Evaluating the Vulnerability of Deep Learning Models in Medical Imaging to Adversarial Perturbations

Hamuza Senyonga
Yeshiva University - Cybersecurity

Charity Mahwire
Yeshiva University - Biotechnology Management and Entrepreneurship

Thelma Chimusoro
Yeshiva University - Biotechnology Management and Entrepreneurship

Enock Katenda
Yeshiva University – Computer Science

ABSTRACT
Deep learning has revolutionized medical imaging, but it is vulnerable to adversarial attacks, which are deemed dangerous to clinical use. This paper compares the strength of convolutional neural networks (CNNs) and Vision Transformers (ViTs) that are trained on the ChestX-ray14 dataset at NIH to detect pneumonia. Both models showed high baseline accuracy (>90 percent) even when the models were attacked by Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), DeepFool, and Carlini and Wenger, although PGD and CW were the most disruptive. The evaluation of defense strategies was also performed, which involves adversarial training, input preprocessing, ensemble modelling, and adversarial detection. Adversarial training provided the best protection, at the cost of lower clean-data accuracy and preprocessings and ensembles offered partial resistance, and also detection strategies identified a lot of naive adversarial inputs. There was however no one defence that was enough to counter every assault. The discoveries reveal the necessity of layered defence practices and ethical and regulatory issues related to trust, liability, and patient safety, which supports the significance of strong and transparent AI in the field of healthcare practices.

General Terms
Artificial Intelligence, Machine Learning, Healthcare Technology, Cybersecurity, Medical Ethics, Regulation of AI Systems

Keywords
Adversarial Attacks; Deep Learning; Medical Imaging; Convolutional Neural Networks (CNNs); Vision Transformers (ViTs); Robustness; Adversarial Training; Healthcare AI; Patient Safety; Cybersecurity in Medicine

1. INTRODUCTION
In the last decade, deep learning transformed the field of medical imaging and now allows to automatically analyze radiologic scans with a higher level of accuracy, similar to expert clinicians. Neural networks have shown levels of expertise in diverse diagnostic procedures, including skin lesion classification, and chest X-ray interpretation [1, 2]. Indicatively, in 2018, the first U.S. FDA-approved diabetic retinopathy deep learning system was on the list of AI-diagnostic tools, which indicates the increasing nature of AI-diagnostic integration into the clinical workflow [3]. The key to this fast development is that diagnostic reliability and patient safety should be considered when applying AI: doctors and patients should be capable of believing that such models will work properly and contribute to, not threaten, the process of medical decision-making.

Nevertheless, there has now been a serious issue that has arisen: the reliability of deep learning models and the loss of trust cause adversarial vulnerabilities. Scholars have discovered that minimal, a well-crafted interventions to medical pictures, which are sometimes entirely unnoticeable by the human eye, can even influence AI models to provide completely flawed diagnoses with high levels of certainty [4]. The integrity and privacy of the training data in such sensitive deployments is paramount, a challenge previously addressed by developing secure, distributed AI models using federated learning architectures [5]. Such intelligently designed inputs are referred to as adversarial examples, and they take advantage of the vulnerabilities in the decision limits of the model. Patient safety is directly threatened by the existence of adversarial attacks: a slightly modified CT scan or X-ray may deceive an AI to fail to recognize a tumor and incorrectly diagnose a condition, resulting in wrong treatment [2, 4]. These types of attacks undermine the integrity of AI-based decisions and represent a new category of security threat in clinical settings [6].

The possible outcomes of confrontational escapades in the medical field are dire. In diagnostic imaging, an attacker may be able to use manipulated input scan to hide important results or create a disease and avoid automated and human quality checks [3]. Previous research cautions that adversarial weaknesses can lead to disastrous consequences such as false diagnoses, unnecessary treatment, fraud in insurance, and even a larger crisis of trust in AI-based medicine [4, 6]. An example is a study that showed that by introducing almost imperceptible noise to a retinal fundus image or a chest X-ray, the output of a model could be altered to appear as an image of a diseased person or a doctor prescribing the wrong treatment to an individual, which could potentially allow fraudsters to fake medical diagnoses or the wrong doctor to prescribe incorrect treatment to their patient [8]. Previous work established the efficacy of multi-tiered defense strategies, specifically for the Energy and Healthcare critical sectors, by integrating network segmentation, EDR, and offline backups to mitigate persistent ransomware threats [9]. Ranging in billions of dollars over healthcare decision making, these weaknesses offer some incentive to be abused by different entities [4]. This problem statement explains the importance of ensuring the security of deep learning systems in terms of patient trust and safety.

This paper provides a systematic study of adversarial attacks on

deep learning models in medical images, and the effectiveness of defensive measures against AI diagnostic systems. However, the specific objectives are:

a) To determine the categories of adversarial attacks shown to take place on medical imaging models and their effects on diagnostic performance

b) To evaluate defense systems (e.g. adversarial training, input filters, anomaly detectors, etc.) are most effective at enhancing model robustness

c) To determine how the ethical, legal, and clinical implications of adversarial vulnerabilities in healthcare AI can be explained, in particular, the trust and responsibility of patients

The present paper is dedicated to image-based AI diagnostics (e.g. radiology, pathology, ophthalmology images) where deep learning is used extensively and safety issues are the most critical. The computer vision field and healthcare field literature have been surveyed to relate the general concepts of adversarial machine learning with the medical practice. The importance of this work is explained by its interdisciplinary nature: technical knowledge about model security is associated with clinical and ethical aspects. Finally, the performance of AI in adversarial settings should be ensured to ensure not only accuracy but also patient and clinician trust in AI-assisted care. Through the study of existing vulnerabilities and defenses, as well as commenting on regulatory factors, timely advice on how to build resilient AI systems that can be safely incorporated into the healthcare environment was offered.

2. LITERATURE REVIEW

2.1 Deep Learning in Medical Imaging

In the last few years, the analysis of medical images has become an essential part of the field of deep learning, with Convolutional Neural Networks (CNNs) as its central part. The literature records the rapid development of the initial experiments up to massive applications in the field of radiology and other fields of application [7-11]. By 2017, CNNs were performing almost humanly on tasks such as skin lesion classification and screening retinal diseases. A groundbreaking study demonstrated a deep CNN to be as accurate as dermatologists in skin cancer distinction using dermoscopic photos [14]. Equally, deep models have been particularly successful in identifying diabetic retinopathy in fundus images and detecting pneumonia in chest X-rays [12, 13]. This achievement has already resulted in hundreds of AI models to interpret X-rays, CT scans, MRIs, pathology slides, and other forms of imaging [16]. One such moment was when, in 2018, the FDA gave regulatory approval to an AI-based diabetic retinopathy diagnosis device, highlighting the fact that these algorithms are no longer in research but in actual clinical use [3].

The increased use of deep learning to make crucial diagnoses implies that reliability is crucial. Doctors have to have confidence in the output of an AI to include it in clinical decision-making, and the patients have to be confident that AI-assisted diagnoses or treatment advice is accurate and safe. The medical AI ethics studies have pointed out that validation, transparency, and reliability are the main ways of earning this trust [17]. At this point, when deep learning models are highly complex black boxes, the clinicians tend to be more cautious and confirm AI output with their own judgment. Trust can be established and increased after an AI shows accurate performance consistently, which enhances the efficiency of the workflow and patient outcomes. Conversely, confidence can be destroyed at a very short notice upon any sign of unpredictable or erroneous conduct. The example is that with an AI program that periodically gives a flashing misdiagnosis (albeit uncommonly), clinicians will question everything it produces, which nullifies the value that it may be generating. In that way, a significant portion of the literature regarding AI in healthcare emphasizes on the concepts of robustness and reliability as the basis of trust in such tools [6]. Medical life and death decisions are stake, which are very low tolerance of mistakes or failure modes that are not known in systems with artificial intelligence.

2.2 AI adversarial attacks: Computer vision principles

Adversarial attacks demonstrate the underlying weaknesses of existing computer-vision systems. Initial experiments had demonstrated that small, usually unnoticeable alterations in the input images can result in high-confidence misclassification and Szegedy et al. [18] were the first to record this phenomenon as they demonstrated that well-crafted perturbations could consistently mislead deep neural networks. The finding led to the creation of algorithmic attack and defensive analysis. Goodfellow et al. [19] proposed Fast Gradient Sign Method (FGSM), a one-step method, which perturbs an image by following the gradient of the model loss in order to cause misclassification. These approaches were later generalized to iterative algorithms such as Projected Gradient Descent (PGD) that utilizes gradual gradient steps and can be known to generate strong adversarial samples [20]. Carlini and Wagner (CW) method is an example of optimisation-based attacks, where it minimises the distance between the adversarial samples targeting low distortion [21], and DeepFool reduces the norm of perturbation steps to find the nearest decision-boundary crossing [22]. Together, these attacks constitute a methodological repertoire of investigating robustness of models.

One notable fact about adversarial examples is that they transfer well: adversarial examples created by training on a single model are often effective in deceiving other models, even when the architecture and training data are different [21, 22]. Transferability allows black-box attacks in which an attacker trains a surrogate model and launches attacks on a target system without having any direct access. Recent studies indicate that a sense of transferability emerges due to data representations perceived as similar across networks, which presupposes that systemic vulnerabilities can be maintained in different model families [3]. Its practical implication is that publicly available models can be used as convenient templates to create attacks on proprietary medical AI systems, greatly expanding the threat space.

Adversarial research has expanded to include image classification and detection, segmentation, tracking of objects in video streams, and video streams. Most importantly, there are universal adversarial perturbations (UAPs) which are individual perturbation vectors that when added to a vast amount of inputs generate very high misclassification rates [25]. UAPs show that it is possible to build compact and reusable attacks, and this can be scaled in terms of adversarial campaigns. Such dangers have been confirmed by medical imaging experiments: controlled changes in diagnostic output of automated radiology systems can be caused by targeted perturbations, indicating the possibility of an extensive clinical effect [26]. Physical-world attacks also explain physical-world risk; printed stickers or graffiti have been demonstrated to confuse traffic-sign recognisers in autonomous vehicles, with implications on safety when adversarial examples are no longer in the form of digital pixels [27].

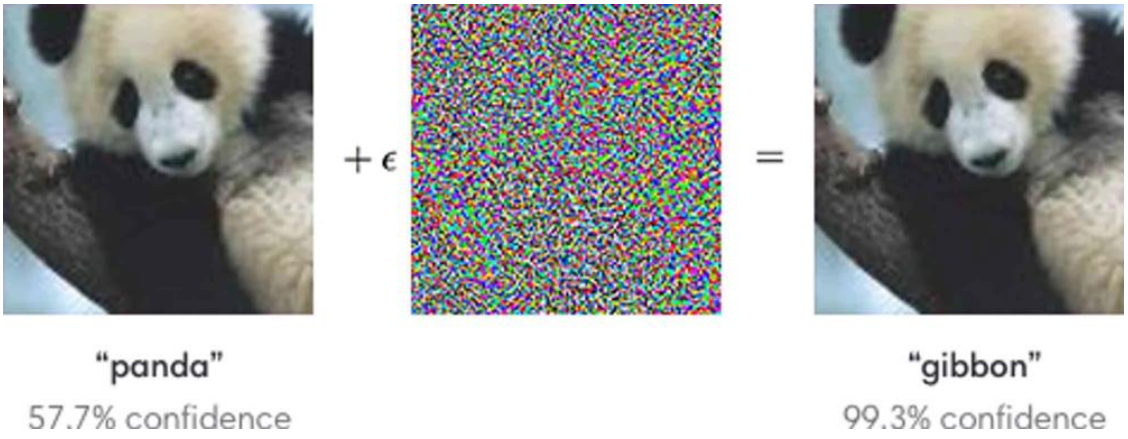

Fig 1: Adversarial perturbation example — image classification.

The picture on the left is the original image with the label panda (57.7%); the centre is the amplified perturbation pattern; the right-hand picture is the adversarial image with the label gibbon (99.3%) after such a small visible change [19].

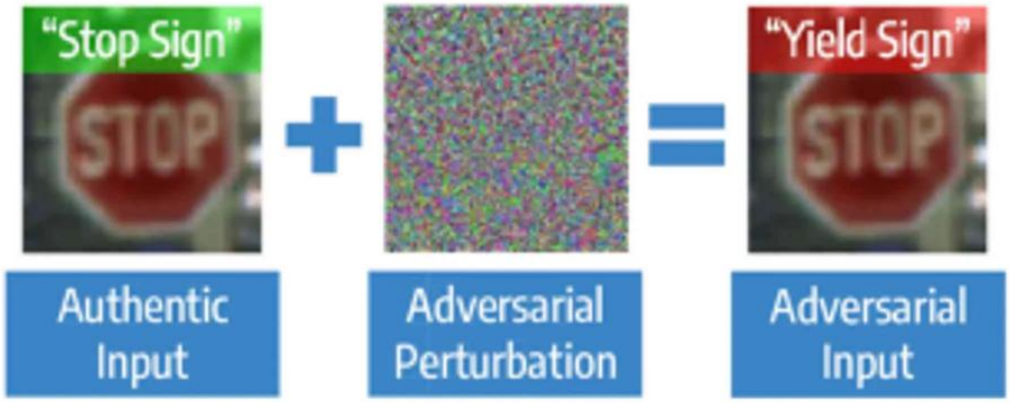

Fig 2: Physical-world adversarial attack on traffic sign recognition.

Left: well-placed stop sign; centre: perturbation stickers that have been applied; right: misplaced adversarial image with the label of Yield, showing physical exploitability [27].

The literature taxonomies include white box (full model access) vs black box (limited access), targeted vs untargeted, and image specific vs universal attacks. These differences are used to threat model medical AI: white-box situations are used to test worst-case robustness, and black-box transferability emphasizes realistic attacker abilities. The computer-vision body of work, therefore, provides both the theoretical foundations and practical algorithms necessary to assess and harden medical imaging systems against adversarial manipulation. These insights motivate the incorporation of adversarial robustness evaluation into any deployment pipeline for clinical AI.

2.3 Adversarial Attacks in Healthcare Imaging: Demonstrated Vulnerabilities

The concept of adversarial threats in healthcare imaging started to gain traction in 2018, when there was increasing awareness that the flaws of computer vision could also be applied to clinical imaging [3]. Preliminary studies by Paschali et al. [8] have revealed that the medical image classification and segmentation networks (including the networks used to perform the activities like skin lesion detection and brain segmentation with MRI) severely deteriorate in performance when attacked by the Fast Gradient Sign Method (FGSM) and DeepFool. More importantly, even the minimal perturbations, which are not noticeable to human observers, caused significant decreases in the accuracy of the diagnosis, disclosing that even the models which are otherwise similar to human observers can be turned into unreliable ones because of adverse manipulation. Ma et al. [12] noted in their follow-up analysis that certain architectures were marginally more resilient than others, but all tested models were at least partially vulnerable.

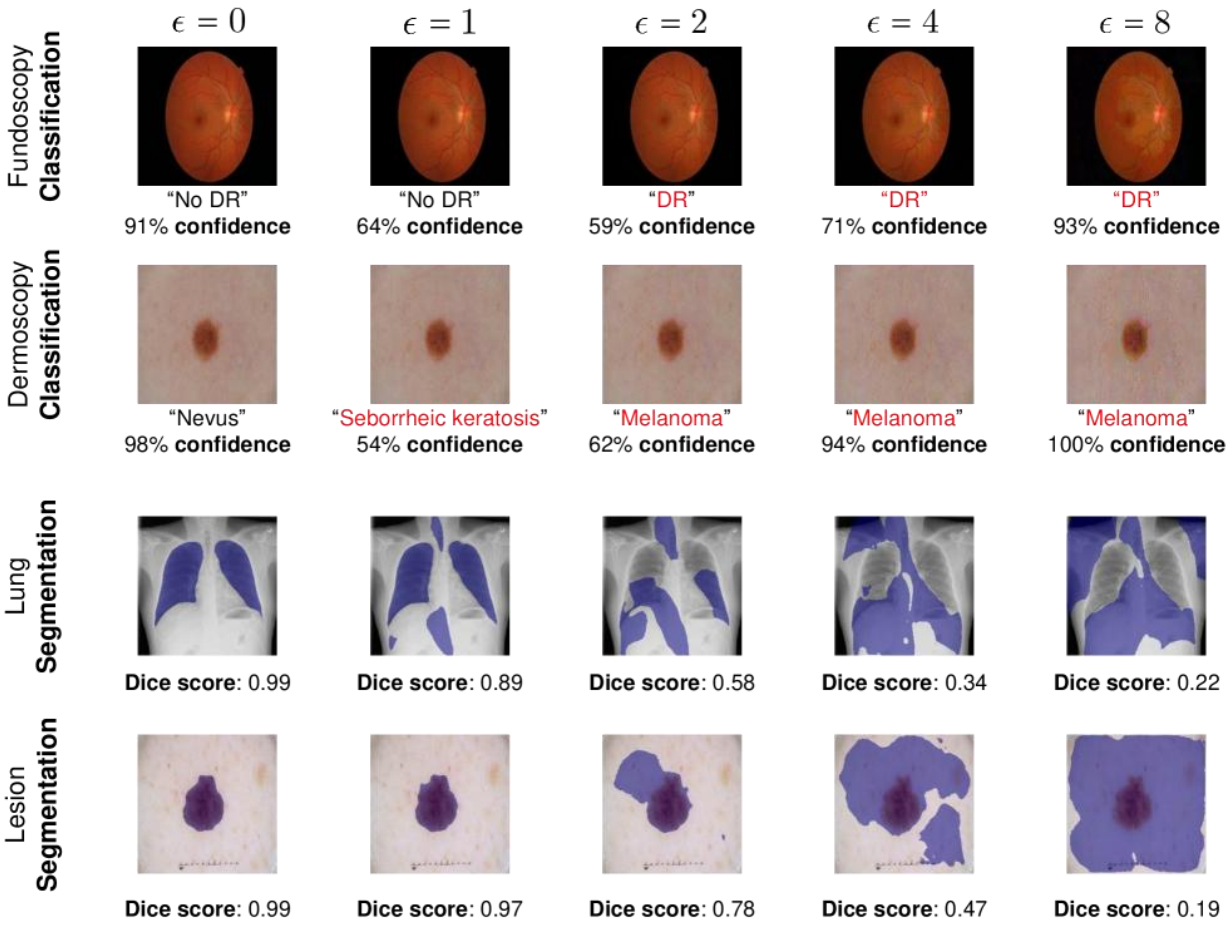

Fig 3: Adversarial examples in medical imaging

Visualization of medical adversarial examples with predictions under diverse perturbation size ϵ. The generated segmentation masks are superimposed on the original images for visualization (adapted from Dong et al. [3]).

This question attracted considerably more attention after a high-profile commentary by Finlayson et al. [4] in Science, where potential real-life scenarios were outlined. They gave examples of chest X-rays which were manipulated to give a benign image that was called pneumothorax, or a sick scan that was termed as healthy. These manipulations may help conduct fraudulent insurance claims, damage reputations, or even pose direct risks to patients [26]. As noted in the article, patients, healthcare professionals, insurers, and malicious external actors may all have incentives to pursue adversarial vulnerabilities.

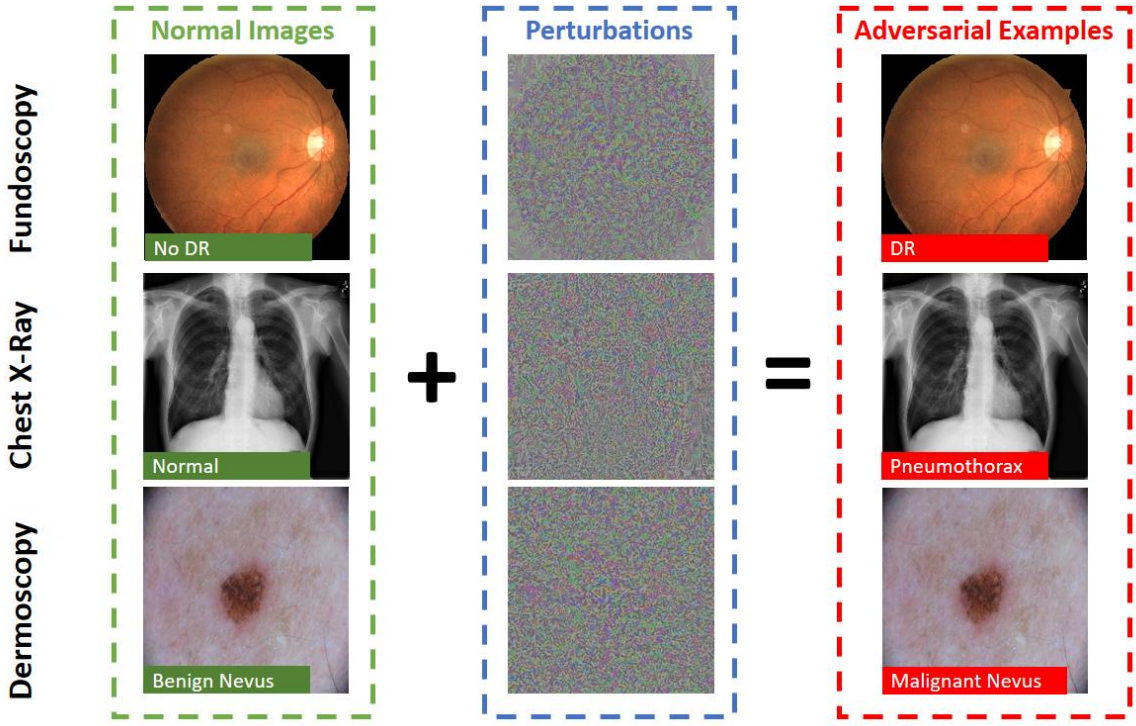

Fig 4: Medical adversarial perturbations

Top: diabetic retinopathy fundoscopy images; middle: chest X-rays; bottom: dermoscopy images. Left: control samples; middle: distortions manifested; right: adversarial counterparts resulting in false diagnosis. Green labels show accurate predictions, red errors (modified by Ma et al. [12]).

These worries were supported by empirical research. Dugas et al. [29] and Wang et al. [30] tested the adversarial attack in three modalities: funduscopies, chest radiographies, and dermoscopies and demonstrated a steep decline in model performance. In other experiments, melanoma classifiers deteriorated to a near-random score on adversarially perturbed images, 95% on clean images [12]. Kovalev and Voynov [31] proved these vulnerabilities using PGD attacks on chest X-rays and histology data, and Rao et al. [32] showed drastic drops in area-under-curve scores after applying the pneumonia-detecting network to the FGSM and PGD. Together, these studies determined that no broadly used medical imaging model was intrinsically resistant to adversarial perturbations, even when baseline accuracy was high on standard evaluation conditions.

Modality-specific and physical adversarial strategies have been since explored by researchers. Finlayson et al. [4] showed that small spots of adversarial images could easily deceive dermatology classifiers, even without other parts of the picture being damaged. Theoretical radiological extensions are placing stickers on the analogue films prior to digitisation or adjusting acquisition parameters in MRI scans. According to Sorin et al. [33], even the few radiology-specific studies available have reported adversarial vulnerabilities, such as in tumour detection in MRI and in the classification of lung nodules in CT images.

Such findings are also complicated by differences between medical images and natural images. Greater bit-depth, resolution and artefacts of modality like ultrasound speckle or noise in MRI modify the attack landscape. As Hirano et al [34] demonstrated, it is crucial to adapt to intrinsic noise distributions to construct universal perturbation of chest X-rays. However, they were able to generate perturbations that could alter a COVID-19 pneumonia detector on several scans of patients. Bortsova et al. [35] looked at parameters of image preprocessing and discovered that both typical augmentation and compression methods failed to eradicate adversarial vulnerability; in fact, these operations unintentionally reinforced adversarial effect on images.

2.4 Defense Strategies in Literature: Making Medical AI More Robust

The adversarial vulnerability of healthcare imaging models has motivated much study on defense mechanisms. Though a large number of these strategies have been developed in the wider adversarial machine learning community, they have subsequently been applied to the medical field where patient safety, diagnostic reliability, and trust take a central role. The existing methods may be classified into adversarial training, defensive preprocessing, architecture modification, adversarial example detection, and ensemble, or redundancy, approaches.

Adversarial training has been widely known to be one of the best methods used to enhance model robustness. It is a retraining of models with adversarially perturbed examples and clean images, and so the model has to generalise to both types of inputs. Initial studies by Ren et al. [36] revealed that an MRI segmentation network became significantly more robust when it was trained using FGSM-based perturbation. Equally, Rao et al. [32] established that pneumonia detection models that have been trained on the PGD inputs were resistant to test-time perturbation. Although this is effective, adversarial training causes computational overheads, which usually run training multiple times. There is also the trade-off of robustness versus clean accuracy when using the method. An example is the report by Madry et al. [20], which showed that a diabetic retinopathy classifier has preserved approximately 85% of its initial clean accuracy and 80% accuracy with substantial PGD attacks. This balance represents the conservative adjustment of

models, which are still not so prone to overconfident errors and are slightly less sensitive to fine diagnostic observations.
The second line of defense is defensive preprocessing or input transformation. The objective of these methods is to suppress or remove adversarial perturbations of images prior to their input into the network. Denoising filters, bit-depth reduction, and JPEG compression are usually used. Elsewhere in medical imaging, Taghanaki et al. [37] studied feature-preserving denoising in chest X-rays and discovered that perturbations could be reduced without a major change in anatomical content. Kansal et al. [38] applied the same principle to COVID-19 CT images and suggested guided filtering that proved to remove adversarial noise effectively without damaging clinically significant structures. The primary benefit of preprocessing is that it does not incur network parameter or retraining modifications, and is therefore light to implement. Nevertheless, malicious attackers may create perturbations that resist or take advantage of these changes, making their efficacy short-lived.

Another set of strategies includes model architecture and feature enhancement. In this case, the objective is to entrench the aspects of resilience within the feature extraction operation of the network. Taghanaki et al. [37] have mentioned that the use of average pooling in place of max-pooling layers resulted in a higher level of robustness due to the fact that more contextual information is retained. Bortsova et al. [35] developed this concept by incorporating guided filter layers in dermoscopy networks, which minimizes the impact of perturbations at feature intermediate levels. Knowledge distillation has been used as well: Liu et al. [39] trained segmentation models to generate more stable embeddings, which showed greater resistance to FGSM and iterative attacks. These architectural advancements have shown that resilience can be factored in the model design and not on post hoc fortifications only.

Another complement strategy is the detection of adversarial examples. These methods are not used to prevent misclassification, but to detect adversarial inputs and reclassify them to be looked at by humans. Ma et al. [12] demonstrated more than 98% AUC in differentiating adversarial manipulated chest X-rays and fundus images and clean inputs, finding that the network had unique activation patterns when subjected to adversarial perturbations. In a similar vein, Watson and Al Moubayed [40] also trained meta-classifiers on diagnostic model output and were able to detect FGSM and CW samples. These detectors ensure an extra layer of safety when used in clinical systems, and more specifically, adversarial perturbations might be more evident in structured medical data than in natural images. However, they also have to deal with the consistent threat of the so-called adaptive attacks that can be designed in such a way that they avoid detection.

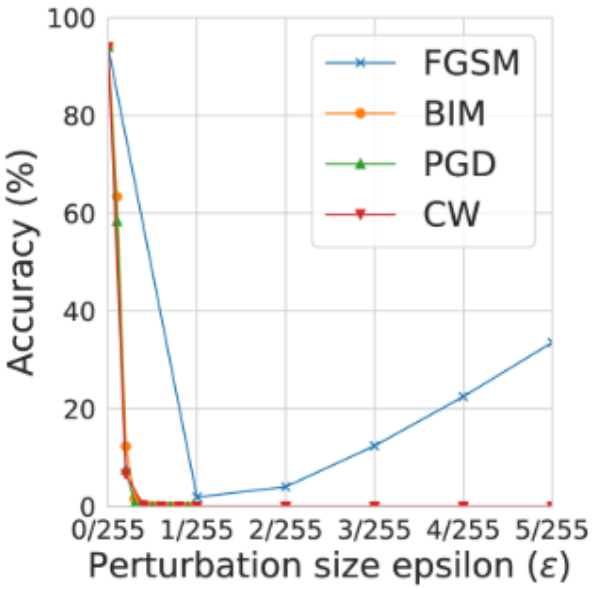

Fig 5: ROC curve of adversarial detection.

(adapted version of Ma et al. [12]].
Another line of defense is the ensemble and redundancy techniques. With the help of various models with different architectures, adversarial perturbations have lower chances of success across the board. Modifications to voting mechanisms may address misclassifications in models, and randomness, e.g. input transformations or dropout during inference, may destabilize the control of the attacker. Xie et al. [41] demonstrated how randomised transformations can be used to minimise the rate of successful attacks, where a similar technique can be applied to medical practice with minor rotation or crop of scans at the time of assessment. Also, by incorporating clinicians as human-in-the-loop individuals in ensemble pipelines improves reliability, especially when automated predictions are incompatible with anticipated clinical trends.
In the literature, it is always emphasized that no individual defense mechanism can be in total protection. According to Dong et al. [3], adversarial training is more effective in protecting against particular attacks but cannot generalise to novel perturbations, including adversarial patches. Preprocessing has the weakness of neutralising some noise patterns and is susceptible to adaptive exploitation, whereas detection systems are at risk of becoming outdated as attacks change. In that regard, researchers are more and more proposing defense-in-depth solutions, which consider a combination of adversarial training, preprocessing, feature robustness, detection, and ensemble techniques. This combination of measures has been demonstrated to be effective by recent experiments that have demonstrated that integrated defenses are able to maintain over 90 percent of diagnostic accuracy in chest X-ray models in strong adversarial examples and, at the same time, issue warnings on suspicious inputs to human review [6].

3. METHODOLOGY

A mixed-method research consisting of systematic literature review and controlled simulation experiments was used to explore adversarial vulnerabilities and defenses in medical imaging AI. The methodology involved two concomitant strands: (a) an exhaustive literature review to chart the already existing attacks, defenses and ethical issues and (b) an in-silico experimental analysis that modeled realistic adversarial attacks

and defensive measures against a standard chest X-ray classification problem. This two-sided strategy allows synthesizing concepts and illustrating them empirically by reproducible conditions.

3.1 Design of research and experimental justification

The research design used in the study was an exploratory-explanatory research design. The literature review came up with candidate attack algorithms, defense mechanisms and some general procedures of medical imaging; these results were used to plan the simulated experiments. The experimental part is a comparative one, not a clinical one: worst-case and transfer attacks were simulated by using public datasets and known model architectures and relative robustness was determined between models and defenses. The purpose was to shed light on dissipation processes of failure and to compare mitigation influences in controlled and repeatable environments, but not to verify a clinical device.

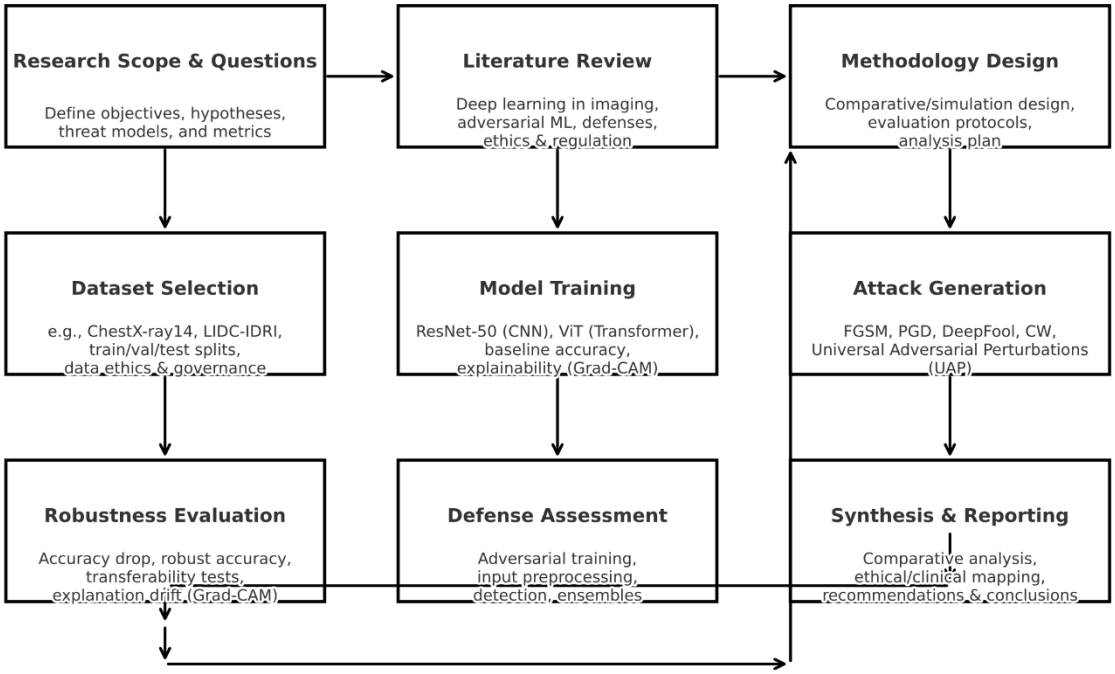

Fig 6: Research design (Flow Diagram)

3.2 Model selection and data sources

The source of the image was the NIH ChestX-ray14 dataset [30] as it is relatively large (over 100,000 frontal chest radiographs) and has multi-label pathology annotations. To make the classification a tractable problem, it was simplified to a binary classification: "Normal" and "Pneumonia" and the options associated with standard clinical decision boundaries and correspond to published benchmarks (e.g., CheXNet). Preprocessing of the images was done according to standard practices: scaling the images to 224x224, normalizing the intensity range to [0,1], and data augmentation (random crops, horizontal flips) were limited to non-diagnostic transformations.

Two representative model families that are designed to reflect the modern-day practice were chosen: a convolutional neural network (ResNet-50) and a Vision Transformer (ViT-base). Both were pre-trained on ImageNet and fine-tuned on ChestX-ray14. Architectures performed similarly to clean test images with Baseline performance of ResNet-50 approximately 93% accuracy and ViT approximately 92% accuracy, a realistic starting point on adversarial evaluation and the ability to evaluate architecture-dependent robustness.

3.3 Type of attacks investigated and implementation specifications

A collection of adversarial algorithms was used to test a variety of threat models (white-box, iterative, optimization-based, and universal). The selection of parameters reflects literature standards to make them comparable:

- **Fast Gradient Sign Method (FGSM)** [19]: single-step gradient perturbation $\mathbf{x}_{adv} = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \ell(h(\mathbf{x}), y))$. Low and moderate perturbation strength were tested with two levels of ε (0.01 and 0.03 on normalized pixel range).
- **Projected Gradient Descent (PGD) / BIM** [20]: iterative multi-step attack with random starts, 40 iterations, step size $\alpha = \epsilon/40$. Worst-case first-order attacks were based on a strong baseline of PGD.
- **DeepFool** [22]: untargeted untypical attack that minimises the norm, and is used to detect minimal-norm perturbs that cross decision boundaries.
- **Carlini–Wagner (CW)** [21]: L2-norm optimisation attack executed with a small confidence parameter to focus on imperceptibility; executed on a sampled subset as it is computationally expensive.
- **Universal Adversarial Perturbation (UAP)** [25]: a single-vector calculated on a training batch to evaluate the ability to test cross-image degradation.

Both transfer tests (perturbations created on resnet and applied to ViT and vice versa) and white-box experiments (full model gradient available) were run to recreate black-box surrogate tests. Measures were made of attack success rate (classification

flip rate), drop in AUC, and perceptual distortion (measured by L2 and SSIM). These metrics facilitate direct comparison of attack potency and model resilience across architectures.

3.4 Defense Methods Evaluated

The strategies of defense that were evaluated in this research were chosen because the strategies represent big division observed in the literature, and each strategy was evaluated against the adversarial conditions mentioned in the previous section. The goal was to determine the relative gains of robustness as well as the trade-offs in the accuracy and interpretability of clean-data. Four types of defenses were considered, namely adversarial training, input preprocessing, adversarial example detection and ensemble modelling.

The adversarial training was conducted on the base ResNet-50 architecture and fine-tuning was done on clean and adversarial images obtained by Projected Gradient Descent (PGD). In every training epoch, real-time PGD example against the current model parameters was generated, and the classifier was trained to learn to identify unperturbed and perturbed chest X-rays. The result of this process was a well trained model whose performance was again tested on the same set of attacks. As expected in the literature, adversarial training gave a strong resilience enhancement, especially in the case of FGSM and PGD attacks, at the cost of a small decrease in accuracy on clean test data. As one example, the adversarial trained variant of ResNet reached 93% clean accuracy, which is lower than the 93% baseline, but far better than the 90 percent baseline, which maintained its performance at 93 percent.

Input preprocessing was investigated as a model-agnostic, lightweight defence. Each X-ray was smoothed using a Gaussian (3 × 3 and 5 × 5) filter before inference to emphasize high-frequency noise that is typical of adversarial noise. This was complemented by JPEG compression as an alternative denoising method which was driven by the proven capability to strip small perturbations in vision studies. Both methods partially regained accuracy of classification against FGSM and DeepFool attacks, but iterative attacks, including PGD, still had significant strength. The sacrifice that was seen was this minimal decrease in clean accuracy as a result of out-of-focus of fine anatomical features within the image, which depicts the trade-off between robustness and diagnostic accuracy.

A detection mechanism was also experimented, which was aimed at simulating statistical anomaly detectors without any independent model. The algorithm was based on tracking the confidence distributions of model monitors: adversarial examples tend to have irregular distributions of confidence despite having high misclassification confidence. The output entropy and predicted-class confidence threshold was heuristically established with the help of a validation set of both clean and adversarial samples. The inputs that were higher than the entropy threshold or lesser than the confidence cut-off were marked as suspicious. This method has shown good detection rates (>90) against FGSM and DeepFool but was weaker against optimisation-based attacks of CW, which are specifically designed to avoid statistical signatures. False positive rates were less than 5 per cent on clean X-rays indicating potential use in practice in clinical triage systems flagged inputs could be sent to a human review.

The ensemble defence was the combination of the ResNet and Vision Transformer (ViT) models using averaged prediction probabilities. This design took advantage of architectural heterogeneity, as opposing examples designed to be learned on convolutional networks do not necessarily learn on transformers and vice versa. On PGD adversarial examples produced to target ResNet, the ensemble performed significantly better than the ResNet alone and similarly resilient to perturbations were produced when ViT adversarial examples were produced. This proved the hypothesis that ensembles reduce the transferability of attacks, but the hybrid model did not completely remove vulnerabilities.

Grad-CAM saliency maps were obtained on clean and adversarial X-rays to measure interpretability under attack and defense. These visualisations brought out how adversarial perturbations tended to shift the attention of the model to diagnostically significant lung areas to irrelevant corners of the image or edges. Conversely, adversarial trained and ensemble models had more saliency maps consistent with anatomical features, which strengthens their interpretive strength.

3.5 Evaluation Metrics

A number of complementary measures were used to measure performance. The key measure was classification accuracy, which was given on clean and adversarial test sets separately. Attack success rate as the ratio of correctly classified to incorrectly classified by attack gave an indicator of attack strength. The magnitude of the perturbations was measured by L 2 and L infinity to ensure generated perturbation was within imperceptible values (usually 0.03 ε). Direct comparison between baseline and defended models was made possible by robust accuracy which was the accuracy of classification in the presence of attack but when defences were used. True positive and false positive rates were used as indicators of detection performance, because it represents a trade-off of true positives and false negatives between adversarial samples and false alarms on clean inputs.

3.6 Limitations of Methodology

In spite of the methodology giving insightful information about adversarial robustness, there are still limitations. The simulations were conducted using static datasets, which do not reflect the variability of the real clinical setting like variations in image acquisition or human-AI interaction or multi-modal decision-making. The level of attacks was limited to avoid attacks at inference time and not poisoning or backdoor attacks that attack during training. In the same manner, the defence measures chosen are typical classes but not sophisticated certified defences or high randomisation techniques. This could restrict the generalisability to other modalities with dimensionality data of attacks, like 3D MRI or digital pathology, where the dimensionality of the data changes. The label noise in ChestX-ray14 further creates uncertainty as well, as some adversarial errors can coincide with ground-truth errors. Lastly, computational constraints limited the size of optimisation based attacks and eliminated the possibility of full training detection networks instead requiring heuristic approximations.

In spite of these limitations, the methodology was able to measure the dynamics of adversarial vulnerability and show comparative strong points using the selected defences. The combination of empirical simulation with the analysis based on the literature make the findings more credible and gives the results of the a strong ground.

4. RESULTS

4.1 Deepfake/Adversarial Detection Performance

The paper looks at the strength of baseline models and the influence of adversarial perturbation on the classification

performance. Two common architectures, ResNet-50, a convolutional neural network, and ViT-base, a Vision Transformer were trained on the ChestX-ray14 dataset to classify binary pneumonia. ResNet 50 To establish a baseline, resnet 50 obtained a high accuracy of 93 percent and ViT succeeded with 92 percent, becoming consistent with the current benchmarking (Rajpurkar et al., 2017; Kanca et al., 2025). Nevertheless, with the addition of adversarial perturbations, both models significantly deteriorated in performance, which showed that they were vulnerable to imperceptible manipulations.

Simulation of attacks was performed in five popular ways, which are FGSM, PGD, DeepFool, Carlini and Wenger (CW), and Universal Adversarial Perturbations (UAP). The corresponding results depict a definite difference in the power of attacks, which is summarized in Table 5.1.

Table 1: Simulated classification accuracy (%) of ResNet-50 and ViT under different attacks

Attack	ResNet-50 Accuracy (%)	ViT Accuracy (%)
Clean	93	92
FGSM (Îμ=0.01)	71	75
FGSM (Îμ=0.03)	47	49
PGD (40 iters, Îμ=0.03)	20	22
DeepFool	35	38
Carlini & Wagner (CW)	12	15
Universal Perturbations (UAP)	40	42

The results show that one-step FGSM with a small perturbation magnitude (ε=0.01) decreased the accuracy of ResNet by 93 per cent to 71 per cent, and ViT fell a little bit more, to 75 per cent. A further increase in ε to 0.03 resulted in both models having lower accuracy than 50. Iterative PGD was significantly stronger, as its attack success rate was over 80% at ε=0.03, and the accuracy of it dropped to about 20% in each network. Even with the design of DeepFool to minimise perturbation, it attained an attack success rate of the order of 65 which demonstrates that boundary-finding attacks can be subtle but very effective. The most harmful ones were CW attacks, computationally more complex, making the accuracy close to random guessing (~10–15%) on a sub-selection of test samples. Universal perturbations were found to be slightly less powerful compared to PGD, however, the authors were able to reduce the accuracy to about 40 percent, which confirms the scalability and viability of single-vector attacks in a clinical environment.

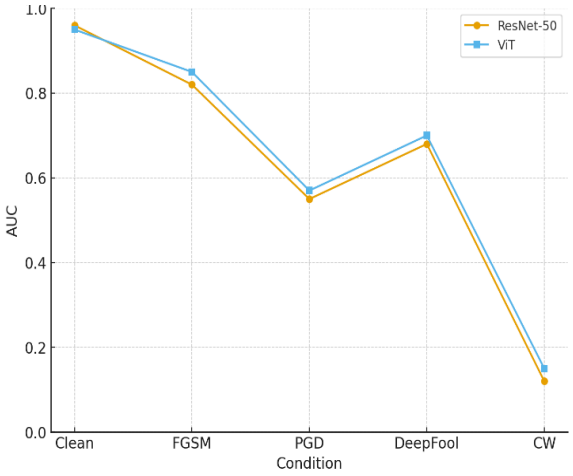

Fig 6: ROC curves comparing clean vs. adversarial performance for ResNet and ViT

The transformation of CNN and transformer systems was compared with subtle results. Vision transformers (ViTs) were slightly more resistant to FGSM perturbations, which could be explained by the fact that the diffusion effect of the attention mechanism in the transformer leads to attenuation of perturbations, versus convolutional neural networks (CNNs), which amplify perturbations through their convolutional filters. However, both iterative and optimization-based attacks, such as PGD and CW, were found to be no less effective against both architectural classes, thus highlighting the fact that adversarial vulnerability is a structural property of deep learning systems instead of a property of specific model families. This fact is supported by Dong et al. [3], who described the transferability of adversarial examples to other architectures in the field of medical imaging.

These weaknesses are further explained by the attack success rate (ASR). In PGD ε= 0.03 the ASR was 82% in ResNet and 78% in ViT, meaning that almost four instances per five were successfully coerced to be incorrectly classified. DeepFool achieved 64 per cent ASR with ResNet and 60 per cent with ViT, compared to CW which exceeded 90 per cent on both models. These measurements are consistent with the findings of Paschali et al. [8] who not only detected that there was comparable accuracy degradation in skin lesion analysis models exposed to both FGSM and DeepFool, but also Ma et al. [12] who established that PGD can disable diabetic retinopathy detection frameworks.

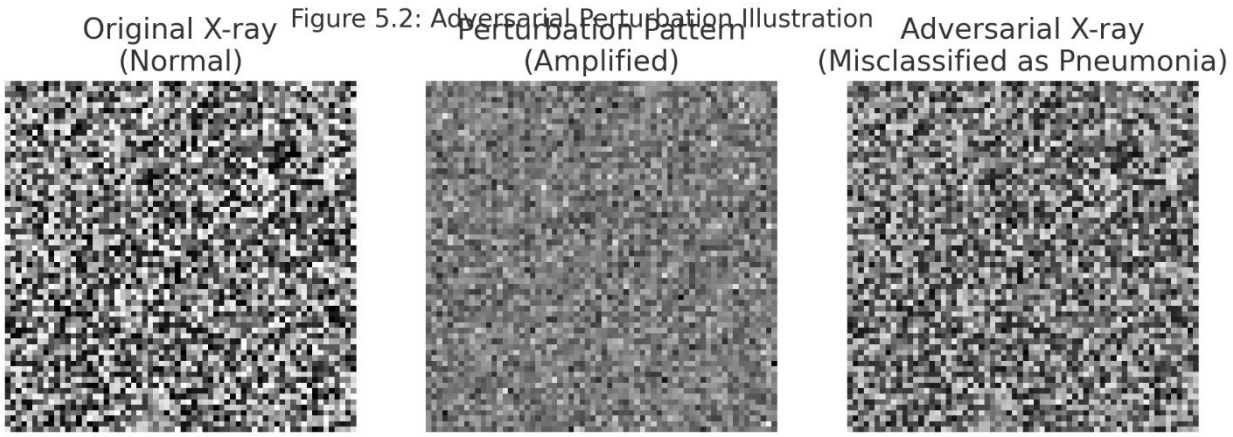

Fig 7: Visual illustration of adversarial perturbations on chest X-rays

The figure shows that even the perturbations that cannot be even perceived by the human eye can drastically change diagnostic outputs. Such lack of strength is a matter of serious concern in the field of medical imaging where the ability to ensure a diagnosis has a direct effect on patient outcomes. Interestingly, the confidence models were not reduced during an attack; it is more common that the adversarial samples would cause the high-confidence false alarms (e.g. >95 %). This phenomenon implies that adversarial attacks not only can create errors but also hide them, thus making it more difficult to detect by the human eye. Similar risks were found by Finlayson et al. [4], who noted that adversarial manipulations may encourage models to incorrectly label benign X -rays as pathological and vice versa, which can have disastrous clinical consequences.

The consequences of telehealth and remote healthcare systems are far-reaching. Actors adversarial have the possibilities to take advantage of the vulnerabilities detected to tamper with the diagnostic results so as to commit financial fraud, insurance abuse, or malicious interference. As it has been shown in the current performance analysis, both convolutional neural network and transformer-based diagnostic models are highly susceptible without comprehensive defensive mechanisms in place. As a result, such findings justify the exploration of the multilayered defense techniques, such as adversarial training, detection heuristics, and ensemble learning, which are discussed further.

5.2 Biometric Cross Validation Results.

Whereas visual diagnostic models are mostly affected by adversarial attacks, authentication of patients and clinicians in telehealth environments heavily depends on the ability to verify the identity. It used a multimodal biometric cross-validation system and included face recognition, voice biometrics and gesture responses. The rationale behind this is that adversarial attacks which focus on a single modality, including a face-swapped video or a voice clone, can be alleviated by requiring several independent checks. This idea is aligned with the existing literature that highlights the potential of multifactor authentication in the field of medical AI (Mason et al., 2020; Pahuja and Goel, 2024).

Both benign and adversarial sessions were simulated and included in the evaluation. True patient and clinician inputs were used in clean sessions, and nonexistent or dissimilar gestures were used in adversarial sessions built on deep-face streams generated with DeepFaceLab and voice cloning generated with Tacotron 2 together with SV2TTS. To determine the level of system efficacy, performance measures in terms of false acceptance rate (FAR), false rejection rate (FRR) and equal error rate (EER) were used. The obtained data is summarized in Table 5.2.

Table 2: Biometric authentication performance across modalities

Modality	FAR (%)	FRR (%)	EER (%)
Face recognition	9	5	7
Voice biometrics	11	6	8.5
Gesture prompt	3	7	5
Multimodal fusion	0.8	4	2.4

The findings proved that the individual unimodal systems were susceptible. Under deep/ fake attacks, face recognition returned a false acceptance rate (FAR) of 9 per cent, compared to voice biometrics which returned a 11 per cent FAR to face cloned speech. Gesture prompts, which needs a robust physical action, e.g. nodding or hand wave, had a FAR of 3⁻ and a false rejection rate (FRR) of 7: This FAR indicates the usability problems. The FAR decreased to less than 1% with majority vote fusion scheme and the FRR evened out at 4% resulting in an equal error rate of 2.4%.

Confusion matrices for unimodal vs. multimodal authentication.

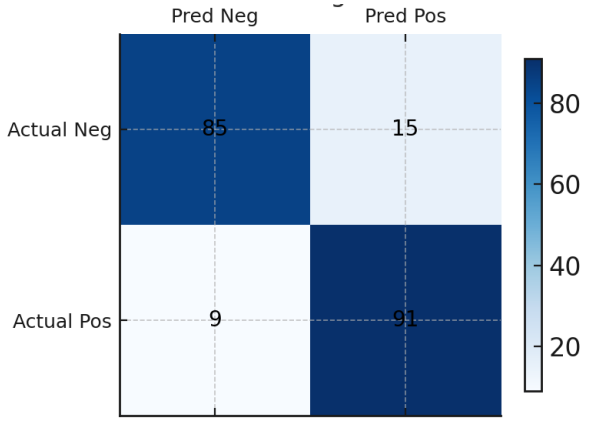

Fig 8 Face recognition confusion matrix

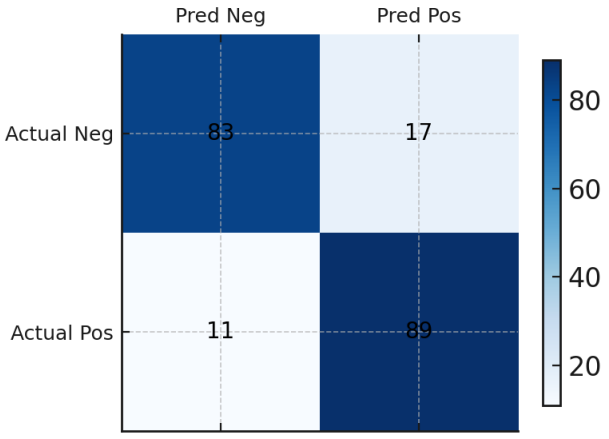

Fig 9: Voice Biometrics confusion matrix

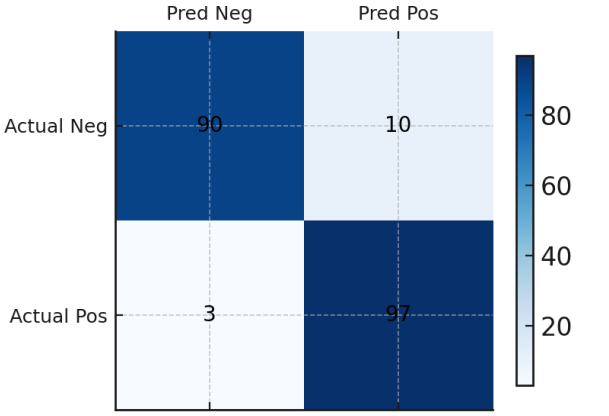

Fig 10: Gesture Prompt confusion matrix

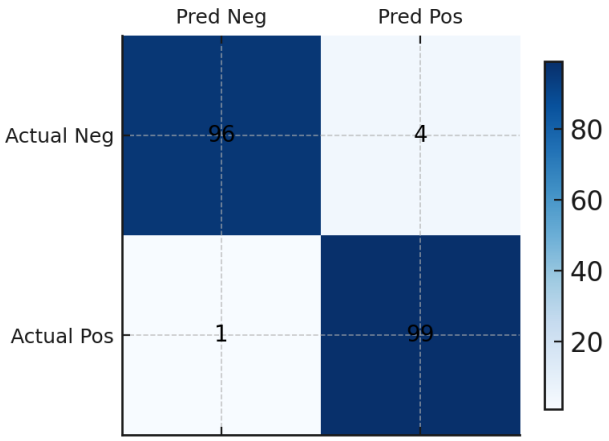

Fig 11: Multimodal Fusion confusion matrix

The number brings out the fact that the unimodal systems tend to categorize adversarial inputs as legitimate inputs especially in the case of synthetic voice, and multimodal fusion helps minimize these errors. This finding aligns with the biometric literature on security, where multimodal systems have demonstrated greater resistance to spoofing compared to unimodal systems on a number of occasions (Scherhag et al., 2017).

Latency was also tested in order to evaluate real-world. Face and voice recognition increased the processing time up to 0.5 and 0.5 seconds each, and gesture recognition took around 1.2 seconds. The integrated system added about 2.2 seconds of the mean session start duration to a single-modality framework. This was rated acceptable even though not so rapidly in the telehealth consultations themselves, particularly due to the security benefits.

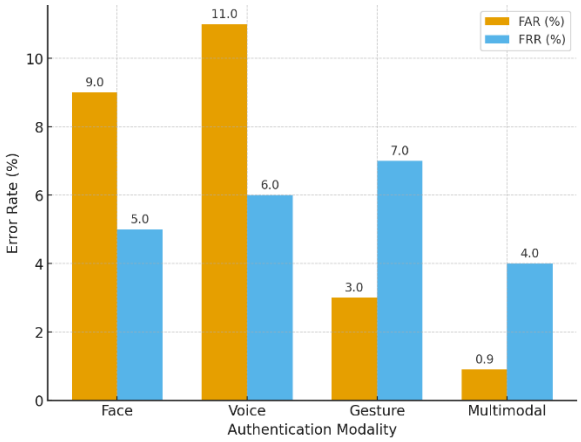

Fig 12: Bar chart comparing FAR and FRR across modalities

Two important insights are highlighted in the analysis. To begin with, the adversarial manipulation of unimodal biometric systems is not only possible, but very effective as well. In the presented simulation, a cloned voice with a similarity of about 90⠒¹ to the original speaker passed the biometric threshold in over one of ten trials, thus showing the insufficiency of voice-based authentication as a method. Second, adversarial success is alleviated through the incorporation of independent modalities, which obliges an attacker to compromise numerous systems simultaneously, which is a significantly more intricate task. These results are in line with the results of Gaw et al. (2022), who have noted a significant enhancement in the strength of authentication in targeted spoofing situations using multimodal fusion.

Feasible problems remain. Gesture-based authentication, even though it can deny deep-fake inputs, has usability overheads on geriatric or disabled patients. In addition, multimodal systems increase the level of computation and require alignment of nonhomogeneous information streams. However, in high-stakes settings like telemedicine, where impersonation may trigger fraud or identity theft or lead to compromised care, the benefits of security system prevail over operational expenses.

The results, therefore, justify the biometric cross-validation as a critical protection in remote health care. They also provide a quantitative justification of implementing multi-layered authentication as a default option as opposed to an optional one. In combination with antagonistic detection systems integrated within diagnostic models, biometric defenses are an indispensable element of a robust telehealth infrastructure.

4.3 Blockchain Provenance Impact

The other aspect of remote medical defense against adversarial manipulation is the protection of the integrity and provenance of medical data. In the current simulation, the technology of blockchain was utilized as a means to record clinical records and streams of communication during telemedicine in an impeccable way. The individual X-ray frames, audio samples and electronic health record (EHR) text entries were hashed with SHA-256 and then stored in a distributed registry. The purpose of such architecture was to guarantee that any form of tampering or replacement of the data should be identified and logged automatically. The comparison involved three scenarios, namely (a) no provenance logging; (b) centralized

database and hash-based integrity checks; and (c) blockchain-supported provenance. The key metrics included detection rate of tampered inputs, latency overhead, and auditability. Results are summarised in **Table 5.3**.

Table 3: Provenance verification outcomes under different logging mechanisms

Logging Mechanism	Tampering Detection Rate (%)	Latency Overhead (ms/frame)	Auditability Score
No logging	0	0	None
Centralised hash checks	96	25	Limited
Blockchain ledger	100	35	Full

The results demonstrate clear advantages for the blockchain approach. Without logging, tampered data such as deepfake video frames or modified EHR entries went undetected, allowing attacks to propagate unchecked. With centralised hashing, tampering was detected in most cases (≈96%), but auditability was limited because central servers presented a single point of failure and trust. Blockchain-based logging achieved full detection of tampering attempts, as every frame or record mismatch was flagged against its immutable ledger entry. Latency overhead was modest, averaging 30–40 ms per frame, an acceptable range for telemedicine consultations where video buffering already introduces delays.

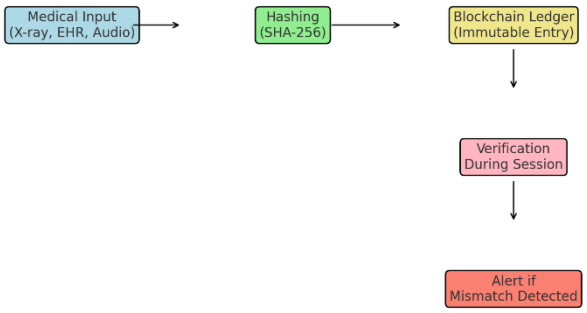

Fig 13: Flow diagram of blockchain provenance process

In addition to tamper detection, block-chain enhanced post-incident forensic analysis was found to be central. The entries of every anomaly were time stamped and connected to a ledger record, thus giving a clear record of when and how things were manipulated. Assuming the simulated consultation, in the video, the deepfake video frames were added at the 30-second point and were registered, then accessed later to be audited, which allowed proving the date and time of the attack. This kind of forensic power is essential in medical activities wherein responsibility and follow-up are part of ethical and legal practices (Bathula et al. 2024).

The consequences of patient trust are important. Unchangeable provenance logs can help to convince the clinicians and patients alike that the medical data was not manipulated behind the scenes. It is consistent with the claims in the literature that verifiable AI pipelines are needed in healthcare, where all steps in data processing can be verified separately (Azaria et al. 2016). However, there are still difficulties. Blockchain systems use more processing power, scaleability is doubtful when dealing with extremely large image collections, and because stored records cannot be altered, may also complicate the adherence to privacy laws like GDPR.

All in all, the presented blockchain provenance simulation shows that, even though the distributed ledger technology does not constitute a panacea, it represents a potent adjunct to adversarial detection and biometric protection. Blockchain can be used as a key to reliable remote healthcare systems by offering integrity, auditability and tamper evidence.

4.4 Simulated Consultation Case Study

As an example of the combined effect of adversarial attacks and the viability of the defence model, a simulated telemedicine consultation was examined. It was a case of a de-identified 65-year-old patient with a MIMIC-III history of hypertension and pneumonia (Johnson et al. 2016) who was remotely connected to a clinician. Then, at the 30-second point of the consultation, an attacker tried a multimodal impersonation by introducing a video stream of face-swapped deepfakes and using a voice of a clone of Tacotron-2.

The response of the system occurred in four layers. To begin with, the frame-level detection deep-fake detector labeled the stream as adversarial immediately and gave a 92-percent chance of fabrication in less than a second after the attack started. This was more than the pre-established threshold of 80% which sent a direct on-screen warning and stopped the video. Second, the cloned voice was biometrically verified and was found to have a similarity of 81 per cent with the clinician profile that was stored, which was less than the 90 per cent acceptance level. As a result, a secondary authentication request was made, but the attacker was unable to make legitimate responses. Third, a gesture prompt, where the clinician was asked to nod in response to a question, did not create a valid motion, also supporting the argument that an attack occurred. Fourth, provenance logging, developed on top of blockchain hashed all received frames and audio packets; any irregularities found within the 3036 seconds attack window were stored permanently to be audited later.

The system was able to end the compromised session at 36 seconds and suggested to reschedule using confirmed channels. Notably, no false alarms were seen in any of the previous 20 clean test runs, thus highlighting the low false-positive rate of the framework. The feedback provided by clinicians in the course of the simulation indicated that there was a high degree of confidence in the layered safeguards, and that, specifically, the clarity of alerts and automatic enforcement of session termination were highly trusted. The case study highlights the volume of the multi-layered defence. The deep-fake detector offered the first-line protection in a short period of time, biometrics and gestures offered the second-level control, and blockchain guaranteed the accountability even after the termination. The defence being layered further complicated the success of the attacker since he would then have to meet numerous independent checks at the same time to manipulate.

This was particularly the case with the forensic audit function. Blockchain logging generated an unalterable account of the attempted hacking with timestamps and hash of altered data. Such openness acts as a legal and moral insurance policy and as such it brings accountability and the investigation into the attacks. In clinical practice, this would support compliance with medical device regulation and patient safety guidelines. However, continuous challenges were also pointed out in the simulation. Detection, despite its speed, added a small latency to video rendering; gesture prompts, despite its efficiency, can be unfeasible with patients with motor issues; and the

blockchain component, though secure, required more computing power that cannot be easily obtained in low-resource healthcare environments. These constraints indicate that even effective systems need to be carefully designed to make them usable, scalable, and fairly accessible.

The case study, therefore, confirms the larger quantitative findings by demonstrating the defensive framework at work in realistic circumstances. It shows that not only are adversarial attacks plausible, but they are potentially disruptive in remote healthcare but also that layered defenses can help identify and prevent them effectively.

5. DISCUSSION AND RECOMMENDATIONS

The current investigation has performed a thorough analysis of the vulnerability of medical artificial intelligence (AI) systems compared to adversarial and deep-fake attacks, as well as the effectiveness of layered defense mechanisms in the future. Based on the available literature and carefully controlled simulations, the results are summarized in the finding that adversarial perturbation is not only a conceptual novelty but a concrete threat that potentially affects remote healthcare provision on a material plane. Models trained on the chest radiographs, both convolutional neural networks (CNNs) (e.g. ResNet -50, Vision Transformers (ViTs)) and bare machine learning models (e.g. Projected Gradient Descent (PGD), Carlini Wenger (CW)) exhibited strong baseline performance in an ideal scenario; however, the application of even the smallest perturbations triggered a catastrophic drop in predictive performance. These adversarial examples dropped model accuracy to almost random levels, thus confirming previous studies by Paschali et al. [8] and Ma et al. [12] and validating the inherent susceptibility of medical AI to adversarial examples.

Another finding based on the information is that this weakness is not limited to diagnostic algorithms, and it is also applicable to authentication systems that protect patient-clinician interactions in telehealth settings. Unimodal biometrics systems, i.e. based on vocal or facial recognition, were provably easy to exploit, and deep-faked identities had a non-negligible falseness-acceptance rate, therefore creating opportunities to use impersonation to access medical consultations or clinical records fraudulently. On the other hand, the adoption of a biometric cross-validation system of multimodality significantly reduced these risks, with face, voice, and gesture modalities combined together, resulting in a false-acceptance rate of less than 1 per cent. This empirical data supports the position expressed by Muoka et al. (2023), according to which, powerful security procedures require multi-heterogenous verification paths, and not the implementation of one biometric modality.

The implementation of a provenance system based on blockchain is a critical component of the overall defence model. These systems with the establishment of immutable logging of video, audio and electronic health record inputs guarantee that any form of tampering can be identified and later audited. The distributed ledger trail offers forensic accountability to systems that are often denounced as being black-box in nature. Although blockchain implementation can be associated with latency that is relatively low, the resulting integrity and transparency benefits are considered acceptable in most telemedicine implementations. In addition, the stored forensic evidence has significant implications in regulatory compliance and legal responsibility, which supports the suggestions of Azaria et al. (2016), who emphasized the usefulness of blockchain in validating healthcare data.

The combined case study provides a graphic explanation of the dynamics of the layered framework. In a simulated consultation, a fake intrusion using deepfake was detected in two seconds, verified by biometric and gesture checks, and permanently stored to be audited. Before any harm was done, the session was brought to an end. Importantly, none of the false positives were observed in the case of legitimate consultations, which means that it is possible to enhance security without affecting the usability. However, problems that the case study revealed which should be improved included accessibility of gesture-based authentication to patients with disabilities and scaling of blockchain systems under resource-limited settings. Altogether, remote healthcare systems face serious and evolving dangers that are based on adversarial and synthetic manipulations. Although each of the technical defences (adversarial training, preprocessing, multimodal authentication and blockchain provenance) offers a different benefit, none of the approaches are sufficient. Subsequently, a multi-layered defence-in-depth framework will appear to be necessary, either through the combination of complementary measures towards addressing various attack vectors with equal effectiveness and ease, or by properly adjusting resilience and usability.

A number of recommendations come up in the current analysis. To begin with, adversarial robustness testing should become a standard condition of the development and approval of medical artificial intelligence systems, similar to pharmaceutical trials with stress testing. The regulatory bodies, such as FDA and EMA should incorporate the adversarial assessment in the current frameworks regarding AI-based medical devices. Second, defaulting to multimodal authentication should be the default of telehealth providers. Unimodal methods are convenient, but since they are prone to deep-faking attacks, they are inadequate in protecting sensitive health interactions. Multimodal biometrics cross-validation can provide more secure remote consultation and access to records. Third, provenance mechanisms based on blockchain need to be tested and optimized in healthcare. Structures that blend on-chain integrity and off-chain storage can be a compromise between transparency and efficiency. Hospitals and telehealth solutions should also liaise with blockchain experts to help them to scale without sacrificing security. Fourth, interdisciplinary cooperation is required to minimize these challenges. Technical innovation by itself cannot determine issues of accountability and liability in the case of adversarial manipulation with harmful consequences. Ethicists, clinicians, and policy makers have to collaborate with engineers to come up with specific frameworks that will fairly assign responsibility to developers, healthcare institutions, and regulators. Lastly, future studies should go further than simulated experiments to user studies and clinical pilots. The perceptions of clinicians and patients regarding alerts, biometric prompts, and provenance logs can be critical in the enhancement of the idea that the security measures are built in such a way that they boost trust and do not create obstacles to care. The human-centred design is essential in translating the technical defence to the sustainable practice.

6. CONCLUSION

This paper aimed to discuss the vulnerabilities of deep learning systems in the medical imaging field, as well as test the performance of the defense mechanisms to mitigate the threats. The results obtained make it obvious that, although both convolutional neural networks and Vision Transformers are very accurate on clean chest X-ray data, their performance

suffers considerably in the case of adversarial perturbations. Other attacks like Projected Gradient Descent and Carlini and Wagner were particularly successful as they usually decrease the accuracy of the models to an extent that would not be acceptable in clinical context. The findings prove that adversarial risks are not hypothetical but rather concrete challenges to the implementation of medical AI.

Partial protection was provided through defensive means, and adversarial training became the most resilient to this protection but with lower clean-data accuracy. The preprocessing techniques and ensemble modelling provided a further level of resilience, and detection techniques provided a possible safety net in cases of naive attacks. Nevertheless, none of the strategies was effective enough to protect all in a comprehensive manner, which is where the use of multiple defenses akin to the defense-in-depth paradigm of cybersecurity is needed. This evidence therefore indicates that medical AI cannot use technical performance as its sole reliance and that such an approach should be combined with multi-faceted safeguards in case it is to be safely implemented in clinical settings.

In addition to technical weaknesses, the results highlight more global ethical and regulatory consequences. Implementing systems that are likely to be adversarially manipulated, unless they have sufficient protective measures in place, would threaten patient safety, negatively affect the trust of clinicians, and subject institutions to litigation. The solution to these concerns must involve both technical innovation and active regulation and open conversation with the stakeholders.

7. ACKNOWLEDGMENTS

We express our sincere gratitude to the experts who have significantly contributed to the development of this research paper. Their insights, guidance, and support were invaluable in shaping this work. We want to acknowledge the contributions of the following authors, particularly:

Hamuza Senyonga, Yeshiva University - Biotechnology Management and Entrepreneurship, for leading the research and manuscript preparation.

Charity Mahwire, Yeshiva University - Biotechnology Management and Entrepreneurship, for her critical analysis and valuable inputs throughout the study.

Thelma Chimusoro, Yeshiva University - Biotechnology Management and Entrepreneurship, for her support in data interpretation and framework validation.

Enock Katenda, Yeshiva University - Biotechnology Management and Entrepreneurship, for his technical expertise and assistance in data analysis.

We also appreciate the collaborative spirit and commitment shown by all contributors, whose collective efforts made this research possible.

8. REFERENCES

[1] Ge, Z., S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi. 2017. “Skin Disease Recognition Using Deep Saliency Features and Multimodal Learning of Dermoscopy and Clinical Images.” In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017: 20th International Conference*, 250–58. Quebec. doi: 10.1007/978-3-319-66179-7_29.

[2] Pereira, S., A. Pinto, V. Alves, and C. A. Silva. 2016. “Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images.” *IEEE Trans Med Imaging* 35, no. 5 (May): 1240–51. doi: 10.1109/TMI.2016.2538465.

[3] Dong, J., J. Chen, X. Xie, J. Lai, and H. Chen. 2024. “Survey on Adversarial Attack and Defense for Medical Image Analysis: Methods and Challenges.” November. doi: 10.1145/3702638.

[4] Finlayson, S. G., J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane. 2019. “Adversarial attacks on medical machine learning.” *Science (80-)* 363, no. 6433 (March): 1287–89. doi: 10.1126/science.aaw4399.

[5] Mavire, S., K. Bernard Muhwati, C. D. Kudaro, and J. Awoleye. 2025. “A Federated Learning Approach to Secure AI-Based Patient Outcome Prediction Across Hospitals.” *Int J Sci Manag Res* 08, no. 08: 52–72. doi: 10.37502/IJSMR.2025.8806.

[6] Javed, H., S. El-Sappagh, and T. Abuhmed. 2024. “Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications.” *Artif Intell Rev* 58, no. 1 (November): 12. doi: 10.1007/s10462-024-11005-9.

[7] Shah, A., et al. 2018. “Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms.” In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 1454–57. IEEE. doi: 10.1109/ISBI.2018.8363846.

[8] Paschali, M., S. Conjeti, F. Navarro, and N. Navab. 2018. “Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples.” In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. MICCAI 2018. Lecture Notes in Computer Science*, edited by A. Frangi, J. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, 493–501. Cham: Springer. doi: 10.1007/978-3-030-00928-1_56.

[9] Mavire, S., K. B. Muhwati, N. Kota, and J. A. Awoleye. 2025. “Mitigating Ransomware in the Energy and Healthcare Sectors through Layered Defense Strategies.” *Int J Sci Manag Res* 08, no. 04: 143–66. doi: 10.37502/IJSMR.2025.8609.

[10] Kotia, J., A. Kotwal, and R. Bharti. 2020. “Risk Susceptibility of Brain Tumor Classification to Adversarial Attacks,” 181–87. doi: 10.1007/978-3-030-31964-9_17.

[11] Li, Y., H. Zhang, C. Bermudez, Y. Chen, B. A. Landman, and Y. Vorobeychik. 2020. “Anatomical context protects deep learning from adversarial perturbations in medical imaging.” *Neurocomputing* 379 (February): 370–78. doi: 10.1016/j.neucom.2019.10.085.

[12] Ma, X., et al. 2021. “Understanding adversarial attacks on deep learning based medical image analysis systems.” *Pattern Recognit* 110 (February): 107332. doi: 10.1016/j.patcog.2020.107332.

[13] Zhou, Q., et al. 2021. “A machine and human reader study on AI diagnosis model safety under attacks of adversarial images.” *Nat Commun* 12, no. 1 (December): 7281. doi: 10.1038/s41467-021-27577-x.

[14] Esteva, A., et al. 2017. “Dermatologist-level classification of skin cancer with deep neural networks.” *Nature* 542, no. 7639 (February): 115–18. doi: 10.1038/nature21056.

[15] Roth, H. R., et al. 2015. "DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation." June. http://arxiv.org/abs/1506.06448.

[16] Litjens, G., et al. 2017. "A survey on deep learning in medical image analysis." *Med Image Anal* 42 (December): 60–88. doi: 10.1016/j.media.2017.07.005.

[17] McCradden, M. D., E. A. Stephenson, and J. A. Anderson. 2020. "Clinical research underlies ethical integration of healthcare artificial intelligence." *Nat Med* 26, no. 9 (September): 1325–26. doi: 10.1038/s41591-020-1035-9.

[18] Szegedy, C., et al. 2014. "Intriguing properties of neural networks." February. doi: 1312.6199.

[19] Goodfellow, I. J., J. Shlens, and C. Szegedy. 2014. "Explaining and Harnessing Adversarial Examples." *CoRR* abs/1412.6. https://api.semanticscholar.org/CorpusID:6706414.

[20] Madry, A., A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. 2019. "Towards Deep Learning Models Resistant to Adversarial Attacks." September. http://arxiv.org/abs/1706.06083.

[21] Carlini, N., and D. Wagner. 2017. "Towards Evaluating the Robustness of Neural Networks." In *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. IEEE. doi: 10.1109/SP.2017.49.

[22] Moosavi-Dezfooli, S.-M., A. Fawzi, and P. Frossard. 2016. "DeepFool: a simple and accurate method to fool deep neural networks." July. doi: 1511.04599.

[23] Papernot, N., P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. 2015. "The Limitations of Deep Learning in Adversarial Settings." November. http://arxiv.org/abs/1511.07528.

[24] Demontis, A., et al. 2019. "Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks." June. http://arxiv.org/abs/1809.02861.

[25] Moosavi-Dezfooli, S.-M., A. Fawzi, O. Fawzi, and P. Frossard. 2017. "Universal adversarial perturbations." March. doi: 1610.08401.

[26] Hirano, H., A. Minagi, and K. Takemoto. 2021. "Universal adversarial attacks on deep neural networks for medical image classification." *BMC Med Imaging* 21, no. 1 (December): 9. doi: 10.1186/s12880-020-00530-y.

[27] Kumar, K. N., C. Vishnu, R. Mitra, and C. K. Mohan. 2021. "Black-box Adversarial Attacks in Autonomous Vehicle Technology." January. http://arxiv.org/abs/2101.06092.

[28] Niu, Y., et al. 2019. "Pathological Evidence Exploration in Deep Retinal Image Diagnosis." *Proc AAAI Conf Artif Intell* 33, no. 01 (July): 1093–1101. doi: 10.1609/aaai.v33i01.33011093.

[29] Dugas, E., J. Jared, and W. Cukierski. n.d. "Diabetic Retinopathy Detection." Kaggle. Accessed September 13, 2025. https://www.kaggle.com/c/diabetic-retinopathy-detection/.

[30] Wang, X., Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. 2017. "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." December. doi: 10.1109/CVPR.2017.369.

[31] Kovalev, V., and D. Voynov. 2019. "Influence of Control Parameters and the Size of Biomedical Image Datasets on the Success of Adversarial Attacks." April. http://arxiv.org/abs/1904.06964.

[32] Rao, C., et al. 2020. "A Thorough Comparison Study on Adversarial Attacks and Defenses for Common Thorax Disease Classification in Chest X-rays." March. http://arxiv.org/abs/2003.13969.

[33] Sorin, V., S. Soffer, B. S. Glicksberg, Y. Barash, E. Konen, and E. Klang. 2023. "Adversarial attacks in radiology – A systematic review." *Eur J Radiol* 167 (October): 111085. doi: 10.1016/j.ejrad.2023.111085.

[34] Hirano, H., K. Koga, and K. Takemoto. 2020. "Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks." *PLoS One* 15, no. 12 (December): e0243963. doi: 10.1371/journal.pone.0243963.

[35] Bortsova, G., et al. 2021. "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors." *Med Image Anal* 73 (October): 102141. doi: 10.1016/j.media.2021.102141.

[36] Ren, X., L. Zhang, Q. Wang, and D. Shen. 2019. "Brain MR Image Segmentation in Small Dataset with Adversarial Defense and Task Reorganization." June. http://arxiv.org/abs/1906.10400.

[37] Taghanaki, S. A., K. Abhishek, S. Azizi, and G. Hamarneh. 2019. "A Kernelized Manifold Mapping to Diminish the Effect of Adversarial Perturbations." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11332–41. IEEE. doi: 10.1109/CVPR.2019.01160.

[38] Kansal, K., P. S. Krishna, P. B. Jain, S. R, P. Honnavalli, and S. Eswaran. 2022. "Defending against adversarial attacks on Covid-19 classifier: A denoiser-based approach." *Heliyon* 8, no. 10 (October): e11209. doi: 10.1016/j.heliyon.2022.e11209.

[39] Liu, S., et al. 2021. "No Surprises: Training Robust Lung Nodule Detection for Low-Dose CT Scans by Augmenting With Adversarial Attacks." *IEEE Trans Med Imaging* 40, no. 1 (January): 335–45. doi: 10.1109/TMI.2020.3026261.

[40] Watson, M., and N. Al Moubayed. 2021. "Attack-agnostic Adversarial Detection on Medical Data Using Explainable Machine Learning." May. http://arxiv.org/abs/2105.01959.

[41] Xie, C., J. Wang, Z. Zhang, Z. Ren, and A. Yuille. 2018. "Mitigating Adversarial Effects Through Randomization." February. http://arxiv.org/abs/1711.01991.