Securing AI-Driven Medical Diagnostics: Detecting Adversarial Examples in Healthcare Imaging

Saurav Raj Pandey
Department of Computer Science, UNC Chapel Hill

Scott Merill
Department of Computer Science, UNC Chapel Hill

Abstract

Machine Learning (ML) models have revolutionized healthcare diagnostics but are vulnerable to adversarial attacks, where subtle input modifications can lead to inaccurate diagnosis. In this work, we evaluate the vulnerability of AI-driven medical diagnostics to gradient-based attacks – Universal Adversarial Perturbation (UAP) and Projected Gradient Descent (PGD). We find a roughly 36% and 60% drops in accuracy in popular deep learning models like VGG16 and EfficientNetB0 respectively due to the adversarial perturbations. To mitigate these threats, we build a simple yet accurate detection framework that is 95% accurate for identifying UAP perturbations and 72% for PGD perturbations. Our findings support the urgent need for robust defenses to secure AI systems in high-stakes areas such as healthcare.

1 Introduction

Machine Learning (ML) models have become an indispensable part of modern healthcare, significantly enhancing the speed, accuracy, and efficiency of medical diagnoses. In radiology, these models can easily identify statistical patterns in X-Rays, MRIs and CT scans to diagnose diseases like COVID-19, pneumonia, and various cancers. Healthcare providers also use AI models with tabular data. By combining biochemical data such as Fasting Blood Sugar (FBS) with lifestyle data such as physical activity, for example, ML models can be used to predict and diagnose diabetes. However, while these models can quickly process vast amounts of data and identify patterns not immediately apparent to human clinicians, they also introduce new risks. Medical image diagnostic systems are easily fooled by adversarial examples. Adversarial examples exploit the intricate decision boundaries that ML models learn during training. By applying small, carefully crafted perturbations to an image, adversaries can manipulate the model's output without significantly altering the image.

The growing accessibility of medical data further amplifies these risks. In many healthcare systems, patients now have direct access to their own medical data, including diagnostic images, via patient portals, second opinion services, or by request. While this accessibility empowers patients by giving them more control over their health, it also creates a potential avenue for misuse. In particular, a patient could intentionally modify their medical data—such as altering an X-ray image—to manipulate the diagnostic outcome in their favor. This manipulation might be motivated by a variety of reasons, such as gaining access to specific medical treatments, qualifying for government assistance programs, or obtaining financial benefits. For example, a diagnosis of COVID-19 could allow a patient to receive government support programs, specialized treatments, or paid sick leave. Similarly, patients may alter their diagnosis to gain eligibility for subsidized medications, participation in clinical trials, or access to disability benefits. Being classified as having diabetes for example may allow patient's to be prescribed Ozempic or other lucrative diabetese drugs at substantial discounts. These motivations, coupled with the relative ease with which adversarial examples can be generated, raise critical concerns regarding the security of AI-driven medical diagnostic systems.

In this work, we aim to emphasize the risks posed by adversarial examples in medical diagnostics. We demonstrate how an adversary could easily manipulate their own X-ray scans to change their medical diagnosis (e.g. to obtain a COVID-19 positive diagnosis) using two gradient-based attacks: Universal Adversarial Perturbation (UAP) [23] and Projected Gradient Descent (PGD) [4]. We show how these attacks lead significant drops in accuracies across two popular deep learning architectures - VGG16 [19] and EfficientNetB0 [20]. This attack vector poses a highly practical and significant threat to healthcare providers. It is motivated both personally and financially, is easy to execute with low barriers and is highly scalable. By distributing tampered data to multiple healthcare providers simultaneously, attackers significantly increase their chances of achieving favorable outcomes. We review existing methods for mitigating such attacks and discuss the limitations of each method. Furthermore, in the field of medical X-ray imaging, much of the work in the literature, such as in [13] [7] [8], has focused on only analyzing the impact on models of perturbations produced by UAPs and PGDs. We go one step beyond by not only analyzing the impact of these attacks but also proposing a simple, lightweight but highly accurate detection framework designed to identify when X-ray images have been perturbed. This enables healthcare professionals to detect adversarial examples in the first place before beginning to analyze them. Furthermore, we also explore the generalizability of our detector trained on UAP to PGD, thereby experimenting the effectiveness of our detection method against a variety of adversarial attacks. Overall, our paper contributes to the growing body of research on improving the security and robustness of AI systems in healthcare.

2 Background and Related Work

2.1 Adversarial Examples in Images

As many have pointed out, deep learning models are highly vulnerable to adversarial attacks, especially in image processing. Attacks can be broadly categorized into white-box attacks and black-box attacks. White-box attacks assume complete knowledge of the target model, including its architecture, parameters, and training data. Black-box attacks, in contrast assume no knowledge of the target model. Thus, while black-box attacks may provide a more realistic threat model, they are also more difficult to implement.

Adversaries may want to perform two types of attacks on ML models: targeted and untargeted. In a targeted attack, the goal is to manipulate input data to force the model to predict a specific incorrect class, such as making a COVID-19 X-ray appear normal. In contrast, an untargeted attack aims to cause any misclassification without preference for a particular outcome; for instance, it may result in a COVID-19 X-ray being classified as either normal or pneumonia. Untargeted attacks are generally easier to implement since they do not require controlling the specific misclassification outcome. While untargeted attacks may suffice in many scenarios, some adversaries may need to target a specific class to achieve their objectives. Ultimately, the choice between targeted and untargeted attacks depends on the adversary's goals and the specific context of the situation.

Many of the most successful attacks are untargeted white-box attacks that leverage model gradients. Goodfellow et al. were the first to propose gradient based attacks with their Fast Gradient Sign Method (FGSM). Their attack generates adversarial examples by applying a small perturbation in the direction of the gradient of the loss with respect to the input image [5]. While an impactful idea, the FGSM's effectiveness can be easily defended. The Projected Gradient Descent (PGD) attack improves on FGSM by applying multiple smaller perturbations iteratively [10]. This technique works surprisingly well and PGD is considered one of the strongest first-order attacks. There also exist non-gradient-based attacks. DeepFool for example attempts to identify a model's decision boundaries and apply the smallest perturbation to shift an example to a different decision boundary [12]. The authors emphasize the subtly of these perturbations compared to the aforementioned gradient based approaches.

When direct access to a model is not feasible black box attacks can be a successful alternative. Many of the most successful black box attacks attempt to estimate the victim model's gradients and use them to perform gradient based attacks. One approach to estimate the victim's gradients is by training a surrogate model and use the gradients of this model directly [9]. These transfer-based attacks can be quite effective and difficult to defend against. Another method to estimate gradients is with query-based attacks [2]. These attacks query a black-box model many times and approximate gradients with finite-differences methods. If this is feasible in the threat model, and many queries can be performed efficiently, query-based attacks can also be quite effective.

In healthcare, adversaries can't observe the hospital's model and are thus limited to black-box attacks. Furthermore, querying the model many times is unrealistic. Our methodology assumes access to a publicly available dataset, enabling an adversary to train a surrogate model that approximates the gradients of the hospital's model. With the increasing availability of public medical datasets, our methodology is both highly practical and adaptable to a wide range of scenarios.

2.2 Adversarial Defenses on Images

Despite extensive research in controlled settings, real-world defenses against adversarial attacks remain challenging. A major hurdle is the transferability of adversarial examples, where perturbations effective on one model can often succeed on others [21]. A common approach to defending against adversarial examples is with adversarial training [5]. This involves augmenting the training data with adversarial examples to expose the model to potential attacks. While this can improve robustness against attacks similar to those seen during training, it's not a universal solution. Furthermore, training in this way can be computationally expensive and negatively impact the model performance on unperturbed examples. Another common approach is to use knowledge distillation [14]. By training a model to predict soft label outputs from another model, gradients become less informative to prospective attackers. However, such is also computationally expensive as it requires training both a teacher and student model. Furthermore, regularizing the teacher model in this way can also impact performance. Another approach is adversarial detection [11]. Rather than try to make a model robust against adversarial examples, detection involves training an auxiliary model to detect which examples have been altered. These models can be used in conjunction with existing models without impacting their performance. They can also be trained easily and tuned to a desired false positive rate. Given the high impact of incorrect diagnosis and the low cost for a hospital's to classify images as being altered when it's really not, detection methods are a promising solution. In our work, we demonstrate how detection systems can be trained easily and yield impressive results.

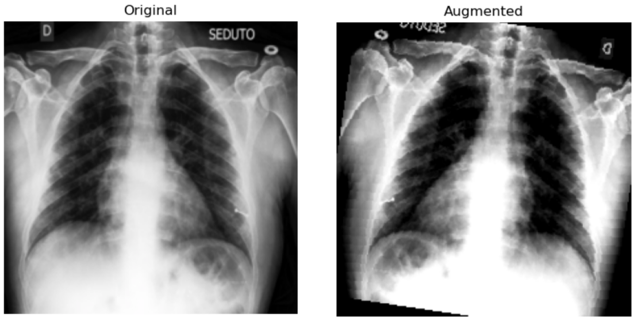
2.3 Adversarial Attacks in Medical Imaging

Adversarial attacks are an important consideration in medical imaging due to the critical nature of the domain. A variety of the aforementioned attack strategies. FGSM and PGD have been adapted to fool models trained on tasks such as tumor detection and disease classification [16]. In the black-box setting, transfer and query-based attacks have also been shown to be effective against medical imaging systems [6]. These findings emphasize the need for robust and domain-specific defenses to protect medical imaging applications from adversarial threats. Given the high impact of this domain, defending against these attacks is a prominent area of research. Many features of healthcare systems seem to signify a domain specific solution is necessary. Popular methods such as knowledge distillation and adversarial training can negatively impact critical diagnosis decisions making them less ideal in this high impact setting. Researchers have instead looked to utilize preprocessing methods, such as noise filtering or adversarial perturbation removal, to restore tampered images to their original state without degrading diagnostic quality [1]. Detection based systems are also prominent approach in the medical realm as they preserve model performance on clean data, have low computational overhead, better interpretability and maintain compatibility with existing models [3]. There seems to be little downside in training a detection system and for these reasons this is the approach we recommend to healthcare practitioners.

3 Threat Model

We consider a scenario where an adversary gains access to their own chest X-ray scan. They can easily obtain this scan by requesting a hard copy or digital version from a healthcare provider. This is a standard practice for patients seeking second opinions or maintaining personal medical records. The adversary's objective is to maliciously modify this scan and submit it to another healthcare provider, tricking their diagnostic system into classifying the image as COVID-positive. Motivations for such an attack may include obtaining paid sick leave, accessing medical treatments, or exploiting pandemic-related benefits. By altering their X-ray, the adversary aims to achieve these goals without detection, presenting a significant security challenge for AI-driven diagnostic systems.

In this threat model, we assume the adversary has access to a dataset containing chest X-rays paired with medical diagnoses. With the increasing democratization of ML resources on platforms like Kaggle and HuggingFace, this assumption is realistic and reflects current trends in data availability [22]. The adversary's dataset need not match the one used by healthcare providers; we hypothesize that fundamental features used for COVID-19 classification—such as lung opacity patterns—are consistent across datasets. While variations in imaging equipment or procedures may introduce differences in scale, resolution, or orientation such don't impact the core features required to trick a classification model.



(a) Original X-ray Image (b) Augmented X-ray Image

Figure 1: Original versus Augmented X-ray Image

With a training dataset, the adversary doesn't require black or white box access to the target's model. Instead the adversary attempts to exploit the consistencies present in COVID Chest X-Rays by training a surrogate model to mimic the victim's model. Using their proxy model, the adversary generates adversarial examples through gradient-based attacks. These carefully crafted perturbations are imperceptible to human observers, ensuring that the altered X-ray appears unchanged while successfully misleading the diagnostic model. The attack leverages the surrogate model's decision boundaries to manipulate the output, targeting the classification label for COVID-19. Because the perturbations are based on fundamental features of Chest X-Rays they are designed to transfer effectively to any healthcare provider's diagnostic system. Moreover, this attack can be performed at scale by submitting altered images to many providers simultaneously. This significantly increases the likelihood of obtaining the desired diagnosis from at least one provider, with no apparent downside from the adversary's perspective.

4 Experimental Setup

4.1 Dataset

We use a publicly available dataset from Kaggle [17], which contains X-ray images categorized into three classes: COVID-19, Pneumonia, and Normal. The goal is to train a model to classify images accurately into these categories. The training dataset contains 111 instances of COVID, and 70 instances each of Normal and Pneumonia. The test set contains 66 total examples, out of which 26 are COVID cases and 20 each are

Normal and Pneumonia. This dataset presents a challenge in terms of both size and class distribution, making it an interesting case for evaluating adversarial attacks and the robustness of AI models.

In real-world healthcare scenarios, hospitals often collaborate using federated learning approaches to train ML models across decentralized datasets without sharing sensitive patient data. Instead of transmitting patient data to a central server, hospitals train models locally on their own datasets and send only the model updates to a central server; these updates are aggregated to refine a global model. This approach allows hospitals to build powerful models while maintaining patient privacy and adhering to data protection regulations.

To simulate an environment where data coming from multiple hospitals are leveraged to train a more powerful model, we apply several data augmentation techniques such as random horizontal flips, rotations and adding random color jitters. We show an example of these augmentations in Fig. 1. Observe how the augmented image is horizontally flipped, is rotated by a small angle clockwise, and is also brighter than the original image. In image classification tasks, these augmentations are commonly employed to improve model robustness by expanding the training data and enhancing decision boundaries. In our case, these augmentations represent the variation in imaging equipment and protocols across different hospitals. We train the victim model (the hospital's model) using multiple augmented versions of the training data. In contrast, the adversary's model is trained using only the raw, training data. This setup mimics real-world conditions and ensures the victim and adversarial models are not trained on the same data distribution, which would otherwise make it easier to identify adversarial examples. We train our models for up to 50 epochs, each representing a full pass over the training set. We employ early stopping with a patience of 5, halting training if performance fails to improve for 5 consecutive epochs.

4.2 Implementation Details

We use two popular deep learning architectures that excel at classification tasks to train our victim model – VGG16 and EfficientNetB0.

We suspect that adversarial perturbations are most effective if the victim model and the adversary's model have the same architecture. However, to simulate a realistic scenario, we assume that the attacker does not have this knowledge. Thus, the adversary uses a ResNet18 architecture to train their surrogate model and generate adversarial perturbations. Given the consistency in X-ray data, we hypothesize different architectures rely on a similar set of high level features when making their decisions, making adversarial attacks highly transferable.

To generate the adversarial perturbations, we leveraged the gradients from the adversary's model to perform gradient based attacks. We implemented two attacks: Universal Adversary Perturbations (UAP) and Projected Gradient Descent (PGD) [23]. UAP generates a single perturbation vector that, when added to multiple inputs, can cause misclassification. Thus, it generates one perturbation that is effective across multiple inputs, which makes it versatile, robust to transformations, and also transferrable across many neural networks. Prior to being fed to a deep neural network, the adversarial example may undergo transformations, such as changes in pixel intensity and scaling when an X-ray hard-copy is being scanned by the hospital. However, UAP attempts to build perturbations that are robust to these shifts. On the other hand, PGD generates perturbations by optimizing for misclassifications while staying within a specified perturbation budget [4]. Both UAP and PGD rely on hyperparameters such as epsilon (ε) and the maximum number of iterations, which we set to 0.01 and 10, respectively, following standard practices. The value $\varepsilon = 0.01$ specifies the maximum allowable perturbation magnitude, ensuring changes remain imperceptible. The maximum iterations determine the number of steps the attack takes to refine the perturbation for maximum effectiveness. Additionally, PGD uses α, set to 1, which controls the step size of perturbation adjustments at each iteration.

5 Results and Analysis

5.1 Adversarial Examples

Our next goal is to observe how effective our perturbations are in impacting our two victim models. For UAP, the process is simple. Based on the training set, we produce a single, general perturbation that we now add to each of the examples in our test set. For PGD, the approach is slightly different. Since PGD generates a unique perturbation specific to each input, we iterate through each example in the test set, generate its specific perturbation, and add apply it. Ultimately, this process yields two sets of perturbed images: one generated by UAP and the other by PGD, which we then evaluate for effectiveness.

As seen in Fig. 2, we observed significant drops in accuracy for both VGG16 and EfficientNetB0 under UAP and PGD, with the drops being notably similar across attacks. We hypothesize this high similarity could be due to choosing the same hyperparameters for UAP and PGD to ensure consistency as well as both being gradient based attacks. Most importantly, despite the perturbations being generated on a completely different surrogate model (ResNet18), they still demonstrated high transferability by greatly impacting two different deep learning architecture, highlighting the transferability of our adversarial perturbations.

We use confusion matrices to evaluate the robustness of hospital models against these transfer attacks. We assign the labels as follows:

- 0 = COVID
- 1 = Normal

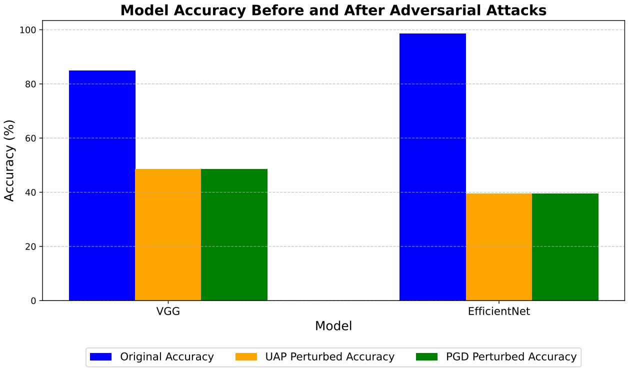
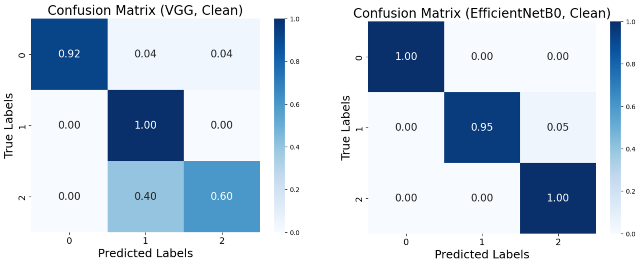


Figure 2: Model Accuracy Before and After Adversarial Attacks. This figure shows the original accuracy and the accuracy after being perturbed by UAP and PGD attacks for two victim models - VGG16 and EfficientNetB0.



(a) VGG16: Clean Images

(b) EfficientNet: Clean Images

Figure 3: Confusion matrices for VGG16 and EfficientNetB0 on clean images.

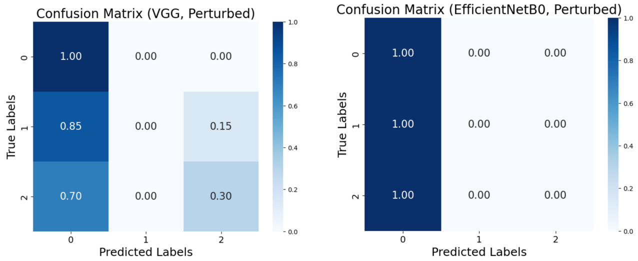
- 2 = Viral Pneumonia

In each confusion matrix, the y-axis represents the true label and the x-axis represents the predicted label. For instance, in Fig. 3a, the first row shows that:

- 92% of all COVID cases were correctly classified as COVID
- 4% of all COVID cases were incorrectly classified as Normal
- 4% of all COVID cases were also incorrectly classified as Viral Pneumonia

The confusion matrices for the clean images highlight our model's strong ability to detect ailments in patients across all three classes.

In contrast, we show the confusion matrix for the perturbed images using UAP on our models in Fig. 4. Observe that misclassificaiton is severe across all all cases, across both models. Notably, UAP on EfficientNetB0 misclassifies 100% of Normal and Viral Pneumonia cases as COVID—ideal for someone seeking a false positive. Similar patterns are observed with PGD, so its confusion matrices are omitted.



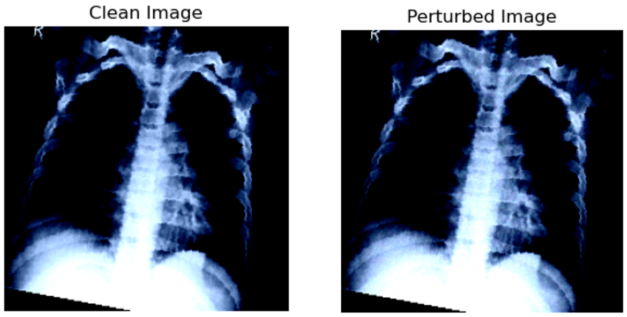
(a) UAP on VGG16

(b) UAP on EfficientNetB0

Figure 4: Confusion matrices for VGG16 and EfficientNetB0 on UAP perturbed images.

5.2 Detection Network

The applied perturbations are so minimal that they are imperceptible to the human eye. The clean and perturbed images appear virtually identical, as illustrated in Figure 5.



(a) Clean X-Ray Image

(b) Perturbed X-ray Image

Figure 5: Confusion matrices for VGG16 and EfficientNetB0 on UAP perturbed images.

This motivates us to train a model capable of classifying X-ray images as either perturbed or unaltered. To achieve this, we augment the training set to simulate X-rays from diverse hospital settings. We then apply the previously generated UAP perturbation to each image, creating a balanced dataset with equal representation of clean and perturbed images.

Using this set of clean and perturbed images, we train a simple deep learning model with only linear layers and ReLU activation functions. To test out this detector, we pass our test set, which the model has not seen before, to our trained model and achieve a **95% accuracy**; this is very impressive given the simplicity of our model.

As a separate experiment, we take our test set, add PGD perturbations, and see if the model can still differentiate between the PGD-perturbed images and the clean images. The model still achieved an impressive **72% accuracy**, despite only being trained on UAP-perturbed images. This showcases the generalization of our model to another gradient-based attack, which is PGD. However, to make the model as robust as possible, training on images perturbed by a wide range of attacks is crucial and we save this for future work. It might also be interesting to explore more sophisticated models, such as those using convolutional neural networks (CNNs) or Vision Transformers [15], to improve our model's detectability.

5.3 Model Interpretations with Grad-CAM

As another line of defense, hospitals can consider using visualization techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) [18] in addition to our detection framework. Grad-CAM highlights the pixels that most affect model predictions and can help clinicians visually assess whether the model made its decision based on relevant features or meaningless patterns caused by adversarial perturbations. This would provide extra support in addition to our detection framework to ignore adversarial images.

By taking the gradients of the target class flowing back to the final convolutional layer, Grad-CAM generates a heatmap on top of the original image. Warmer colors (e.g., red) represent areas with higher importance than those with cooler colors (e.g., blue). Figure 6 highlights what regions of the images were most focused on by the model when making the predictions on a sample X-ray image. We observe significant differences in heatmaps between the original and perturbed images. For instance, the area around the neck appears warmer in the original image, indicating it plays an important role in the model's prediction. In contrast, the neck area appears cooler for the perturbed image, suggesting the model doesn't place much emphasis on it when making its prediction. This type of decision-making, where regions expected to be significant in diagnosis not appearing as important, may seem illogical to healthcare professionals, helping them identify that the original image has been altered.

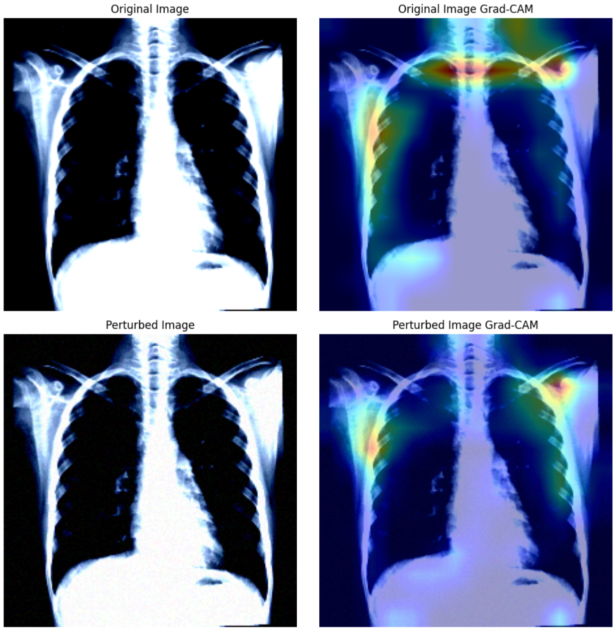


Figure 6: Pixels deemed to be most significant by the model in making predictions or the original (top) and perturbed (bottom) images.

6 Discussion

6.1 Implications for Healthcare

While we specifically study how an adversary can modify their chest X-ray scans to be falsely diagnosed as having COVID-19, our results have wide reaching impacts. With small modifications, similar attacks can be applied to any healthcare image based classifier so an adversary can receive a desired diagnosis. Furthermore, we suspect that slightly different techniques could extend these vulnerabilities to tabular medical AI models as well.

Our attacks only require access to widely available training datasets, which are abundant in healthcare. Examples include LUNA2016 for lung nodule detection, ISIC for melanoma classification, and UCI's Diabetes 130-US Hospitals dataset for diabetes diagnosis and treatment outcomes, among many others. A determined adversary could easily download one of these datasets and manipulate the input data to influence the diagnosis in their favor.

Our study highlights the adversary's motivations for manipulating COVID-19 chest X-rays, driven by personal or financial incentives such as obtaining paid leave, disability benefits, avoiding quarantine restrictions, or circumventing travel disruptions due to a positive COVID-19 diagnosis. In healthcare, financial motivations are significant. For instance, the diabetes medication Ozempic (semaglutide), used for type 2 diabetes and weight management, costs nearly $1,000 per month without insurance in the U.S. Insurance typically covers this drug for diabetes patients, making it financially accessible. Given the obesity epidemic and its association with greater health risks than smoking or poverty, many individuals are eager to access this life-saving medication. However, for many, this drug is only affordable with insurance coverage. Consequently, adversaries have a clear financial incentive to deceive hospital diabetes classifiers to secure insurance coverage for Ozempic.

In our study, we emphasize how adversaries can easily execute targeted attacks. Untargeted attacks are equally relevant in healthcare settings and often easier to carry out. As mentioned previously, an attacker might seek any diagnosis other than COVID-19 to avoid travel disruptions and restrictions. Another scenario involves a pilot diagnosed with depression. Under strict Federal Aviation Administration (FAA) regulations, such a diagnosis can result in the loss of their pilot's license. Regaining it is a complex process, requiring evaluation by an FAA-approved specialist, passing extensive tests, and waiting up to six months for medical clearance. By tricking an ML model into assigning any other mental health diagnosis, a pilot could potentially bypass these hurdles entirely.

For both targeted and untargetd attacks, the barrier to entry is alarmingly low. Adversaries can leverage freely available resources, like Google or ChatGPT, to learn how to generate adversarial examples. If needed, they could even hire ML professionals via platforms like Upwork for a modest fee (e.g., $50/hour). Depending on the financial or personal benefits of a misdiagnosis, such an investment could easily pay off. Moreover, given the significant incentives and ease with which an attack can be performed there is a significant need to ensure medical AI systems are robust to adversarial examples.

6.2 Limitations

Our detection network demonstrated impressive results, but we suspect two main factors contributed to this outcome. First, the perturbations introduced may have been excessively large, making the altered images easily detectable. This is supported by the fact that nearly all UAP-perturbed images were classified as COVID-19, suggesting that the single perturbation vector generated by UAP was large and conspicuous. However, as shown in Figure 1, these perturbations are not visually apparent and would likely go unnoticed without a detection model. Second, the network's strong performance may stem from its training and evaluation being conducted with gradient-based perturbations. Although we used different types of gradient-based attacks for training and testing, there was likely some learnable overlap. While gradient-based attacks are the most common and successful in the literature, adversarial attacks are not limited to these methods, and our detection network might not perform as well against other strategies, such as decision-boundary-based attacks. To ensure robustness, detection networks should be trained on a diverse range of attack strategies, a task we leave for future work.

Our work is also limited in that we don't consider practical transformations on the adversarial examples. As mentioned previously, an adversary may print his X-ray that is then scanned by a hospital exposing it to potential shifts in pixel intensities and scales. While we considered UAP to ensure attacks are robust to these transformations, we lacked the time and resources to adequately test the robustness of these examples. In future work we plan to study the impact transformations of the adversarial examples have on both the hospital's model and the hospital's detection model.

Finally, our analysis is limited to a single dataset of images. It remains to be seen whether these results can be replicated in other datasets and with tabular data. Moreover, in future work we plan to consider building adversarial examples and detecting them on a wider range of datasets as well as in tabular data.

7 Conclusion and Future Directions

This paper demonstrates the vulnerability of AI-based medical diagnostics to adversarial examples and proposes a detection framework to address this issue. In addition, we also explore explainable AI techniques, such as Grad-CAM, that supplements our adversarial detection framework and assists healthcare professionals to make X-ray diagnostics. For the future, we can consider using a more comprehensive dataset with a lot more training examples from various hospitals instead of creating a diverse, simulated environment. Additionally, exploring how our UAP and PGD perturbations affect other deep learning architectures could also be of keen interest. Overall, our findings emphasize the need for secure and trustworthy AI systems in healthcare.

References

[1] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S. Sara Mahdavi, Ellery Wulczyn, Boris Babenko, Megan Wilson, Aaron Loh, Po-Hsuan Cameron Chen, Yuan Liu, Pinal Bavishi, Scott Mayer McKinney, Jim Winkens, Abhijit Guha Roy, Zach Beaver, Fiona Ryan, Justin Krogue, Mozziyar Etemadi, Umesh Telang, Yun Liu, Lily Peng, Greg S. Corrado, Dale R. Webster, David Fleet, Geoffrey Hinton, Neil Houlsby, Alan Karthikesalingam, Mohammad Norouzi, and Vivek Natarajan. Robust and efficient medical imaging with self-supervision, 2022.

[2] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, CCS '17. ACM, November 2017.

[3] Samuel G. Finlayson, Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam. Adversarial attacks against medical deep learning systems, 2019.

[4] Simon Geisler, Tom Wollschläger, MHI Abdalla, Johannes Gasteiger, and Stephan Günnemann. Attacking large language models with projected gradient descent. *arXiv preprint arXiv:2402.09154*, 2024.

[5] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[6] Kim Hee, Alejandro Cosa, Nandhini Santhanam, Mahboubeh Jannesari, Mate Maros, and Thomas Ganslandt. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22, 04 2022.

[7] Hokuto Hirano, Kazuki Koga, and Kazuhiro Takemoto. Vulnerability of deep neural networks for detecting covid-19 cases from chest x-ray images to universal adversarial attacks. *Plos one*, 15(12):e0243963, 2020.

[8] Hokuto Hirano, Akinori Minagi, and Kazuhiro Takemoto. Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging*, 21:1–13, 2021.

[9] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks, 2017.

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[11] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations, 2017.

[12] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks, 2016.

[13] Shantanu Pal, Saifur Rahman, Maedeh Beheshti, Ahsan Habib, Zahra Jadidi, and Chandan Karmakar. The impact of simultaneous adversarial attacks on robustness of medical image analysis. *IEEE Access*, 12:66478–66494, 2024.

[14] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks, 2016.

[15] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[16] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, page 493–501, Berlin, Heidelberg, 2018. Springer-Verlag.

[17] Pranav Raikokte. Covid-19 image dataset. https://www.kaggle.com/datasets/pranavraikokte/covid19-image-dataset, 2023. Accessed: 2023-11-23.

[18] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[21] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses, 2020.

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

[23] Changming Xu and Gagandeep Singh. Robust universal adversarial perturbations, 2023.