Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Defense against Artificial Intelligence Hacking Model

Valarmathi K[1], Kousalya S[1], Janani R[2], Devi M[2]

[1]Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, Tamil Nadu, India
[2]Department of Computer Science & Business Systems, Panimalar Engineering College, Chennai, Tamil Nadu, India
Email: valarmathi_1970@yahoo.co.in

The widespread adoption of deep learning technology has led the way in a new era of AI- powered capabilities that hold the potential to revolutionize numerous aspects of our society. From healthcare to autonomous vehicles, these AI systems offer powerful solutions to complex problems. However, with this incredible power comes a vulnerability that could be exploited by malicious actors. Adversarial attacks represent a method of fooling or undermining AI technology, particularly deep learning models. Adversarial examples are crafted inputs, often images, that undergo subtle, imperceptible alterations. To the human eye, these modifications are virtually undetectable, yet they have the remarkable ability to lead AI models to make erroneous and highly confident predictions. The consequences of such a misclassification could be dire, resulting in traffic accidents and posing a severe safety concern. This is achieved through the application the Fast Gradient Sign Method (FGSM), which involves computing gradients of the loss function in relation to input images. By perturbing these images to maximize the loss, the aim is to induce misclassification. The second facet of the methodology entails Training Using a Pretrained Model, utilizing the well-established AlexNet architecture as a foundational tool. The pretrained model is specifically engineered to distinguish between original and adversarial images. It excels at recognizing intricate patterns and subtle alterations introduced by adversarial attacks. Leveraging its classification prowess, the model accurately identifies the classes to which both original and adversarial images belong. This capability positions it as a robust tool for image classification and bolsters security against adversarial manipulation.

Keywords: Adversarial attack, Hacking, Defense, AlexNet, Imperceptible, misclassification.

1. Introduction

In the ever-evolving landscape of artificial intelligence, deep learning models stand as the highpoint of computational prowess, demonstrating remarkable abilities to recognize patterns, make predictions, and process complex data. However, with great power comes an inherent vulnerability - the susceptibility of these models to adversarial attacks, where malicious entities exploit the model's inherent weaknesses to deceive, manipulate, and compromise its

predictions. This intersection of artificial intelligence and cybersecurity has given birth to a formidable realm known as 'Hacking Deep Learning Models.'

The concept of hacking deep learning models delves into the sinister art of subverting the very systems designed to enhance our understanding of the world. These attacks, leveraging sophisticated techniques, aim to manipulate the model's decision-making process, steering it towards incorrect predictions or classifications. At the heart of this manipulation lies the creation of 'adversarial examples' - subtly altered inputs crafted to deceive the model while remaining inconspicuous to human observers. The adversarial examples, often generated using methods like the Fast Gradient Sign Method (FGSM) or more complex algorithms like Variational Autoencoders, introduce imperceptible perturbations into input data, leading the model astray.

The motivation behind hacking deep learning models varies, from academic research aiming to fortify models against attacks to malicious intents such as digital impersonation, misinformation campaigns, or unauthorized access. Adversarial attacks raise critical questions about the security and robustness of AI systems, prompting researchers, developers, and cybersecurity experts to delve into the intricate dynamics of these attacks.

This exploration goes beyond mere technical prowess; it delves into the ethical and societal implications of AI security breaches. The repercussions of a hacked deep learning model can be devastating, leading to erroneous medical diagnoses, flawed autonomous vehicle decisions, or compromised cybersecurity systems. As society increasingly relies on AI for decision-making across sectors, understanding the vulnerabilities and defenses against adversarial attacks becomes paramount.

In this comprehensive work, we embark on a detailed journey into the methodologies, motivations, and countermeasures associated with hacking deep learning models. Through a nuanced examination of attack vectors, defensive strategies, and real-world implications, we aim to shed light on this clandestine domain, fostering a deeper understanding of the challenges that lie at the intersection of artificial intelligence and cybersecurity. By unraveling the dark art of hacking deep learning models, we empower ourselves to navigate the future of AI in a secure, resilient, and ethically sound manner.

2. LITERATURE SURVEY

Y. Ding et al [1],proposed a method for protecting the privacy of medical imaging data is crucial, leading to the proposal of DeepKeyGen—a deep learning-based key generation network and made a generative adversarial network (GAN) to create private keys for medical image encryption and decryption. Y. Chen et al.[2], strives to offer readers a comprehensive overview of the current state of adversarial attacks and defenses within the realm of image classification. In this context, it acknowledges the rapid advancements in deep learning and the vulnerabilities that machine learning models face when exposed to carefully crafted adversarial examples data points designed to deceive these models. K. Khullar et al [3],evaluated the performance of two state-of-the-art model architectures when subjected to adversarial attack techniques designed to deceive well-trained machine learning models and these models excel in classifying images from the CIFAR-10 dataset. Pedraza A et al.[4], made

a unique approach leveraging chaos theory, positing that deep networks exhibit chaotic behavior, with adversarial examples as a prime manifestation and he demonstrated that combining Lyapunov exponents and entropy significantly improves accuracy in detecting adversarial examples, achieving impressive results across various attack types for datasets like MNIST, Fashion-MNIST, and CIFAR 19. Y. Wang et al.[5], has concentrated on identifying adversarial inputs within image classifiers constructed using deep neural networks. His approach, which is independent of specific models, relies on observing distinctions in logit semantics between adversarial and original inputs.

A. Pandya et al.[6],explored how these models are vulnerable to adversarial attacks. Through an iterative targeted attack and the application of explanation algorithms, the author uncovers insights about the impact of attacks on interpretability methods and suggests ways to enhance their adversarial robustness. C. Xiao et al.[7],proposed a novel approach by embracing adversarial examples, turning them into a defense mechanism and the author introduced a groundbreaking approach to address this vulnerability.

Y. Wang et al.[8],introduced an innovative adversarial example detector that excels in identifying the latest adversarial attacks on image datasets. This method employs sentiment analysis to gauge the impact of adversarial perturbations on DNN feature maps. Y. Liu et al.[9], provided embedding attacks for multiple resizing methods and a universal attack for adaptability and the adversaries can embed a small target image into a benign one to create adversarial examples without directly querying the target network. Zhang YA et al.[10],introduced a defense strategy based on image compression and reconstruction and compressed input images to eliminate adversarial perturbations and then uses a super-resolution network to restore image quality, effectively countering adversarial examples.

Tian, Xuejun et al.[11],examined deep learning's limitations in image classification and introduces a game that lets users manipulate images with deep neural networks. In the game, players select a product category, causing the model to misclassify others as the chosen one, revealing AI's limitations. A.Agarwal et al.[12], proposed a well across databases and unseen attacks, achieving impressive detection accuracy and also demonstrates the effectiveness of wavelet decomposition-based denoising to neutralize adversarial perturbations across various image databases and perturbation methods. J. Ji et al.[13],proposed a multi-scale defense method to combat this challenge. It involves evolving input images using Gaussian kernels of different strengths, resulting in a multi-scale image representation and the method also monitors confidence changes during image evolution, offering insights into potential attacks. C. -Y. Lin et al.[14], introduced a method to generate adversarial face images that are virtually identical to the source images and it also leverages facial landmark detection and super pixel segmentation to guide the insertion of imperceptible adversarial noise. J. -H. Choi et al[15], demonstrated the vulnerability of deep image- to-image models, which generate output images from input images, to adversarial attacks and examined 16 deep models across five common image-to-image tasks, considering factors like the impact of attacks on output quality, transferability of adversarial examples between tasks, and the nature of perturbations.

D. Vyas et al. [16] proposed a novel feature-map approach to analyze how adversarial attacks impact individual object feature-maps in convolutional neural networks (CNNs). Inspired by the effects of adversarial perturbations on image pixels, the method calculates the weight of

each object's class activation map and generates a consolidated activation map using an innovative detection technique. S. Niu et al. [17], categorized transfer learning into four main categories and explores four learning types within each. It provides a thorough survey of the field, covering the current state of the art, emerging trends, applications, and unresolved challenges. S. Rezaei et al. [18], demonstrated that with only access to the publicly available pre-trained model, an attacker can launch a brute force attack that efficiently crafts inputs to trigger any target class with high confidence. The attack is designed to be target- agnostic, making it distinct from previous methods. Experiments on face and speech recognition tasks confirm the attack's effectiveness, revealing a security weakness in the SoftMax layer used in transfer learning scenarios.A.Abdelkader et al. [19], introduced "headless attacks" that successfully transfer adversarial attacks against victim networks by targeting their feature extractors without needing access to the classification head. The label-blind adversarial attack does not rely on class-label information and significantly reduces the accuracy of a ResNet18 model trained on CIFAR10 by over 40%. Meenakshi et al. [20] presented a taxonomy of these threats, discusses defense techniques and countermeasures, and introduces security policies to enhance the robustness of machine learning models.

B.Pal et al. [21], demonstrated that transfer learning from state-of-the-art teacher models increases the susceptibility of student models to misclassification attacks, with high attack accuracy achieved in various tasks. This underscores the importance of designing robust training techniques for transfer-learned models.

Z. Yan et al. [22], introduced," a training strategy called “deep defense” that incorporates an adversarial perturbation-based regularization into the classification objective, enhancing the model's resistance to attacks. P. Yang, J. Chen et al. [23], introduced a novel approach for detecting adversarial examples by analyzing feature attributions. It extends the method to handle mixed confidence levels in attacks and performs exceptionally well in distinguishing adversarial examples across various real datasets, outperforming existing detection techniques. Athalye, N. Carlini et al. [24], highlighted "obfuscated gradients," a phenomenon that can deceive adversarial defense mechanisms. Defenses based on obfuscated gradients may seem effective against iterative attacks, but the study reveals they can be circumvented. Y. Ji, X. Zhang et al. [25], addressed the security risks posed by unregulated, pre-trained primitive models widely used in modern machine learning systems. It introduces model-reuse attacks, where malicious models trigger predictable misbehavior in host systems when exposed to specific inputs.

3. PROPOSED SYSTEM

Nowadays, artificial intelligence takes primary control in autonomous working because of its high learning performance. But hackers introduce different adversarial attacks like white box attacks, in which the attacker knows everything about the deployed model, including its architecture, inputs, and specific internal components like coefficient values or weights, and "black box attacks, in which the attacker has no knowledge of the structure or parameters of the target model and the only ability of the attacker is to input the chosen data to the target model and examine results labelled by the target model, to hack the AI system and produce the misclassification results. The adversarial image black box attack looks like the original

image in the human eye, which cannot be identified because both are viewed in an identical manner. This project proposed a pre-trained model for analyzing the hacking deep learning model. It delves into Adversarial Image Attacks, a technique involving the manipulation of images in a manner imperceptible to human observers but highly misleading to deep learning models. This is achieved through the application of the Fast Gradient Sign Method (FGSM), which calculates the gradients of the loss function concerning input images and perturbs them to maximize the loss, ultimately causing misclassification. The second element of the methodology is Training Using a Pretrained Model, where the well-known Alex Net architecture is harnessed as a foundational tool. This involves retaining the lower layers of the model, which have already acquired general features from a large dataset like ImageNet and fine-tuning the upper layers to adapt to a new dataset. This process significantly expedites training, as the model has already grasped fundamental features from its initial training on ImageNet. The final phase centers on building a Predictive Model, with Alex Net as its backbone. This pretrained model is designed to classify both original and adversarial images, allowing it to discern patterns and subtle perturbations introduced by adversarial attacks. This classification capability enables the model to determine the respective classes to which both original and adversarial images belong, thus serving as a robust tool for image classification and security in the face of adversarial manipulation.

4. ARCHITECTURE DIAGRAM

ALGORITHM DESCRIPTION

The convolutional layer acts as the central processing unit in a convolutional neural network, housing numerous convolutional kernels. This layer is essential for extracting features and recognizing patterns within the network. Different convolutional kernels can be used to obtain different attributes. A convolution neural network just collects features from the local perception of the previous layer, hence minimizing the number of parameters, as opposed to a typical neural network, where each neuron must be connected to every neuron in the previous layer, requiring a significant amount of calculation. Multi-core convolution may learn a wide range of features because the number of convolution kernels correlates with the number of output feature maps. Increasing the number of convolution kernels can aid in obtaining more features and, to a certain extent, improve the expressiveness of the model. The "convolution" process is carried out by the convolutional layer. A set of weights are multiplied by the input in a linear process called convolution. A set of weights makes up a kernel or filter. The input data is too large for the filter to process. A dot product is created by multiplying the filter by a portion of the input that is the size of the filter. To obtain a single value, the dot product is then combined.

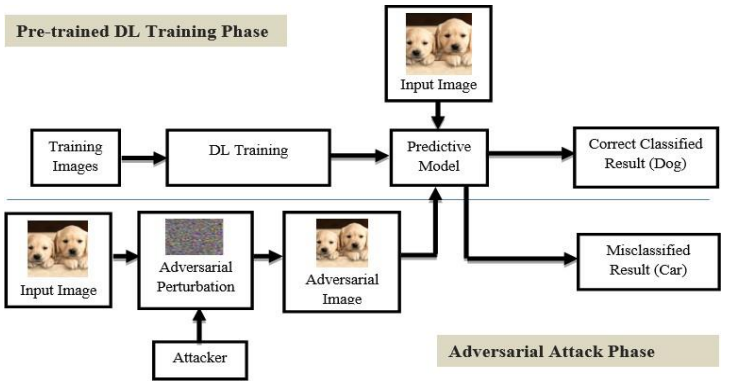

Fig. 1: Architecture diagram

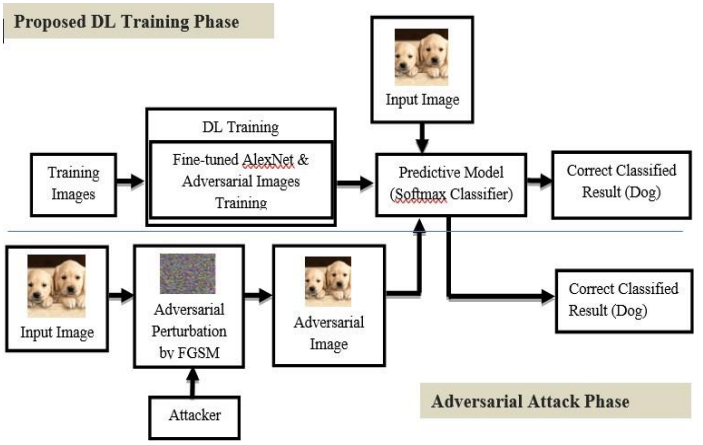

Fig. 2: Architecture diagram

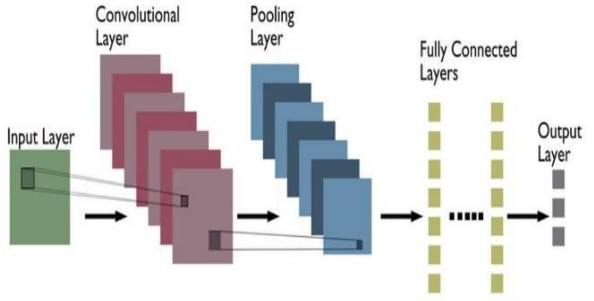

Fig.3: CNN Architecture

The pooling layer serves as a crucial element between convolutional layers, effectively reducing network complexity and computational demands. Through subsampling operations, it systematically diminishes the input size in all depth sections, a process integral for preventing overfitting during network training. Notably, this pooling method focuses on spatial size reduction while keeping the depth dimension constant. The two most popular kinds of pooling operations are max pooling and average pooling. Equations (1), (2) and (3) allow

for the breadth and height of the output in the pooling layer.

Wn, Hn and Dn in both equations stand for the input's width, height, and depth, respectively. K stands for the kernel size, whereas L stands for the stride size (the amount by which kernels are shifted on the input image).

$$W_n = \frac{W_o - K}{L} + 1 \qquad (1)$$

$$H_n = \frac{H_o - K}{L} + 1 \qquad (2)$$

$$D_n = D_o \qquad (3)$$

Fully Connected Layer

The final layer in a CNN architecture is the fully connected layer, resembling a traditional artificial neural network (ANN). Here, every neuron in the layer establishes connections with each neuron in the preceding layer. Throughout the training process, the fully connected layer computes the overall dataset score for each class, offering a comprehensive representation of class-related information.

The output of the previous layer's activation is applied to the output of a rectified linear unit (ReLU), which enhances the CNN by applying an element-wise activation function, such as sigmoid. ReLU function shown in Equation (4).

$$Rel(v) = \max(0, v) \qquad (4)$$

Due to its short calculation size and quick training time, the SoftMax classifier is commonly utilized in the output layer to solve multi-classification issues. CNNs undergo a two-step process during learning: feature extraction and classification. In the feature extraction stage, convolution is applied to input data using a filter or kernel, generating a feature map. In the subsequent classification stage, the CNN calculates the probability of the image belonging to a particular class or label. Since CNN automatically learns features rather than requiring manual feature extraction, it is especially helpful for classifying and recognizing images. Additionally, CNN can be retrained and deployed in a different domain via transfer learning. Transfer learning enhances classification performance, as has been demonstrated.

5. PROPOSED ALGORITHM

Alex Net, a CNN introduced by Alex Krizhevsky, secured the top position in the LSVRC competition by achieving a remarkable increase in accuracy compared to previous models. This includes surpassing the accuracy of the second-place model from the competition.

(1) Alex Net enhances performance through the use of the ReLU activation function for faster training convergence.

(2) It addresses overfitting with dropout and data augmentation, introducing randomness and diversifying the training set.

(3) Additionally, Alex Net leverages parallel GPU processing to significantly accelerate computational throughput during training, expediting experimentation and optimization of the neural network.

Adversarial Image Attacks

- Adversarial Image Attacks involve manipulating images in a way that's imperceptible to humans but misleads deep learning models.
- In this context, the Fast Gradient Sign Method (FGSM) is employed.
- FGSM calculates the gradients of the loss function concerning the input image, then perturbs the image in the direction that maximizes the loss.
- Essentially, it crafts subtle, purposeful alterations to input images to deceive the model. For AlexNet, these attacks might involve modifying specific pixel values within the image to cause misclassification.

Training using Pretrained model

- In this module, the focus is on exploiting transfer learning with the pretrained AlexNet model.
- AlexNet is a popular deep learning architecture known for its effectiveness in image-related tasks. Here, the pretrained AlexNet model is utilized as a foundation. The lower layers of the model, having learned general features from a large dataset like ImageNet, are retained, while the upper layers, responsible for more specific features, are further trained or modified based on the new dataset. This process accelerates training significantly, as the model has already grasped fundamental features from its initial training on ImageNet.

Predictive Model

- In this phase, a predictive model is constructed, leveraging Alex Net as its backbone. This pre-trained model is intended to classify both original and adversarial images.
- The model learns to recognize patterns in both data types, understanding the subtle perturbations introduced by adversarial attacks. By determining the respective classes to which both original and adversarial images belong.

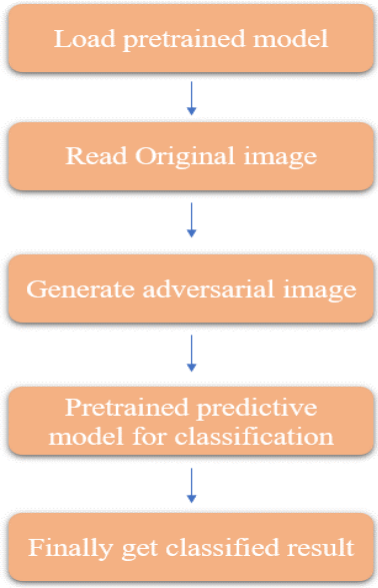

Fig. 4: Flow diagram of proposed work

Develop Defense Model

- Adversarial attacks are effective against models with unchanged pretrained base models. However, transfer learning via fine- tuning renders these attacks ineffective. We develop defense model by,

- Training with aversive images with different epsilon values to avoid black box attacks.

- Fine-tuning the structure of AlexNet pretrained architecture to avoid white box attacks. The Fine-Tuned AlexNet model was generated by modifying the fully connected layers to adapt to our specific dataset.

- With hyperparameters set to 50 epochs and SGDM optimizer, image augmentation techniques like horizontal reflection, translation, and scaling were applied to enhance model robustness.

Training using Defense Model

- The defense approach, termed 'fine-tuned AlexNet model with adversarial images training', involves mitigating vulnerabilities by training the model on adversarial examples. This process incorporates aversive images with varying epsilon values to counter black box attacks. Additionally, fine- tuning the structure of the AlexNet pretrained architecture helps deter white box attacks.

- During training, the model is exposed to both original and adversarial images.

- The training dataset includes examples with subtle adversarial perturbations, enabling the model to learn to differentiate between genuine and manipulated inputs.

- By iteratively adjusting model parameters based on training data, the model gradually improves its ability to classify images accurately, even in the presence of adversarial attacks.

Defense Predictive Model

- The defense predictive model is designed to classify both original and adversarial images accurately.
- Leveraging the insights gained from training on adversarial examples, the model employs sophisticated algorithms to identify patterns indicative of adversarial manipulations.
- Through meticulous analysis of image features, the model determines the respective classes to which original and adversarial images belong, ensuring robust and reliable classification results.

6. RESULTS & DISCUSSIONS

The experiments conducted to explore Adversarial Image Attacks on the AlexNet model revealed the model's susceptibility to imperceptible alterations. Utilizing the Fast Gradient Sign Method (FGSM), subtle changes in specific pixel values within the images led to misclassifications. Adversarial Image Attacks successfully deceived the model, highlighting the vulnerability of deep learning systems to such manipulations. In the current epoch of rapid technological advancement, artificial intelligence (AI) has emerged as a transformative force, profoundly influencing diverse facts of our existence. Nevertheless, this ascent of AI technologies has ushered in an era of heightened vulnerability, as sophisticated systems increasingly find themselves targeted by malicious actors. This study delves deep into the intricate realm of hacking AI models, meticulously scrutinizing the methods employed by nefarious entities to compromise these highly intricate systems. Through a meticulous dissection of adversarial techniques, including data poisoning, model inversion, and evasion attacks, this research sheds light on the multifaceted vulnerabilities inherent in AI systems. The exploration extends beyond mere identification, venturing into the realm of innovative defense strategies. By comprehensively understanding the intricacies of adversarial attacks, this work pioneers the formulation of robust defense mechanisms, essential for safeguarding AI systems against evolving threats. The proposed defenses not only provide immediate solutions but also lay the foundation for future advancements in AI security paradigms. This research serves as a pivotal contribution to the field of AI security, offering a profound understanding of the nuanced challenges posed by adversarial attacks. In an age where AI is ubiquitous, this study acts as a beacon, guiding researchers, developers, and policymakers towards responsible innovation. The insights gleaned from this exploration are paramount for the ethical development and deployment of AI technologies, ensuring a secure and trustworthy future amidst the evolving landscape of cyber threats.

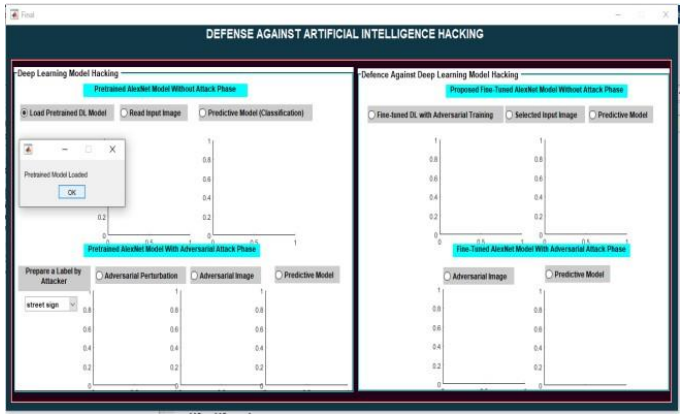

Fig. 5: Loading pretrained model and reading original image

Fig. 6: Prepare a label (Class)for Attack the original image

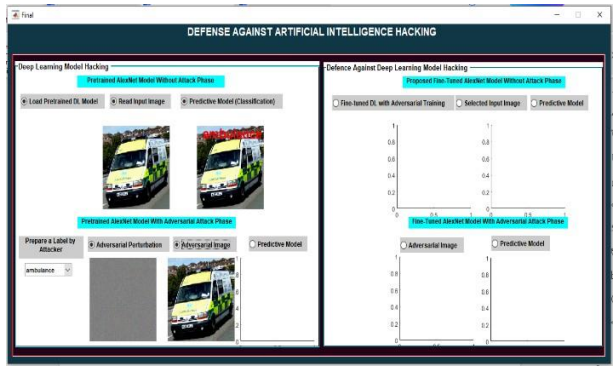

Fig. 7: Generation of Adversarial Image by Combining Generated Adversarial Perturbation and Original Image

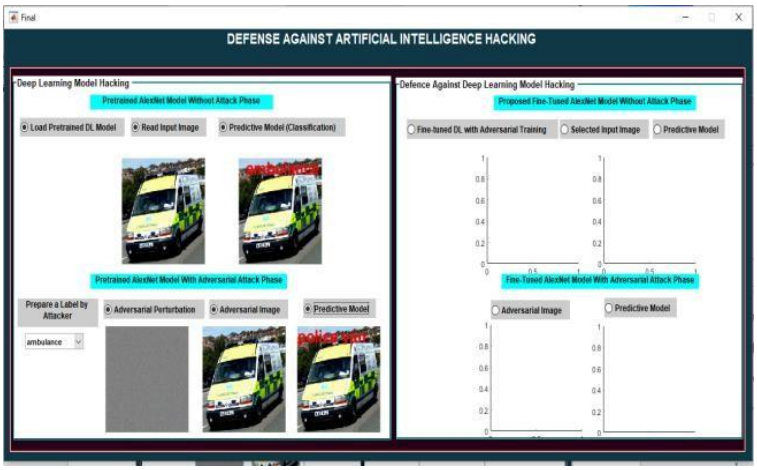

Fig. 8: Predictive Model (Classification Result for Generated Adversarial Image based on pretrained model)

7. CONCLUSION & FUTURE ENHANCEMENTS

In the ever-expanding frontier of technological innovation, the assimilation of artificial intelligence (AI) has ushered in an era of unparalleled progress, reshaping the fabric of our societal and individual experiences. However, this transformative journey is not without its intricacies, particularly as we grapple with the multifaceted challenges posed by adversarial attacks on deep learning models. Unraveling the nuances of hacking AI reveals the potency of techniques like the Fast Gradient Sign Method (FGSM) and the strategic manipulation of pretrained models, accentuating the critical need for comprehensive security frameworks and ethical guidelines in AI research and deployment. As we stand at the crossroads of innovation and ethical responsibility, our exploration into adversarial attacks serves as a pivotal moment for understanding the delicate balance required in this technological landscape. It is not merely about acknowledging vulnerabilities but also about championing proactive defense mechanisms.

Our proposed defense framework goes beyond mere acknowledgment, presenting a holistic strategy to counter deep learning model hacking. The emphasis on generating efficient models through adversarial training with fine-tuned pretrained models signifies a commitment to staying ahead of the curve in the face of emerging cyber threats.

Furthermore, this endeavor recognizes the imperative of education and awareness. As we delve into the complexities of AI security, fostering a community- wide understanding of ethical considerations becomes paramount. It is not just about developing resilient models but also about cultivating a collective responsibility to ensure the ethical deployment of AI technologies. This requires interdisciplinary collaboration, involving experts in AI, cybersecurity, and ethics, to collectively shape a future where innovation aligns seamlessly with ethical standards. In conclusion, our pursuit is not only about fortifying AI against potential breaches but also about fostering a culture of responsible innovation. Through our

endeavors, we aspire to contribute to the ongoing dialogue on the ethical dimensions of AI, shaping a future where technology enhances human experiences while upholding the values of security, transparency, and ethical integrity.

References

1. Y. Ding, F. Tan, Z. Qin, M. Cao, K. -K. R. Choo and Z. Qin, "DeepKeyGen: A Deep Learning-Based Stream Cipher Generator for Medical Image Encryption and Decryption," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 9, pp. 4915-4929, Sept. 2022, doi: 10.1109/TNNLS.2021.3062754.
2. Y. Chen, M. Zhang, J. Li and X. Kuang, "Adversarial Attacks and Defenses in Image Classification: A Practical Perspective," 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi'an, China, 2022, pp. 424-430, doi: 10.1109/ICIVC55077.2022.9886997.
3. K. Khullar, S. Kathuria, N. Chahar, P. Gupta and P. Kaur, "A Quantitative Comparison of Image Classification Models under Adversarial Attacks and Defenses," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2021, pp. 6-10, doi: 10.1109/SPIN52536.2021.9565948.
4. Pedraza A, Deniz O, Bueno G. Approaching Adversarial Example Classification with Chaos Theory. Entropy (Basel). 2020 Oct 24;22(11):1201. doi: 10.3390/e22111201. PMID: 33286969; PMCID: PMC7712112.
5. Y. Wang, L. Xie, X. Liu, J. -L. Yin and T. Zheng, "Model-Agnostic Adversarial Example Detection Through Logit Distribution Learning," 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 2021, pp. 3617-3621, doi: 10.1109/ICIP42928.2021.9506292.
6. M. A. Pandya, P. Siddalingaswamy and S. Singh, "Explainability of Image Classifiers for Targeted Adversarial Attack," 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022, pp. 1-6, doi: 10.1109/INDICON56171.2022.10039871.
7. C. Xiao and C. Zheng, "One Man's Trash Is Another Man's Treasure: Resisting Adversarial Examples by Adversarial Examples," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020 pp. 409- 418. doi: 10.1109/CVPR42600.2020.00049
8. Y. Wang, T. Li, S. Li, X. Yuan and W. Ni, "New Adversarial Image Detection Based on Sentiment Analysis," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2023.3274538.
9. Y. Liu, W. Zhang and N. Yu, "Query-Free Embedding Attack Against Deep Learning," 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 2019, pp. 380-386, doi: 10.1109/ICME.2019.00073.
10. Zhang YA, Xu H, Pei C, Yang G. Adversarial example defence based on image reconstruction. PeerJ Computer Sci. 2021 Dec 24;7:e811. doi: 10.7717/peerj- cs.811. PMID: 35036533; PMCID: PMC8725667.
11. Tian, Xuejun, Tian, Xinyuan, and Pan, Bingqin. 'Similarity Attack: An Adversarial Attack Game for Image Classification Based on Deep Learning'. 1 Jan. 2023 : 1467 – 1478
12. A.Agarwal, R. Singh, M. Vatsa and N. Ratha, "Image Transformation-Based Defense Against Adversarial Perturbation on Deep Learning Models," in IEEE Transactions on Dependable and Secure Computing, vol. 18, no. 5, pp. 2106-2121, 1 Sept.-Oct. 2021, doi: 10.1109/TDSC.2020.3027183.

13. J. Ji, B. Zhong and K. -K. Ma, "Multi-Scale Defence of Adversarial Images," 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 4070-4074, doi: 10.1109/ICIP.2019.8803408.
14. C. -Y. Lin, F. -J. Chen, H. -F. Ng and W. -Y. Lin, "Invisible Adversarial Attacks on Deep Learning- Based Face Recognition Models," in IEEE Access, vol. 11, pp. 51567-51577, 2023, doi: 10.1109/ACCESS.2023.3279488.
15. J. -H. Choi, H. Zhang, J. -H. Kim, C. -J. Hsieh and J. -S. Lee, "Deep Image Destruction: Vulnerability of Deep Image-to-Image Models against Adversarial Attacks," 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 2022, pp. 1287-1293, doi: 10.1109/ICPR56361.2022.9956577.
16. D. Vyas and V. V. Kapadia, "Evaluation of Adversarial Attacks and Detection on Transfer Learning Model," 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2023, pp. 1116-1124, doi: 10.1109/ICCMC56507.2023.10084164.
17. S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010-2020)," IEEE Transactions on Artificial Intelligence, 2021.
18. S. Rezaei and X. Liu, "A target-agnostic attack on deep models: Exploiting security vulnerabilities of transfer learning," arXiv preprint arXiv:1904.04334, 2019.
19. A.Abdelkader, M. J. Curry, L. Fowl, T. Goldstein, A. Schwarzschild, M. Shu, C. Studer, and C. Zhu, "Headless horseman: Adversarial attacks on transfer learning models," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 3087– 3091.
20. Meenakshi and G. Maragatham, "A review on security attacks and protective strategies of machine learning," in Int. Conf. on Emerging Current Trends in Computing and Expert Technology, Cham, Springer, pp. 1076–1087, 2019.
21. B.Pal and S. Tople, "To transfer or not to transfer: Misclassification attacks against transfer learned text classifiers," arXiv preprint arXiv:2001.02438, 2020.
22. Z. Yan, Y. Guo and C. Zhang, "Deep defens e: Training DNNs with improved adversarial robustness," arXiv preprint arXiv: 1803.00404, 2018.
23. P. Yang, J. Chen, C. J. Hsieh, J. L. Wang an d M. Jordan, "Ml-loo: Detecting adversarial examples w ith feature attribution," Proceedings of the AAAI Confer ence on Artificial Intelligence, vol. 34, no. 4, pp. 6639– 6647, 2020.
24. Athalye, N. Carlini and D. Wagner, "Obfusc ated gradients give a false sense of security: Circumven ting defenses to adversarial examples," in Int. Conf. on Machine Learning, PMLR, Stockholm, Sweden, pp. 27 4–283, 2018.
25. Y. Ji, X. Zhang, S. Ji, X. Luo, and T. Wang, "Model-reuse attacks on deep learning systems," in Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018, pp. 349– 363.