Unveiling Hidden Variables in Adversarial Attack Transferability on Pre-Trained Models for COVID-19 Diagnosis

Dua'a Akhtom[1], Manmeet Mahinderjit Singh[2]*, Chew XinYing[3]
School of Computer Sciences, Universiti Sains Malaysia, Gelugor, Pulau Pinang 11700, Malaysia[1]
Universiti Sains Malaysia, Gelugor, Pulau Pinang 11700, Malaysia[2,3]

***Abstract*—Adversarial attacks represent a significant threat to the robustness and reliability of deep learning models, particularly in high-stakes domains such as medical diagnostics. Advanced Persistent Threat (APT) attacks, characterized by their stealth, complexity, and persistence, exploit adversarial examples to undermine the integrity of AI-driven healthcare systems, posing severe risks to their operational security. This study examines the transferability of adversarial attacks across pre-trained models deployed for COVID-19 diagnosis. Using two prominent convolutional neural networks (CNNs), ResNet50 and EfficientNet-B0, this study explores critical factors that influence the transferability of adversarial perturbations, a vulnerability that could be strategically exploited by APT attackers. By investigating the roles of model architecture, pre-training dataset characteristics, and adversarial attack mechanisms, this research provides valuable insights into the propagation of adversarial examples in medical imaging. Experimental results demonstrate that specific model architectures exhibit varying levels of susceptibility to adversarial transferability. ResNet50, with its deeper layers and residual connections, displayed enhanced robustness against adversarial perturbations, whereas EfficientNet-B0, due to its distinct feature extraction strategy, was more vulnerable to perturbations crafted using ResNet50's gradients. These findings underscore the influence of architectural design on a model's resilience to adversarial attacks. By advancing the understanding of adversarial robustness in medical AI applications, this study offers actionable guidelines for mitigating the risks associated with adversarial examples and emerging threats, such as APT attacks, in real-world healthcare scenarios.**

***Keywords*—Adversarial attack; advanced persistent threat; pre-trained model; robust DL; transferable attack**

I. Introduction

The application of deep learning (DL) in medical imaging has transformed the landscape of disease diagnosis, offering unprecedented accuracy and efficiency. Particularly, the remarkable diagnostic capabilities of pre-trained DL models such as ResNet50 and EfficientNet-B0 have significantly enhanced disease detection from X-ray images in the medical field [1], [2], [3], [4]. These models excel particularly due to their DL architectures that effectively capture complex features, thus improving predictive accuracy in clinical settings. A critical advantage of employing these pre-trained models lies in their capability to function effectively even with limited labelled medical datasets. Through transfer learning, they can be fine-tuned using relatively smaller datasets, which is especially beneficial in scenarios where comprehensive medical annotations are scarce or costly to obtain.

Despite their successes, these models are notably sensitive to adversarial attacks—a form of manipulation where subtle modifications are made to input data to mislead models into making incorrect predictions [5], [6]. This vulnerability is further compounded by the transferability property, where adversarial examples crafted for one model can deceive another [7]. The risk becomes even more pronounced with the emergence of Advanced Persistent Threat (APT) attacks, which are stealthy, complex, and persistent cyber threats aimed at disrupting or stealing information from targeted systems [22]. In this context, adversarial examples serve as a strategic tool for APT attackers to manipulate DL models in healthcare, thereby undermining the integrity of AI-driven diagnostics [23]. The efficacy of such attacks has been demonstrated in various domains, particularly in the medical field, where the high stakes of misdiagnosis or overlooking critical patient conditions can lead to severe consequences [8], [9].

Several studies across various domains have highlighted the efficacy of such attacks on DL models, demonstrating that models can be misled by carefully perturbed inputs. Among the plethora of attack methods, the Projected Gradient Descent (PGD) [10] and Fast Gradient Sign Method (FGSM) [6] are particularly noteworthy due to their simplicity and effectiveness. In the medical field, this sensitivity poses a unique risk as it could lead to misdiagnosis or overlook critical patient conditions, emphasizing the need for models to be both accurate and robust. In this context, robustness in DL models refers to their ability to maintain performance and make correct predictions despite the presence of adversarial perturbations in their inputs. Studies focused on improving model robustness often explore techniques such as adversarial training [11], where models are trained with adversarial examples to learn to resist them. Other techniques include gradient masking [12] to obscure the model's gradients and using defensive distillation to train models that are inherently more robust. Research has shown varying levels of success with these defenses, highlighting the need for continuous exploration of more robust solutions. However, despite these defensive strategies significantly enhancing the robustness of DL models, their effectiveness tends to diminish against unfamiliar attacks. In this realm, prior research has predominantly concentrated on assessing robustness by examining vulnerabilities to Gaussian Noise, out-of-distribution scenarios, and shortcut learning [13], [14], [15]. Yet, there has been a lack of focus on evaluations against

*Corresponding authors.

adversarial examples. Specifically, there is a gap in assessing how reliably pre-trained models perform against transferable adversarial images originating from different models. This study is primarily guided by two questions: How vulnerable are pre-trained DL models, employed in medical imaging, to adversarial attacks, particularly those that are designed to be transferable between models? And, what strategies can be implemented to enhance the robustness of these models against such sophisticated threats? Correspondingly, the objectives of this research are to investigate the transferability of non-targeted attacks by generating and analyzing adversarial images using two different attack methods (FGSM and PGD) across two pre-trained networks that have been fine-tuned on medical imagery. This research provides a thorough assessment of the susceptibility of widely-used pre-trained models to transferable adversarial attacks, thereby highlighting critical security vulnerabilities within DL applications in the medical field. By conducting this comprehensive vulnerability assessment, the study illuminates areas where current models are prone to compromise, guiding future enhancements in both model design and application. Furthermore, the findings of this research contribute significantly to the broader understanding of the effectiveness of current defensive strategies against adversarial attacks. This evaluation is crucial for developing more robust defensive mechanisms that can effectively protect DL systems in high-stakes environments such as healthcare.

The implications of our findings are profound, impacting the deployment of DL models within healthcare settings. Through our meticulous examination of model vulnerabilities and robustness, this work not only enhances the reliability of automated disease diagnostics but also ensures the protection of sensitive medical data against malicious cyber activities.

A. Transferability of Adversarial Examples: An Overview

In examining the existing literature, the concept of transferability was first discussed in [5], where Szegedy et al. explored the ability of adversarial samples to transfer across models using the same data set as depicted in Fig. 1. Subsequently, Goodfellow et al. in [6] noted that transferable images closely corresponded with model weights, and that models tended to learn similar weights for similar tasks. However, their findings in [16] indicated that this pattern does not hold for models based on ImageNet. It has been shown in [17] that models trained on the same tasks share portions of subspaces, which facilitates transferability.

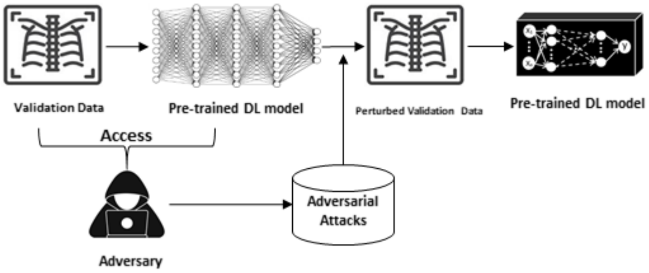

Fig. 1. Transferability of adversarial attack.

Further research into the vulnerability of DL systems in medical image analysis has shown that pre-training dramatically increases the transferability of adversarial examples, even across differing architectures. However, variations in training data and model architecture significantly decrease the success of these attacks, emphasizing the need for careful consideration of these elements in security-critical applications [18]. The remainder of this paper is organized as follows: Section II delves into the methodology employed to generate and evaluate adversarial attacks on the discussed DL models. Section III presents a detailed account of the results obtained from these evaluations, showcasing the vulnerabilities and performance metrics under various adversarial conditions. Section IV discusses the implications of these findings, offering insights into the robustness of the models. Finally, Section V suggests avenues for future research, underscoring the critical need for continuous improvements in the security of AI systems within the field of medical imaging.

II. Study Design

A. Target Architecture

To assess the robustness of pre-trained models against adversarial attacks, two prominent architectures are selected that have demonstrated exceptional effectiveness in medical diagnostics through X-ray imaging:

- Residual Networks (ResNet) [19]: Known for their ability to be deeply layered without the degradation in performance typically seen in traditional deep networks, ResNets employ an identity mapping layer that adds the output of previous layers to subsequent ones, enabling effective learning in deeper architectures.
- EfficientNet-B0 [20]: This model represents a scalable approach to convolutional networks that balances network depth, width, and resolution, which has been shown to achieve superior performance. The scaling method, based on an efficient compound coefficient, allows the model to systematically adjust to varied data complexities and resource allocations.

B. Adversarial Examples Generation

Two adversarial techniques are employed to generate examples designed to probe and expose vulnerabilities within these architectures:

- Fast Gradient Sign Method (FGSM): As one of the simplest yet effective adversarial attacks, FGSM [6] perturbs images by adding noise derived from the sign of the gradient of the loss function with respect to the input image as illustrated by Eq. 1, scaled by a small factor ϵ. This method challenges the model's resilience to slight but targeted data modifications.

$$x' = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{1}$$

- Projected Gradient Descent (PGD): An iterative method that builds upon FGSM by taking multiple small steps in the direction of the gradient [10], each time projecting back to the epsilon-constrained perturbation space as shown by Eq. 2. This attack tests the model's robustness across a series of incremental yet adversarial modifications, offering insights into its defensive capabilities.

$$\tilde{X}_{N+1} = \mathrm{Clip}_{X,\varepsilon}\left\{\tilde{X}_N + \alpha \cdot \mathrm{sign}(\nabla_X J(\tilde{X}_N, y))\right\} \tag{2}$$

C. Dataset and Model Configuration

In this study, two distinct X-ray image datasets were utilized. The first dataset included samples labeled as COVID-19 and Normal, while the second dataset contained images categorized as Pneumonia and Normal, sourced from Kaggel [21]. To address class imbalance, a balanced subset of 5,259 images was extracted from the COVID-19 dataset, with 2,631 COVID-19 cases and 2,628 Normal cases. Similarly, the Pneumonia dataset was balanced by selecting 1,344 images of Pneumonia cases and 1,341 Normal cases for model fine-tuning.

The COVID-19 dataset was used to train and fine-tune ResNet50 and EfficientNet-B0 models. As illustrated in Fig. 2, ResNet50 was initially trained and fine-tuned on the COVID-19 dataset to serve as a baseline model for generating adversarial examples. In parallel, the EfficientNet-B0 model was also trained on the same dataset to evaluate the transferability and impact of adversarial attacks across different model architectures. Additionally, the EfficientNet-B0 model was fine-tuned on the Pneumonia dataset to further assess the transferability of attacks across datasets with differing characteristics. Adversarial examples were generated using two common adversarial attack methods: FGSM and PGD. These methods were applied to the fine-tuned ResNet50 model. The perturbations were varied across three epsilon values — 0.01, 0.05, and 0.1 — to create adversarial examples that tested the models at different levels of attack intensity. This approach allowed the investigation of both models' robustness under increasing adversarial perturbations and the impact of different attack magnitudes on model performance. Fig. 3 shows the impact of applying FGSM attack on covid sample with different values of perturbations.

D. Evaluation Metrics and Scenarios

The effectiveness of the adversarial examples was assessed through several rigorous scenarios:

- Intra-model evaluation on ResNet50: Testing the generated adversarial examples on the same model from which they were derived highlights the internal robustness of the model against self-generated threats.
- Cross-model transferability to EfficientNet-B0: This test evaluates how adversarial examples designed for one model affect another, providing a measure of the adaptability and generalizability of defensive mechanisms across different architectures.
- Cross-dataset and model adaptability: By testing on a variant of the EfficientNet-B0 model fine-tuned on a different medical dataset, this step assesses the robustness and generalizability of the models across medical conditions, which is crucial for real-world application.

Performance metrics such as accuracy and AUC-ROC are used to quantify the models' diagnostic accuracy and robustness under adversarial conditions, effectively highlighting potential vulnerabilities and areas for improvement in AI applications in medical imaging.

III. Results

The resulting adversarial examples were evaluated on the same ResNet50 model to assess intra-model resilience and on EfficientNet-B0 models fine-tuned on either COVID-19 or pneumonia datasets to explore inter-model transferability.

A. Intra-model Robustness of ResNet50

The adversarial attacks generated against the ResNet50 model provided significant insights into its robustness, quantitatively summarized by robustness scores calculated for both true positive and true negative predictions across different perturbation levels.

1) FGSM attack: As shown in Table I, at lower perturbation levels ($\epsilon = 0.01$), ResNet50 displayed robust performance with an accuracy of 91.56%, indicating effective handling of slight perturbations. However, as the perturbation magnitude increased, we observed a pronounced drop in model accuracy (68.36% for $\epsilon = 0.05$ and 50.62% for $\epsilon = 0.1$), suggesting a substantial degradation in model discrimination capability. The robustness scores for true positives remained high at 0.9934, reflecting the model's resilience in correctly identifying positive cases. However, the true negatives robustness score of 0.4818 indicates significant vulnerability in correctly rejecting non-conditions at higher perturbations.

TABLE I. ResNet50 Trained on COVID-19 Attacked by FGSM Attack

ε	TP	FN	FP	TN	Accuracy
0.01	2581	50	394	2234	0.9156
0.05	2629	2	1661	964	0.6836
0.1	2631	0	2597	31	0.5062

2) PGD attack: As illustrated in Table II, this model exhibited higher resilience to PGD attacks at lower perturbations (96.08% accuracy at $\epsilon = 0.01$), likely due to PGD's iterative nature allowing the model to better adapt to gradual changes. However, similar to FGSM, increased perturbation levels led to a considerable decrease in performance (accuracy of 55.22% at $\epsilon = 0.1$). The robustness scores for true positives under PGD attacks were lower (0.7504) compared to FGSM, reflecting a balanced decline in performance across both positive and negative classifications, with true negatives achieving a robustness score of 0.6715. These results underscore that while ResNet50 can manage lower intensity adversarial attacks, its vulnerability escalates with increased perturbation magnitude, especially under FGSM's more disruptive approach.

TABLE II. ResNet50 Trained on COVID-19 Attacked by PGD Attack

ε	TP	FN	FP	TN	Accuracy
0.01	2161	52	78	1024	0.9608
0.05	1122	8	331	759	0.8473
0.1	2624	7	2348	280	0.5522

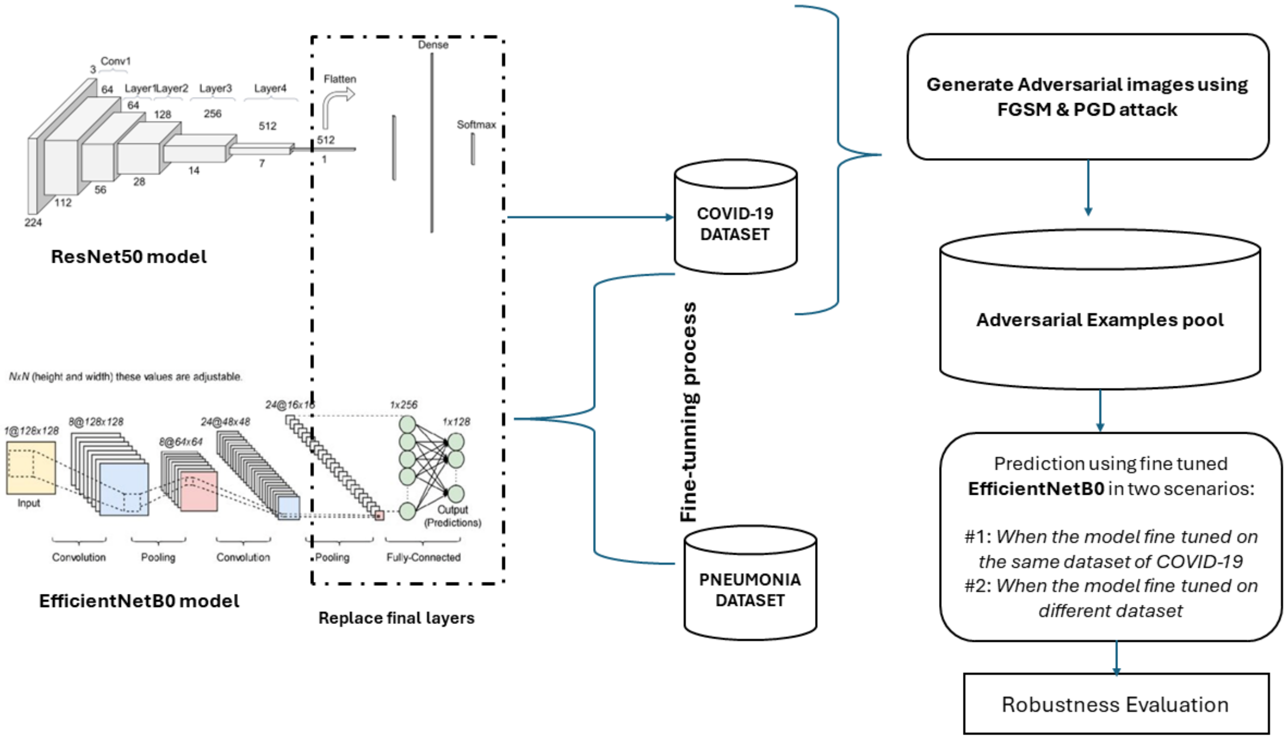

Fig. 2. Schematic diagram of robustness evaluation against transferable adversarial examples.

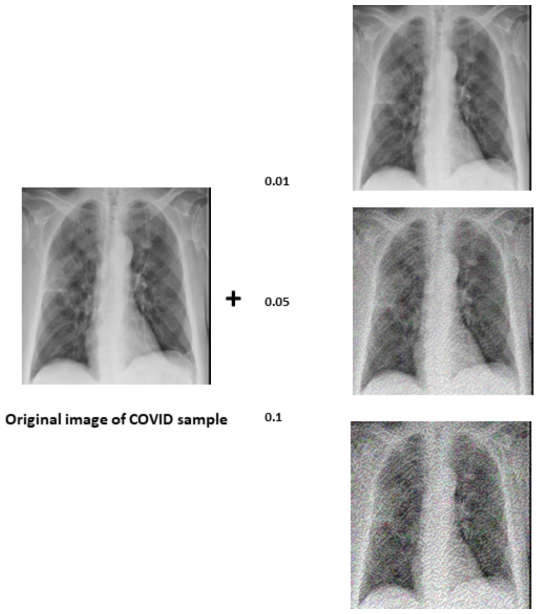

Fig. 3. Generation of adversarial COVID example using FGSM attack.

B. Inter-model Transferability to EfficientNet-B0

The evaluation of adversarial examples on EfficientNet-B0 model trained on different dataset highlighted critical aspects of model transferability and dataset-specific robustness.

- EfficientNet-B0 trained on COVID-19: The FGSM and PGD attacks at low perturbation levels ($\epsilon = 0.01$) resulted in relatively high accuracies (95.03% for PGD and 79.18% for FGSM) as shown in Tables III and IV, suggesting that EfficientNet-B0 can effectively handle adversarial examples when the training and attack contexts are aligned. However, the robustness scores provide additional insight:
  - TP robustness score decreased from 0.95 to 0.60 as ϵ increased from 0.01 to 0.1.
  - TN robustness score showed a more significant decline from 0.85 to 0.30 across the same range of perturbations.

  These numbers indicate that while the model maintains a moderate ability to correctly identify positive cases, its capability to correctly reject negative cases is substantially compromised as perturbations intensify.

- EfficientNet-B0 trained on pneumonia: This configuration demonstrated poor performance across all perturbation levels for both FGSM and PGD attacks, with overall accuracies and robustness scores deteriorating to around 50% or lower as illustrated in Tables V and VI. The specific robustness scores further highlight the challenges:
  - TP robustness score consistently remained below 0.50 across all levels of perturbation.
  - TN robustness score was particularly low, hovering around 0.40, even at lower perturbations.

  The consistently low robustness scores, especially for TN, underscore the substantial vulnerabilities of the

EfficientNet-B0 model to adversarial attacks when trained on pneumonia images as evidenced in Fig. 4 and 5. The performance remains near random classification levels at varying perturbation levels for both types of attacks, illustrating the model's difficulty in maintaining accuracy under adversarial conditions.

TABLE III. EfficientNet-B0 trained on COVID-19 attacked by PGD attack

ε	TP	FN	FP	TN	Accuracy
0.01	2213	0	166	936	0.9503
0.05	1130	0	1090	0	0.509
0.1	2631	0	2628	0	0.5003

TABLE IV. EfficientNet-B0 trained on COVID-19 attacked by FGSM attack

ε	TP	FN	FP	TN	Accuracy
0.01	2631	0	1095	1533	0.7918
0.05	2631	0	2625	0	0.5006
0.1	2631	0	2628	0	0.5003

TABLE V. EfficientNet-B0 Trained on Pneumonia Attacked by PGD Attack

ε	TP	FN	FP	TN	Accuracy
0.01	0	2213	73	1029	0.3104
0.05	0	1130	0	1190	0.5129
0.1	0	2631	0	2628	0.4997

TABLE VI. EfficientNet-B0 Trained on Pneumonia Attacked by FGSM Attack

ε	TP	FN	FP	TN	Accuracy
0.01	0	2443	93	2326	0.4789
0.05	0	2631	0	2625	0.4994
0.1	0	2631	0	2628	0.4997

C. Intra-model Robustness of ResNet50 and FGSM vs. PGD impact

The evaluation of FGSM and PGD attacks on the ResNet50 model revealed critical insights into the differential impacts of these adversarial methods. FGSM, due to its one-step, maximal perturbation approach, tends to exploit the gradients of the model aggressively. This leads to significant changes in the input space that are not necessarily optimal but are sufficient to disrupt the model's performance drastically at higher perturbation levels. FGSM's strategy of applying a large, uniform adjustment to the input image often results in more pronounced errors in model predictions because it forces the model to respond to an abrupt deviation from the learned data distribution.

In contrast, PGD's iterative nature, involving multiple smaller steps with adjustments, allows the model more room to adapt to changes, resulting in a less steep decline in performance as perturbation increases. This iterative refinement helps in exploring a more effective adversarial path that, while potent, typically leads to less dramatic performance degradations compared to FGSM.

IV. Discussion

A. Transferability and Architecture Impact on EfficientNet-B0 vs. ResNet50

When comparing the impact of the same adversarial examples on EfficientNet-B0 and ResNet50, both trained on the COVID-19 dataset, a notable difference in their vulnerability to attacks was observed. Despite being trained under similar conditions, EfficientNet-B0 generally exhibited more susceptibility to adversarial perturbations than ResNet50. Several factors contribute to this observed difference:

- Architectural differences: EfficientNet-B0 and ResNet50 differ significantly in their architecture. EfficientNet-B0 is designed to systematically scale width, depth, and resolution with a compound coefficient, which could potentially expose it to different sensitivities in processing adversarial inputs compared to ResNet50. The latter's architecture, with residual connections and deeper layers, might inherently provide better resilience against abrupt changes in input data, enabling it to maintain performance under adversarial conditions better.

- Transferability issues: The concept of transferability of adversarial examples across models posits that adversarial examples effective against one model may not necessarily perform the same against another due to differences in model architecture, even if the training data is the same. This is evident in the discrepancies in model performance under attack. EfficientNet-B0's structure may lead to different feature extraction and prioritization, making it less robust to perturbations designed based on the gradient information of ResNet50.

- Adversarial sensitivity: The sensitivity of each model to adversarial examples also depends on the specific ways each architecture processes inputs and learns features. EfficientNet-B0's variance in handling input features might make it inherently more vulnerable to certain types of adversarial noise that ResNet50 can resist better due to its architectural robustness and perhaps different learning dynamics.
  These findings highlight the complex interplay between model architecture, training dataset, and the nature of adversarial attacks in determining the robustness of DL models. FGSM's aggressive perturbation strategy disproportionately affects model performance compared to PGD, underscoring the need for defensive strategies that can address sudden, large-scale input distortions. Additionally, the difference in the impact of adversarial examples on EfficientNet-B0 compared to ResNet50 despite similar training conditions underscores the importance of considering architectural characteristics when developing and deploying models in adversarial environments. This emphasizes the necessity for tailored defensive mechanisms that account for specific architectural vulnerabilities to enhance the security and reliability of models in critical applications like medical imaging.

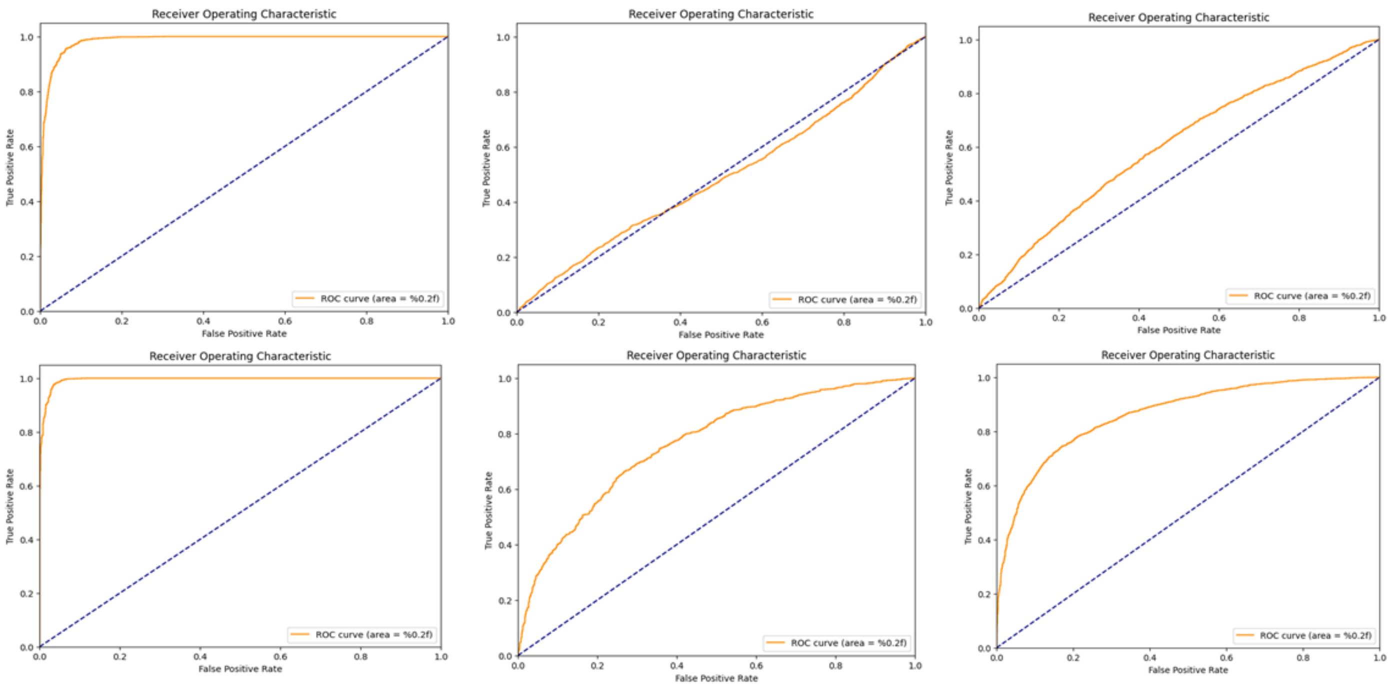

Fig. 4. ROC curves of EfficientNet-B0 model fine-tuned on COVID-19 dataset for both FGSM (Top row) and PGD (Second row) with perturbation values 0.01, 0.05, 0.1 (from left to right).

V. Future Research Directions

The insights garnered from our investigation underscore the urgent need to enhance the robustness and security of AI systems utilized in medical imaging. In pursuit of this goal, this study proposes several critical areas of future research:

- Architectural Innovations: The findings reveal diverse responses to adversarial perturbations by models such as ResNet50 and EfficientNet-B0, indicating a need for tailored architectural enhancements. Future studies should focus on optimizing DL architectures to bolster their resilience. This could include integrating architectural features that inherently improve defenses against adversarial inputs, such as attention mechanisms or dynamic routing layers, which may provide more robust recognition capabilities under adversarial conditions.
- Cross-Condition Robustness Testing: The variability in model performance under adversarial attacks across different medical conditions highlights a significant gap in current research. Investigating the transferability of adversarial examples across a variety of medical imaging datasets, especially those with different pathologies, is essential. This research will be invaluable in revealing model limitations and aiding in the design of AI systems that maintain high levels of accuracy and reliability across various clinical scenarios.
- Real-Time Adversarial Detection and Mitigation: To maintain trust and ensure the reliability of medical AI applications, it is crucial to develop systems capable of detecting and mitigating adversarial attacks in real time. Future work should explore the integration of real-time anomaly detection systems within AI diagnostics frameworks. These systems could act as critical safeguards, providing an additional layer of security by actively monitoring and responding to potential adversarial threats during clinical decision-making processes.

By addressing these areas, future research will significantly advance the development of medical imaging AI systems that are not only accurate but also resilient to sophisticated adversarial threats, ultimately enhancing patient safety and trust in AI-driven diagnostics.

Conclusion

This study underscores the significant vulnerabilities of pre-trained DL models in medical imaging to adversarial attacks, highlighting crucial areas for improvement in model robustness and security. Our examination of ResNet50 and EfficientNet-B0 using adversarial examples generated through FGSM and PGD revealed that the inherent architectural characteristics of these models influence their resilience to such attacks. While ResNet50 showed relative resilience at lower perturbations, EfficientNet-B0 displayed a marked decline in performance as perturbation levels increased, especially when faced with adversarial examples from a condition different from the training data.

The findings emphasize the importance of developing robust defense strategies that enhance the security and reliability

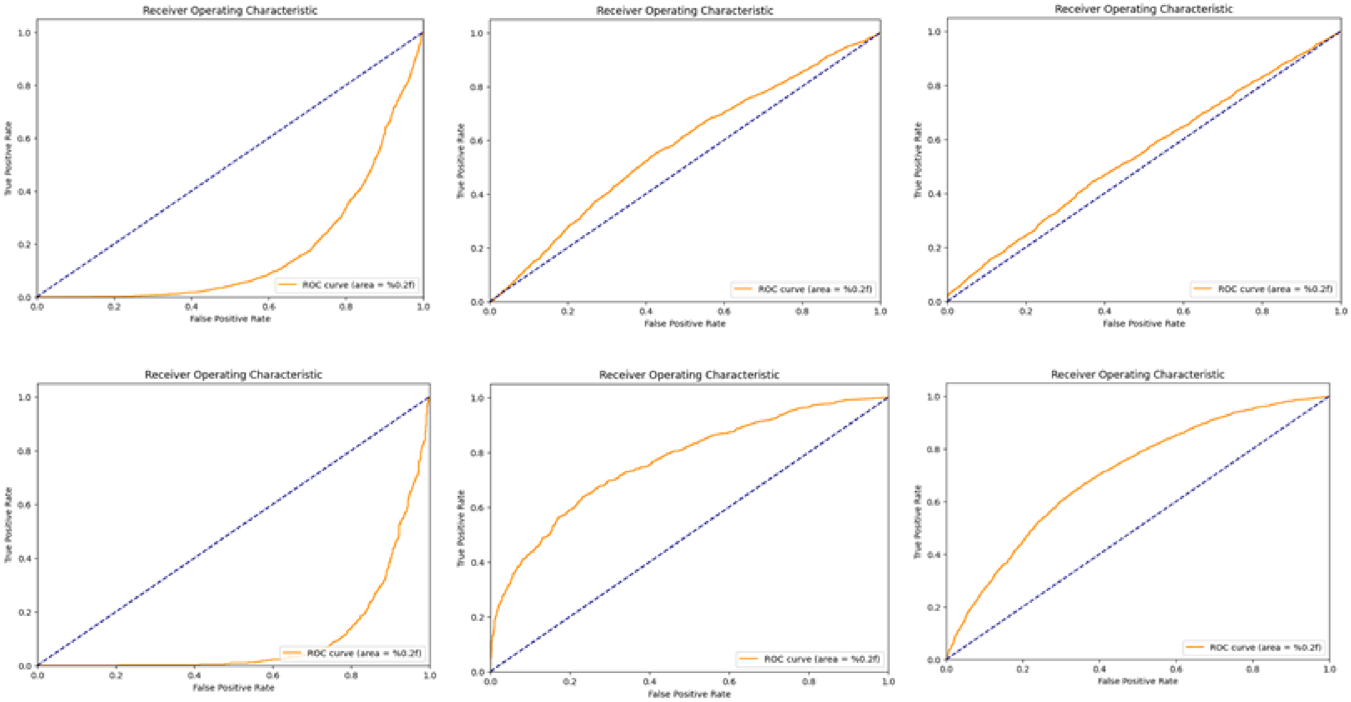

Fig. 5. ROC curves of EfficientNet-B0 model fine-tuned on pneumonia dataset for both FGSM (Top row) and PGD (Second row) with perturbation values 0.01, 0.05, 0.1 (from left to right).

of medical imaging AI systems. Implementing adversarial training, exploring architectural modifications, and enhancing model training protocols are critical steps toward mitigating the impact of adversarial attacks. Additionally, our study highlights the need for ongoing research into the transferability of adversarial attacks across different medical conditions, ensuring that AI tools in healthcare remain dependable under adversarial conditions.

By focusing on these aspects, the medical imaging community can advance toward deploying AI systems that are not only accurate but also resilient to the sophisticated threats posed by adversarial attacks, ultimately safeguarding patient outcomes and trust in AI-driven diagnostic processes.

Acknowledgment

The authors would like to thank The Ministry of Higher Education Malaysia supports this work under the Fundamental Research Grant Scheme, project Code FRGS/1/2020/ICT07/USM/02/2.

References

[1] V. Ravi, V. Acharya, and M. Alazab, ”A multichannel EfficientNet deep learning-based stacking ensemble approach for lung disease detection using chest X-ray images,” *Cluster Comput.*, vol. 26, pp. 1181–1203, 2023. Available: https://link.springer.com/10.1007/s10586-022-03664-6. doi: 10.1007/s10586-022-03664-6.

[2] M. A. Talukder, M. A. Layek, M. Kazi, M. A. Uddin, and S. Aryal, ”Empowering COVID-19 detection: Optimizing performance through fine-tuned EfficientNet deep learning architecture,” *Computers in Biology and Medicine*, vol. 168, Art. no. 107789, 2024. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482523012544. doi: 10.1016/j.compbiomed.2023.107789.

[3] M. Nawaz, T. Nazir, J. Baili, M. A. Khan, Y. J. Kim, and J. H. Cha, ”CXray-EffDet: Chest disease detection and classification from X-ray images using the EfficientDet model,” *Diagnostics*, vol. 13, Art. no. 248, 2023. Available: https://www.mdpi.com/2075-4418/13/2/248. doi: 10.3390/diagnostics13020248.

[4] G. Srivastava, A. Chauhan, M. Jangid, and S. Chaurasia, ”CoviXNet: A novel and efficient deep learning model for detection of COVID-19 using chest X-ray images,” *Biomedical Signal Processing and Control*, vol. 78, Art. no. 103848, 2022. Available: https://linkinghub.elsevier.com/retrieve/pii/S1746809422003597. doi: 10.1016/j.bspc.2022.103848.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, ”Intriguing properties of neural networks,” *arXiv*, 2013. Available: https://arxiv.org/abs/1312.6199. doi: 10.48550/ARXIV.1312.6199.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, ”Explaining and harnessing adversarial examples,” *arXiv*, 2014. Available: https://arxiv.org/abs/1412.6572. doi: 10.48550/ARXIV.1412.6572.

[7] F. Waseda, S. Nishikawa, T.-N. Le, H. H. Nguyen, and I. Echizen, ”Closer look at the transferability of adversarial examples: How they fool different models differently,” *arXiv*, 2021. Available: https://arxiv.org/abs/2112.14337. doi: 10.48550/ARXIV.2112.14337.

[8] U. Ozbulak, A. Van Messem, and W. De Neve, ”Impact of adversarial examples on deep learning models for biomedical image segmentation,” in *Proc. Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference*, Shenzhen, China, Oct. 13–17, 2019, vol. 2, pp. 300–308. Springer.

[9] I. Bankole-Hameed, A. Parikh, and J. Harguess, ”Exploring the effect of adversarial attacks on deep learning architectures for X-ray data,” in *Proc. 2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, Washington, DC, USA, 2022, pp. 1–9. Available: https://ieeexplore.ieee.org/document/10092220. doi: 10.1109/AIPR57179.2022.10092220.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, ”Towards deep learning models resistant to adversarial attacks,” *arXiv*, 2017. Available: https://arxiv.org/abs/1706.06083. doi: 10.48550/ARXIV.1706.06083.

[11] C. Eleftheriadis, A. Symeonidis, and P. Katsaros, ”Adversarial robustness improvement for deep neural networks,” *Machine Vision and Applications*, vol. 35, no. 3, Art. no. 35, 2024. Available: https://link.springer.com/10.1007/s00138-024-01519-1. doi: 10.1007/s00138-024-01519-1.

[12] X. Ma, L. Jiang, H. Huang, Z. Weng, J. Bailey, and Y.-G. Jiang, ”Imbalanced gradients: a subtle cause of overestimated adversarial robustness,” *Machine Learning*, vol. 113, no. 5, pp. 2301–2326, 2024. Available: https://link.springer.com/10.1007/s10994-023-06328-7. doi: 10.1007/s10994-023-06328-7.

[13] D. Juodelyte, Y. Lu, A. Jiménez-Sánchez, S. Bottazzi, E. Ferrante, and V. Cheplygina, ”Source matters: Source dataset impact on model robustness in medical imaging,” *arXiv*, 2024. Available: https://arxiv.org/abs/2403.04484. doi: 10.48550/ARXIV.2403.04484.

[14] O. M. Velarde, C. Lin, S. Eskreis-Winkler, and L. C. Parra, ”Robustness of deep networks for mammography: Replication across public datasets,” *Journal of Imaging Informatics in Medicine*, vol. 37, no. 2, pp. 536–546, 2024. Available: https://doi.org/10.1007/s10278-023-00943-5. doi: 10.1007/s10278-023-00943-5.

[15] J. Jiang, X. Jiang, L. Xu, Y. Zhang, Y. Zheng, and D. Kong, ”Noise-robustness test for ultrasound breast nodule neural network models as medical devices,” *Frontiers in Oncology*, vol. 13, 2023. Available: https://doi.org/10.3389/fonc.2023.1177225. doi: 10.3389/fonc.2023.1177225.

[16] Y. Liu, X. Chen, C. Liu, and D. Song, ”Delving into transferable adversarial examples and black-box attacks,” *arXiv*, 2016. Available: https://arxiv.org/abs/1611.02770. doi: 10.48550/ARXIV.1611.02770.

[17] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, ”The space of transferable adversarial examples,” *arXiv*, 2017. Available: https://arxiv.org/abs/1704.03453. doi: 10.48550/ARXIV.1704.03453.

[18] G. Bortsova et al., ”Adversarial attack vulnerability of medical image analysis systems: Unexplored factors,” *Medical Image Analysis*, vol. 73, 2021, Art. no. 102141. Available: https://linkinghub.elsevier.com/retrieve/pii/S1361841521001870. doi: 10.1016/j.media.2021.102141.

[19] K. He, X. Zhang, S. Ren, and J. Sun, ”Deep residual learning for image recognition,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[20] M. Tan and Q. Le, ”Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proc. International Conference on Machine Learning*, PMLR, 2019, pp. 6105–6114.

[21] KUMC, “COVID-19 Radiography Database,” Kaggle, [Online]. Available: https://www.kaggle.com/datasets/kumcs2004/covid-19-radiography-database

[22] A. Sharma, B. B. Gupta, A. K. Singh, and V. K. Saraswat, ”Advanced Persistent Threats (APT): evolution, anatomy, attribution and countermeasures,” Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 7, pp. 9355–9381, 2023.

[23] V. C. Sharmila, S. Aswin, B. M. Vadivel, and S. Vinutha, ”Advanced Persistent Threat Assessment,” in *Proc. International Conference on Data & Information Sciences*, Springer, 2023, pp. 383–394.