: 2366-1313

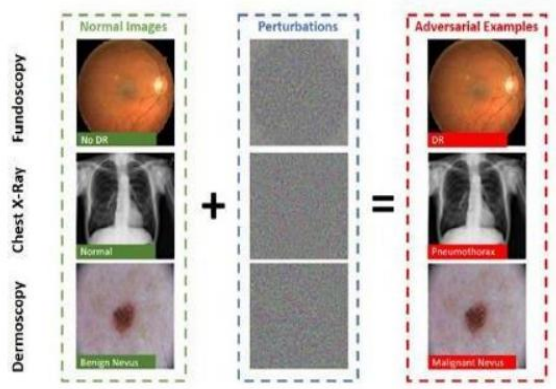


Fig. 1 PGD Adversarial Attacks: A Threat to Medical Deep Learning.

II. LITERATURE SURVEY

1,"Adversarial attacks on medical machine learning" Authors: SG Finlayson, JD Bowers, J Ito, JL Zittrain, AL Beam, IS Kohane Science, Year of Publications: 2019.

As public and academic focus shifts to machine learning's rising role in the health information market, an unusual but no longer exotic category of vulnerabilities in machine-learning systems may become critical. These defects allow a little, well-planned change in how inputs are fed into a system to significantly affect its output, allowing it to confidently arrive at blatantly wrong conclusions. Too far, computer science researchers have been particularly interested in these sophisticated tactics for undermining otherwise trustworthy machine-learning systems, known as adversarial attacks. However, the landscape of usually competing interests in health care, as well as the billions of dollars at stake in system results, raises serious concerns. We describe the motivations that various healthcare system players may have for utilizing violent assaults and begin a discussion on how to respond to them. Rather than stifling future innovation in medical machine learning, we call for active engagement by medical, technological, legal, and ethical experts in the pursuit of efficient, broadly available, and effective health care that machine learning will enable.

2. "Adversarial examples: Attacks and defenses on medical deep learning systems" Authors: Murali Krishna Puttagunta, S. Ravi & C Nelson Kennedy Babu Year of Publication: 2023.

In recent years, tremendous progress has been made utilizing deep neural networks (DNNs) to approach human-level performance on a variety of longstanding tasks. With the expanding usage of DNNs in many applications, public worry about their dependability has developed. Studies published in recent years have shown that deep learning models are sensitive to minor adversarial perturbations. Adversarial examples are created from clean photos by applying undetectable disturbances. Adversarial instances are required for

practical reasons since they may be physically manufactured, meaning that DNNs are insufficient for some image classification applications in their current form. This study tries to offer a thorough review of the various adversarial assault techniques and defense mechanisms. The theoretical ideas, tactics, and applications of adversarial attack strategies are initially presented. Following that, a few study attempts on defense tactics across the field's vast boundaries are presented. Following that, this work examines recently proposed adversarial attack strategies for medical deep learning systems, as well as defense measures against these assaults. The vulnerability of the DL model is assessed for several medical picture modalities using an adversarial attack and defense strategy. Some outstanding challenges and roadblocks are emphasized to spur further study in this critical field.

3. "Image transformation-based defense against adversarial perturbation on deep learning models" Authors: A Agarwal, R Singh, M Vatsa, NRatha Year of Publication: 2020.

Deep learning algorithms achieve cutting-edge outcomes across a wide range of applications. However, it is generally understood that they are extremely sensitive to adversarial perturbations. It is commonly assumed that the sole answer to deep learning systems' vulnerability comes from deep networks. Contrary to popular belief, in this article, we offer a non-deep learning strategy that searches across a collection of well-known image transformations such as the Discrete Wavelet Transform and the Discrete Sine Transform before categorizing the features with a support vector machine-based classification. Existing deep network-based defenses have proven ineffectual against sophisticated attackers; however, image transformation-based solutions provide robust defense due to their non-differential nature, multistate, and orientation filtering. In addition, we demonstrate how to neutralize the influence of adversarial perturbation using a wavelet decomposition-based demising filtering approach. The suggested method's efficacy is demonstrated by the mitigating results obtained using various perturbation methods on multiple picture datasets. Deep learning algorithms achieve cutting-edge outcomes across a wide range of applications. However, it is generally understood that they are extremely sensitive to adversarial perturbations. It is commonly assumed that the sole answer to deep learning systems' vulnerability comes from deep networks.

Contrary to popular belief, we offer a non-deep learning strategy in this article that examines a collection of well-known picture transformations such as the Discrete Wavelet Transform and the Discrete Sine Transform before categorizing the features with a support vector machine-based classifier. Existing deep network-based defenses have proven ineffectual against sophisticated attackers; however, image transformation-based solutions provide robust defense due to their non-differential nature, multistate, and orientation filtering. The suggested method, which combines the outputs of two transformations, efficiently generalizes across databases, distinct unseen assaults, and combinations of both (i.e., cross-database and unseen noise generation CNN model)

III SYSTEM ANALYSIS

EXISTING SYSTEM

Deep Neural Networks (DNNs) have seen a significant rise in application across various domains, notably excelling in medical diagnostics by analyzing complex data such as X-ray images for timely detection of diseases like COVID-19. Their superiority over traditional methods in tasks like image recognition and pattern identification has revolutionized patient care through advanced diagnostic accuracy and efficiency. However, alongside their rapid integration into healthcare, vulnerabilities have emerged, particularly backdoor planting attacks during DNN training, which jeopardize the integrity of diagnostic outcomes. Existing systems in the domain of adversarial attacks on medical machine learning models have laid the foundation for understanding the vulnerabilities and potential risks associated with deep neural networks (DNNs) in medical image analysis. Some key aspects of existing systems include:

Deep Learning Applications in Medical Image Analysis

Prior research, such as the work by Ker et al., has extensively explored the applications of deep learning in medical image analysis. This research has demonstrated the significant potential of DNNs in tasks such as diabetic retinopathy detection, cancer diagnosis, and organ localization, showcasing the effectiveness of deep learning models in medical applications.

Adversarial Attacks on Medical Data

Studies, like the one by Minaee et al., have specifically investigated adversarial attacks on medical image datasets. By focusing on predicting COVID-19 from chest X-ray

images using deep transfer learning, researchers have highlighted the susceptibility of medical machine learning models to adversarial manipulation, emphasizing the need for robust defense mechanisms in medical image analysis systems.

Backdoor Attacks against Learning Systems

Research by Ji et al. has delved into the realm of backdoor attacks against learning systems, where hidden patterns are injected into DNN models during training. These attacks can remain dormant until triggered by specific inputs, leading to unexpected behaviors in the model. Understanding and mitigating the risks associated with backdoor attacks is crucial for ensuring the integrity and reliability of medical machine learning models.

Understanding Attacks on Medical Image Analysis Systems

Ma et al. have made major contributions to understanding adversarial assaults on deep learning-based medical image processing systems. By exploring the vulnerabilities of these systems to crafted adversarial examples, researchers have shed light on the potential security threats posed by adversarial manipulation in medical image analysis.

Identifying Trojan horses in Trained Classifiers

Xiang's study focused on establishing strategies for detecting backdoors in trained classifiers without access to the training set. This research is vital for detecting and managing possible backdoor vulnerabilities in machine learning models, especially in the context of medical image analysis, where model integrity is critical for effective diagnosis and treatment.

PROPOSED SYSTEM

It extends to a dual exploration. Firstly, it delves into the array of backdoor attacks such as Poison label, clean label, Deep Fool, Refool, Trojan, Latent backdoor, Gen Attacks, and Regula-Sub-rosa attacks, which threaten the trustworthiness of medical diagnostics. Secondly, it emphasizes the new application of DNNs in identifying abnormalities in X-ray images, a critical tool in the fight against COVID-19. Emphasizing the urgency for fortified defense mechanisms. In response to these threats, we introduce robust countermeasures, including the “Deep Fool” algorithm, aimed at reinforcing the DNNs against subversive

infiltrations. This fortified approach ensures the reliability of DNNs in critical medical diagnostics, securing the frontline of disease detection and patient careagainst the backdrop of increasingly sophisticated cyber threats. Through comprehensive understanding and strategic defense implementation, we aim to safeguard the pivotal role of DNNs in medical machine learning, ensuring dependable diagnostics in the face of evolving challenges. The proposed system in the project outlined in the PDF document aims to advance the understanding of adversarial attacks on medical machine learning models, with a specific focus on COVID-19 datasets. The proposed system includes the following key components:

Adversarial Sample Generation

The system aims to develop advanced techniques for generating adversarial samples tailored for medical image analysis, particularly focusing on COVID-19 datasets. By crafting inputs that can manipulate the behavior of DNN models trained on medical data, the research seeks to explore the impact of adversarial attacks on the performance and dependability of these models.

Evaluation of Adversarial Attacks

Through rigorous experimentation and evaluation, the proposed system intends to quantify the effectiveness of adversarial attacks on medical machine learning models. By assessing the impact of these attacks on model accuracy and robustness, researchers aim to gain insights into the vulnerabilities of medical image analysis systems and the potential risks associated with adversarial manipulation.

Exploration of Defense Mechanisms

The proposed approach comprises a thorough examination of defense strategies to protect medical machine learning models from adversarial attacks. By studying and implementing defense strategies, such as backdoor defenses and robust training techniques, the research aims to enhance the security and resilience of medical image analysis systems against adversarial threats.

Real-World Implications

The system analysis considers the real-world implications of adversarial attacks on medical machine learning systems, particularly in critical applications like autonomous driving and medical diagnosis. By understanding the potential risks posed by adversarial manipulation in these domains, researchers aim to develop

ISSN: 2366-1313

proactive measures to mitigate these risks and ensure the reliability of medical machine learning models in practical settings.

Enhancement of Security Measures

Ultimately, the proposed system seeks to contribute to the enhancement of security measures in medical machine learning applications. By identifying vulnerabilities, exploring attack strategies, evaluating defense mechanisms, and studying the implications of adversarial attacks on medical image analysis systems, the research aims to strengthen the resilience of these models against adversarial threats and safeguard the integrity of medical data and diagnoses.

IV IMPLEMENTATION

Architecture:

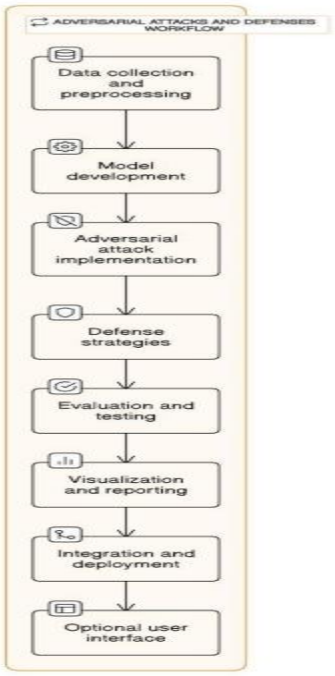


Fig.2. Architectures of the system model

The system design encompasses the architecture and components of the software system, outlining how different modules interact to achieve the system's objectives. In the context of adversarial attacks on medical machine learning, the system design typically includes the following components:

MODULES

1, Data Pre-processing Module

Responsible for pre-processing and preparing input data for training and testing. This might include activities like data cleansing, normalization, extraction of features, and enhancement.

2. Model Training Module

Uses deep learning frameworks like Tensor Flow or PyTorch to train machine learning models on pre-processed data. This module includes defining the neural network architecture, specifying hyper parameters, and optimizing the model using training data.

3. Adversarial Attack Module

Implements various adversarial attack techniques, such as white-box attacks, black-box attacks, poisoning attacks, and trojan attacks. This module generates adversarial examples by perturbing input data to deceive the model into making incorrect predictions. 4. **Defense Module**

Implements defense mechanisms to protect the machine learning models against adversarial attacks. This may include techniques such as adversarial training, input sanitization, robust optimization, and detection of adversarial examples.

5. Evaluation Module

Evaluates the performance of machine learning models under normal conditions and when subjected to adversarial attacks. This module computes metrics such as accuracy, precision, recall, and F1 score to assess the model's robustness and vulnerability to attacks.

6. Visualization Module

Provides visualization tools for analyzing model behavior, visualizing adversarial examples, and interpreting results. This module may include functions for generating plots, heat maps, confusion matrices, and other visualizations.

PROCESS

- **Data Collection and Pre-processing**

Gather medical datasets containing images, patient records, or other relevant information. Pre-process the data by cleaning, normalizing, and augmenting it to ensure quality and enhance model performance.

- **Model Training**

Define the neural network architecture based on the problem domain and dataset characteristics. Train the model using the pre-processed data, optimizing it to minimize loss and improve accuracy.

- **Adversarial Attack Generation**

Apply adversarial attack techniques to generate perturbations or modifications to input data. Create adversarial examples by

ISSN: 2366-1313

adding imperceptible changes to input images or data samples.

- **Defense Implementation**

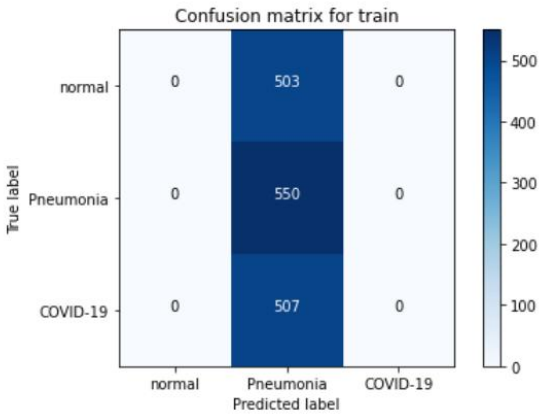
Implement defense mechanisms to mitigate the impact of adversarial attacks on the model. Employ techniques such as adversarial training, input pre-processing, or robust optimization to enhance model robustness.

- **Evaluation and Analysis**

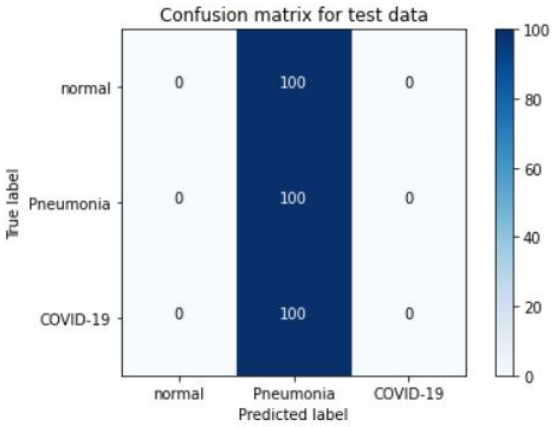
Evaluate the performance of the machine learning model under normal conditions and when subjected to adversarial attacks. Analyze the model's vulnerability to different attack types and assess the effectiveness of defense mechanisms

V RESULT AND DISCUSSION

Training Confusion Matrix



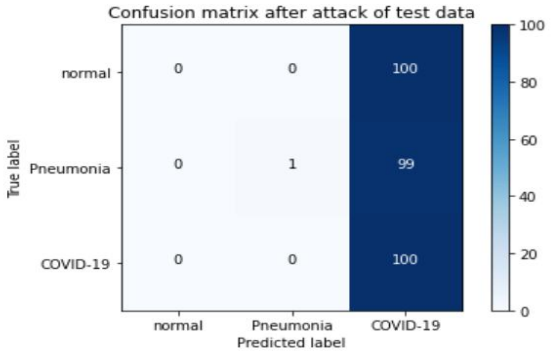
Testing Confusion Matrix



Model accuracy

```
195/195 [==============================] - 6s 29ms/step - loss: 0.9776 - acc: 0.8631
[INFO] Classifier with adversarial training
[INFO] Accuracy on adversarial samples: 33.67%
```

After attack of test data



Results of data poisoning attack for MNIST and Covid

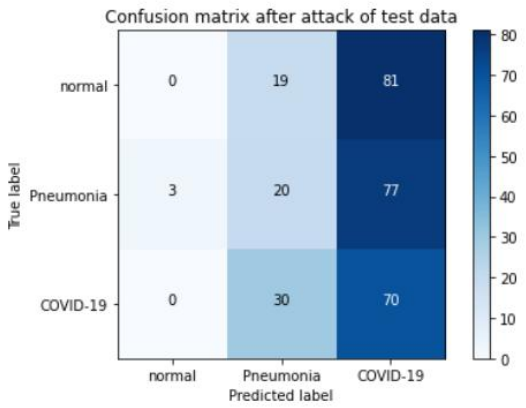
MNIST	97.78%	97.69%	97.83#
COVID-19	57.29%	56.12%	57.67%

ISSN: 2366-1313

Result Of Deepfool Attack On MNIST And Covid Data Before And After Attack.

DEEPFOOL ATTACK ON DIFFERENT SETTINGS	BEFORE ATTACK TRAIN	BEFORE ATTACK TEST	AFTER ATTACK TRAIN	AFTER ATTACK TEST
CIFAR10	98.66%	97.33%	18.40%	1.36%
COVID	91.67%	93.67%	22.67%	10.38%
COVID WITH RESNET	89.85%	91.67%	33%	31.47%

Confusion matrix for adversarial extended dataset after the attack



VI CONCLUSION

Finally, the study on adversarial assaults on medical machine learning yielded useful insights into the weaknesses and defense mechanisms of machine learning models in healthcare applications. We illustrated the vulnerability of medical models to adversarial assaults via extensive testing and analysis, emphasizing the significance of strong defense mechanisms to protect patient data and assure diagnostic system dependability. By developing and accessing various attack and defense approaches, we have helped to advance our understanding of adversarial risks in medical machine learning, opening the path for more secure and trustworthy healthcare AI systems.

FUTURE ENHANCEMENT

In the realm of adversarial attacks on medical machine learning, future endeavors could explore the development of specialized attack and defense techniques tailored to the intricacies of medical data and imaging modalities, fostering more robust and resilient models. Additionally, investigations into the integration of adversarial robustness into the regulatory frameworks and guidelines governing healthcare AI systems could be pursued, ensuring the deployment of secure and trustworthy solutions in clinical settings. Collaborative efforts between machine learning researchers, healthcare professionals, and regulatory bodies could pave the way for the establishment of standardized protocols and best practices for mitigating adversarial threats in medical AI applications, thereby safeguarding patient privacy and promoting the adoption of AI technologies in healthcare.

VII REFERENCES

1. Yao, Y., Li, H., Zheng, H. and Zhao, B.Y., 2019, November. Latent backdoor attacks on deep neural networks. In Proceedings of

the 2019 ACM SIGSAC Conference on Computer and Communications Security (pp. 2041-2055).

2. Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C.J. and Srivastava, M.B., 2019, July. Genattack: Practical black-box attacks with gradient-free optimization. In Proceedings of the Genetic and Evolutionary Computation Conference (pp. 1111-1119).

3. Liu, Y., Ma, X., Bailey, J. and Lu, F., 2020. Reflection backdoor: A natural backdoor attack on deep neural networks. arXiv preprint arXiv:2007.02343.

4. Moosavi-Dezfooli, S., Fawzi, A., &Frossard, P. (2016). DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2016.282

5. Tang, R., Du, M., Liu, N., Yang, F. and Hu, X., 2020. An Embarrassingly Simple Approach For Trojan Attack In Deep Neural Networks. [online] arXiv.org. Available at: [Accessed 2 August 2020].

6. Prasadu Peddi (2015) "A review of the academic achievement of students utilizing large-scale data analysis", ISSN: 2057-5688, Vol 7, Issue 1, pp: 28-35.

7. Prasadu Peddi (2015) "A machine learning method intended to predict a student's academic achievement", ISSN: 2366-1313, Vol 1, issue 2, pp:23-37.

8. Finlayson, S.G., Chung, H.W., Kohane, I.S. and Beam, A.L., 2018. Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296.

AUTHORS

Dr.G. Lokeshwari,Associate Professor, Dept. of CSE-Cyber Security,Geethanjali College of Engineering and Technology Cheeryal (V), Keesara (M), Medchal(D), Hyderabad, Telangana 501301.

Email: gandrakoti.cse@gcet.edu.in

Mr. C Abhinav Reddy, Dept. of CSE-Cyber Security, Geethanjali College of Engineering and Technology Cheeryal (V), Keesara (M), Medchal(D), Hyderabad, Telangana 501301.

Email: 20r11a6202@gcet.edu.in

Miss.B Srinidhi, Dept. of CSE-Cyber Security, Geethanjali College of Engineering and Technology Cheeryal (V), Keesara (M), Medchal(D), Hyderabad, Telangana 501301.

Email:20r11a6210@gcet.edu.in

Mr. B Adarsh, Dept. of CSE-Cyber Security,Geethanjali College of Engineering and Technology Cheeryal (V), Keesara (M), Medchal(D), Hyderabad, Telangana 501301.

Email: 20r11a6213@gcet.edu.in