Adversarial Threats in Healthcare: A Comprehensive Analysis of Vulnerabilities, Defense Mechanisms, and Recent Research

Atrab A. Abd El-Aziz[1], Reda A El-Khoribi[2], and Nour Eldeen Khalifa[2]*

[1]Department of Information Technology, Faculty of Computers and Information, KafrelSheikh University, Egypt
[2]Department of Information Technology, Faculty of Computers and Artificial Intelligence, Cairo University, Egypt

Abstract
Recently, medical adversarial attacks had grave repercussions for both patients and healthcare providers. The resistance of deep learning-based medical models to adversarial attacks in dynamic situations is a topic of great interest since they are easily fooled using simple imperceptible perturbations called adversarial attacks. These attacks can exploit vulnerabilities in medical systems and can take various forms such as manipulating medical images, Medical Question Answering (MQA), or compromising electronic health records. These attacks resulted from some limitations in medical images such as the scarcity of high-quality imaging datasets and image labeling. Medical systems are very challenging due to their safety and accuracy are vital considerations. Although several defense techniques have been introduced to defend against these attacks, there is still a significant demand for more robust defense solutions. Notably, the domain of MQA and medical text is significantly under-researched, with virtually no established defense methods to counter adversarial threats in these areas. This paper contributes to the field by bridging the gap between adversarial attacks in medical images and medical text special MQA. It ensures that the review offers valuable insights and practical recommendations for researchers and practitioners in healthcare and cybersecurity. This paper covers various aspects of medical adversarial attack techniques, sources of vulnerabilities, defense strategies, and provides insights into the motivations and tactics of medical attacks. Finally, we provide recommendations for enhancing the resistance of medical systems to adversarial attacks. This review serves as a valuable resource for a diverse audience includes researchers and practitioners for healthcare and cybersecurity providers. It contributes to the collective knowledge base and offers practical recommendations for safeguarding medical systems against adversarial threats, particularly in the understudied areas of MQA and medical text.

Keywords
Adversarial Attacks, Deep Learning, Defense Methods, Medical Images, Medical Question Answering (MQA), Security.

1. Introduction

There has been a notable surge of interest regarding the application of artificial intelligence (AI) technology in the field of medical imaging for the purposes of disease diagnosis. A branch of AI known as deep learning (DL) enables computational models to learn from experience. DL is composed of many processing layers and uses the backpropagation technique to derive the data representation of each layer from the representation in the preceding layer. Convolutional neural network (CNN) is a DL technique used for the processing of image and analysis [1].

DL has gained popularity for medical imaging tasks such as diabetic identification and cancer diagnosis due to its current success in natural imaging. Medical images are frequently used in patient care as prognostic and diagnostic imaging modalities [2]. Manual evaluations of medical images by human professionals are labor-intensive in practice. Therefore, automated DL systems have been suggested to aid in clinical decision-making and assist with human diagnosis [3]. Recently, several DL models have demonstrated performance better than the human specialist level in a variety of medical imaging jobs [4].

Recent research has shown that despite the increased performance of DL techniques, they are susceptible to a new type of attack called adversary attacks. The vulnerability of high-performance DL techniques to adversarial attacks was initially identified by Szegedy et al. [5]. DL models can be tricked into making inaccurate predictions with a high degree of confidence if their input examples are slightly perturbated [6]. The explanation, this type of attack was caused by the high nonlinearity of DL networks coupled with the regularization of the supervised learning techniques. Medical DL classification and segmentation models have proven susceptible to rapidly enhanced and updated adversarial attacks due to the features of medical images and the sparse number of annotated images. Many techniques are developed to generate these attacks such as Carlini and Wagner (C&W) [7], projected gradient descent (PGD) [8], and fast gradient sign method (FGSM) [9].

DL models have been implemented in safety-critical applications which raise safety concerns. A variety of detection and mitigation strategies have been offered to defend against these attacks. Adversarial training is among the most frequently defended strategies. By augmenting the training dataset by adding adversarial samples. This technique increases the robustness of neural networks against these attacks. It has been shown that one of the main causes of these limitations is the scarcity of a sufficient number of high-quality tagged photos and data [10].

Manuscript received April 5, 2026
Manuscript revised April 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.4.17

Main contributions of this paper:

1. **Comprehensive Coverage:** This paper is the first to examine adversarial attacks and defense mechanisms across both medical imaging and medical text. It provides a holistic understanding of the vulnerabilities present in DL-based medical systems.
2. **Addressing Research Gaps:** The paper identifies and discusses the significant research gaps particularly in the defenses for MQA systems. It highlights the urgent need for robust strategies in this area.
3. **Holistic Analysis:** It offers an in-depth exploration of the goals and motivations of adversarial attackers. It shades light on the various tactics employed against medical systems.
4. **Taxonomy and Variability:** The paper presents a detailed overview of the different branches of adversarial attacks. It includes their taxonomy, current examples, and implications for both medical imaging and text.
5. **Visual Representation of Trends:** This paper provides valuable insights into trends and advancements in the field. It provides an analyses and visualization for the state of current research publications.
6. **Future Research Directions:** The paper outlines potential avenues for future research. It includes exploring more defense and detection techniques specifically tailored for medical text.
7. **Addressing Limitations:** It critically examines the limitations of existing approaches and emphasizes the need for collaborative efforts to enhance the resilience of medical systems against adversarial threats.

In this paper, a complete review of the recently developed adversary attacks and defensive techniques on DL-based medical system diagnosis is introduced. Section 2 explains the motivation to investigate this paper. Section 3 discusses the research gap concerning the defenses for MQA systems and highlights the lack of existing strategies. Section 4 examines the goals of adversarial attackers by providing insight into their underlying motivations and tactics. Section 5 offers an overview of the various branches of adversarial attacks by focusing on their application in medical systems. This section delves into the variability, taxonomy, and current examples of attacks as well as the existing defense and detection strategies for DL-based medical imaging and text. Section 6 analyzes and visualizes the current state of research publications in this field by providing a clear representation of trends and advancements. Section 7 explores potential future research directions and outlines the limitations of current approaches. Finally, Section 8 concludes the paper, summarizing the key findings and implications of the paper.

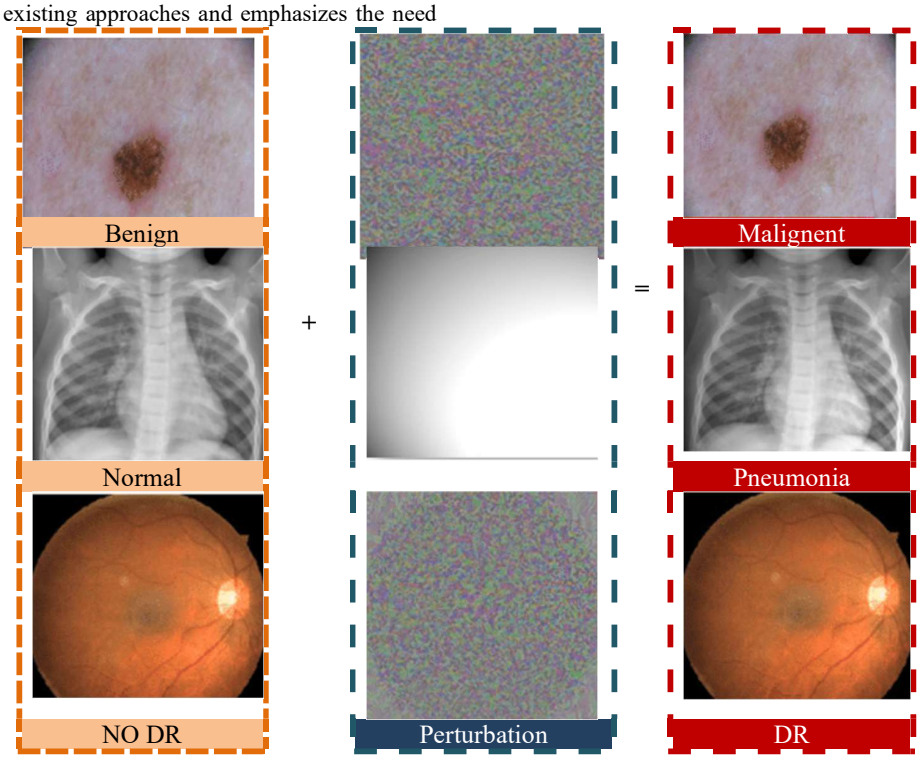

Figure 1. Example Of Adversarial Attacks Developed to Fool Trained Medical Models.

2. Research Motivations

According to a recent industry report reported by Cearley [11], 30% of all attacks on artificial intelligence (AI) systems would use adversarial attacks, poisoning of training data, or stolen models. Given that, the healthcare sector is expected to experience two to three times more cyberattacks than the other sectors. Therefore, adversarial attacks could be a serious threat to the medical industry. Adversarial attacks can lead to fraudulent behavior in fee-for-service healthcare systems such as healthcare professionals and insurance companies falsifying patient diagnosis codes to influence diagnostic decisions. The security of the medical-based deep learning models against adversarial attacks is an important open research area. According to Figure 1, the main objective of the attacker is to significantly reduce the performance of the medical images system and make changes invisible to the human eye [12].

The motivations behind adversarial attacks on MQA systems are varied but generally align with the broader goals of malicious exploitation in healthcare settings. One primary motivation could be to create miscommunication between patients and healthcare providers. By manipulating responses in MQA systems, attackers can degrade the quality of care that potentially led to patient mistrust or legal liabilities for healthcare providers. In other cases, adversaries may seek to introduce errors into automated diagnoses to cause reputational damage to a healthcare provider or institution. Financial gain could also be a driver as attackers might exploit system vulnerabilities to disrupt billing systems, alter medical records or commit fraud. Additionally, state-sponsored attacks or hacktivists may target MQA systems to sabotage healthcare operations or compromise sensitive patient information, leveraging these adversarial tactics for political or ideological purposes.

This paper allows researchers to choose their field of study within the field of medical deep learning-based adversarial attack methods.

3. The Research Gap in Defending Medical Question Answering Systems

The field of adversarial defense for medical systems has made significant strides in recent years, particularly in the context of medical imaging. However, one area that remains notably under-researched is adversarial defense in MQA systems and medical text-based systems. MQA systems which rely on natural language processing (NLP) models to provide responses to patient and clinician inquiries, are especially vulnerable to adversarial attacks. These attacks can subtly manipulate textual inputs which causing the system to generate inaccurate or even dangerous answers, thereby compromising clinical decision-making and patient safety.

Despite the critical role of MQA systems in clinical environments, there is a conspicuous absence of research focused on defending these systems against adversarial text perturbations. Unlike adversarial attacks on medical images which have been the subject of extensive research and development of defense strategies, the domain of medical text remains largely unprotected. This gap poses a significant risk as text-based adversarial attacks can be as detrimental as image-based ones which leading to misdiagnoses, incorrect treatment recommendations, or compromised patient data.

Current defense mechanisms for general question-answering (QA) systems such as synonym replacement or adversarial training haven't been sufficiently adapted or optimized for the medical domain. Medical text often contains highly specific terminology and complex syntactic structures. This makes it particularly susceptible to adversarial modifications. Moreover, the lack of large, well-labeled medical text datasets hampers the development of robust adversarial defense models, further contributing to the research gap.

The absence of tailored defenses for MQA systems suggests a pressing need for novel approaches that can effectively safeguard these systems. Future research should prioritize the development of advanced defense mechanisms that are specifically designed to address the unique challenges of medical text including word-level and character-level attacks. Additionally, interdisciplinary collaborations between NLP, cybersecurity, and healthcare researchers could drive the innovation required to create more resilient MQA systems.

The limited focus on adversarial defense in medical text-based systems presents a significant vulnerability in healthcare. Addressing this research gap is essential to ensuring the safety, accuracy, and reliability of MQA systems, which are increasingly integrated into clinical practice.

4. Goals of Adversarial Attackers

An attacker attempts to feed the target DL model with an input x* which misleads the result to an incorrect class. The attacker's goal is deduced from the misclassification of the model. The attacker objectives can be categorized based on their influence on the accuracy of the classifier output as shown in Figure 2 [13]:

Manuscript received April 5, 2026
Manuscript revised April 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.4.17

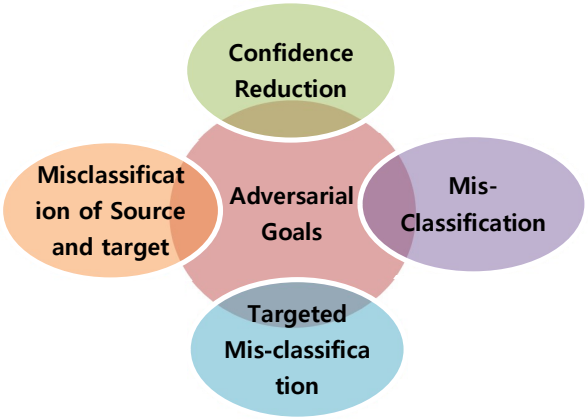

Figure 2. Main Goals of Adversarial Attacks.

1. **Confidence Reduction:** The attacker attempts to decrease the target's confidence model's prediction. For illustration, the "stop" sign image can be predicted with less confidence and a lower probability of belonging to the actual class.
2. **Targeted Mis-classification:** The attacker tries to manipulate the inputs so that the model generates the results of a specific target label. For example, the model of classification can classify that any input belongs to a category of photos with a label "go" sign.
3. **Mis-Classification:** The attacker attempts to modify the output classification of an input instance so that it is assigned to a different class. For example, a real image of represent "stop" image can be classified to any class other than the actual class of the stop sign.
4. **Misclassification of Source and Target:** The attacker attempts to assign a specific input source to a target class that has already been established. For example, the classification model will classify the input sign as "go" while it is actually a "stop" sign.

5. Adversarial Attacks in Medical Systems: Forms and Motivations

Adversarial attacks in medical systems pose significant threats to patient safety and the integrity of healthcare practices. This section provides an overview of the various forms of adversarial attacks, adversarial attacks on medical images and the adversarial attacks on medical text.

5.1 Taxonomy of Image-based Adversarial Attacks

Adversarial samples are inputs that have been generated and modified by adding a small perturbation to the actual input to force the model to classify something incorrectly in DL systems. By creating these samples, the adversaries' primary goal is to significantly alter system performance and make the changes invisible to the human eye. As shown in Figure 3, There are three main classification branches of adversarial attacks [14].

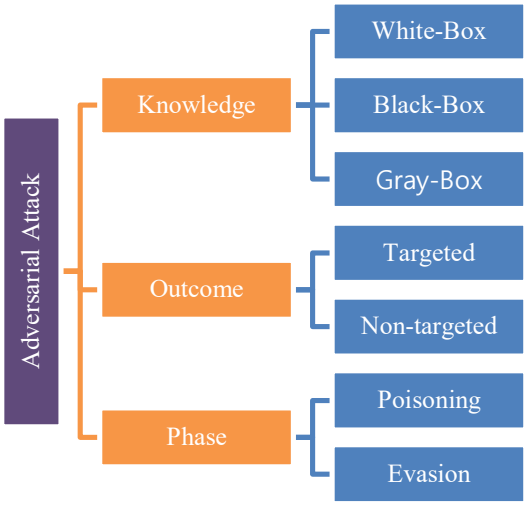

Figure 3. Taxonomy of Adversarial Attacks.

5.1.1 Adversarial knowledge-based Attacks

There are three different forms of attacks that can be used, depending on the attacker's knowledge of the target model [15]. Figure 4 shows the most common techniques applied in each category.

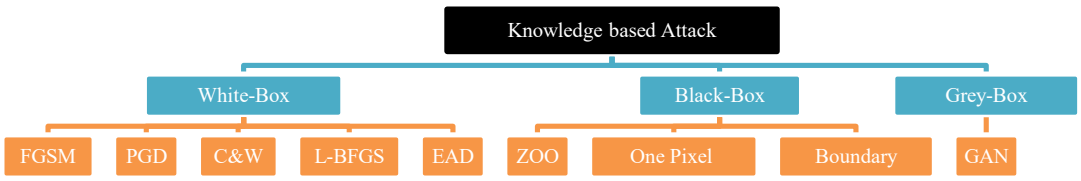

Figure 4. Examples of Knowledge-based Attack Techniques.

1. **White-box attacks:** The attacker understands the target model's parameters and architecture in detail. This is the simplest plan of attack and has the potential to cause the most damage. The attacker may be the model developer or an insider who can consider the worst-case situation of the model being attacked. The most common techniques of this category are the C&W, PGD, and FGSM attacks.

A. Fast Gradient Sign Method (FGSM) Attack

FGSM was first introduced by Goodfellow et al. It is one of the famous white-box, evasion, and a one-step technique. FGSM is used to create adversarial samples by adding a significant amount of perceivable noise (perturbations) in one or more pixels of an image. These perturbations result in model misclassification. This noise is calculated by multiplying the image-related parameters based on gradient signs. Epsilon functions as a minor constant; The model is fooled when epsilon rises. FGSM raises the loss in the most vertical direction by performing the one-step update along the gradient of the opposite loss. Formally, the adversarial sample produced by the untargeted FGSM is formulated by equation 1 while equation 2 serves as the targeted FGSM update rule is as follows:

$$x' = x + \in . sign[\nabla_x J(\theta, x, y)] \quad \text{Eq (1)}$$

$$x' = x - \in . sign[\nabla_x J(\theta, x, y')] \quad \text{Eq (2)}$$

Where J (θ, x, y) represents the gradient loss and $\in$ represents the perturbation magnitude. Targeted FGSM can be extended from the untargeted FGSM by reducing the gradient loss J (θ, x, y). Where y' is the target adversarial sample. If cross-entropy is used as the adversary loss, this update process can reduce the cross-entropy between the target vector and the predicted vector.

B. Projected Gradient Descent (PGD) Attack

PGD is a white-box, and evasion attack that presented by Madry et al. to enhance FGSM's performance. They demonstrate how the performance of FGSM can be improved by repeatedly executing a better iterative optimizer. The BIM trims the modified adversarial sample to a suitable range for T iterations using a reduced step size, with the following update rule at the t_{th} iteration:

$$x'_{t+1} = Clip\{x'_t + \propto . sign[\nabla_x J(\theta, x'_t, y)]\} \quad \text{Eq (3)}$$

Where x'_t is the perturbated sample at iteration t. the iteration number is represented by t, and α is the perturbation magnitude.

C. Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) Attack

L-BFGS is a limited memory, nonlinear gradient-based numerical optimization methodology. L-BFGS white-box attack is used to discover adversarial perturbations with the minimum L_p norm. L-BFGS is used in an effort to lessen the amount of image disturbances. These images mislead the model output. Adversarial samples are produced that are conceptually identical to the original input.

$$\min_x \|x - x'\| p \ \ subject\ to\ f(x') \neq y' \quad \text{Eq (4)}$$

Where the $\|x - x'\|p$ represents the L_p norm perturbations and y' represents the adversarial output. Hybrid loss is minimized by $C\|x - x'\|p + J(\theta, x', y')$ where the hyper parameters are represented by C. Grid search could identify the ideal value of C.

D. Carlini and Wagner (C&W) Attack

C&W is an optimization-based white-box attack that was introduced by Carlini and Wagner. The L-BFGS attack served as the model for this one. It does away with multiple objective functions

and box constraints. Unlike adversarial training and defensive distillation, it more effectively generates adversarial cases to overcome the most recent defenses. It can produce adversarial samples measured according to the norm L0, L2 and L1, or CW0, CW2, CW∞. the optimization objective is formulated as follows:

$$\min_{\delta} D(x, x+\delta) + c.f(x+\delta) subject\ to\ x + \delta \epsilon [0,1] \quad \text{Eq (5)}$$

Where the adversarial perturbation is represented by δ , the L0, L2, or L1 metric is represented by D ($x, x+\delta$), and $f(x+\delta)$ represents the adversarial loss.

E. Elastic-Net Attack (EAD)

EAD was introduced by Chen et al. [16]. It is an adversary technique in which the adversarial sample production process is considered a regularized optimization problem. The overall goal of EAD is to create adversarial samples that can mislead the DL models and cause the least amount of disturbance in terms of L1 and L2 distance metrics. Consequently, this is how the optimization problem is defined:

$$\min_{x'} cJ(\theta, x', y') + \beta \|x' - x\|1 - \|x' - x\|_2^2 \ subject\ to\ x' \in [0,1]^d \quad \text{Eq (6)}$$

Where $J(\theta, x', y')$ is the targeted adversarial loss, β and c are hyperparameters, and $\|x' - x\|1 - \|x' - x\|_2^2$ is used to penalize the L1 and L2 distances between the adversarial samples $x'$and the benign sample x .

2. **Black-box attacks:**

The attacker has no knowledge of the target model's design or parameters, but they can provide input and track model output. This is the most plausible and difficult attack scenario. If two machine learning models are used to solve the same problem, the adversarial example of one machine learning model can successfully fool the other model with a high success rate. Therefore, in black box attacks, the attacker can create adversarial examples from DL models and then attack the original target model by repeatedly querying the target model to create a training set.

Zero-Order Optimization (ZOO) Attack

ZOO is a black-box attack that was first introduced by Chen et al. [17]. ZOO performs an attack against the destination DNN by sending numerous queries and evaluating the output trust values in response. ZOO computes the network gradient through zero-order optimization while improving attack efficiency with hierarchical attack, dimensionality reduction, and importance sampling techniques. ZOO's optimization strategy is comparable to C&W's algorithm; however, because it uses a black box approach, it cannot acquire the gradient of the model. ZOO computes the estimated gradient using the symmetric difference ratio. The best perturbation is created using the ADAM and stochastic coordinate descent approach to increase the efficiency of convergence.

3. **Gray-box attacks:**

this is a partial knowledge attack in which the attacker only knows part of the training data used to modify the target model. In this approach, attackers can compute parameters and scores based on gradients and can only use probability distribution methods to grant access rights to the output labels assigned to them. In most cases, defending against an adversary attack is evaluated in a gray box environment.

Generative Adversarial Networks (GAN)

GAN is an adversarial attack where two different types of neural networks, such as the discriminator and generator are competing together. It was first introduced by Xiao et al. [18] to create adversarial samples. By optimizing both target adversary loss and GAN loss, a generator is specifically trained to learn the adversary's distribution. The generator creates instances that the discriminator classifies incorrectly. The discriminator tries to distinguish between the genuine samples and the fraudulent ones produced by the generator itself in the meantime. Unrestrained adversarial instances are those that are generated because there are no benign samples adjacent to them. This approach doesn't follow the typical constraints set for adversarial samples, making it more successful when attacking models that have been trained through adversarial means using attacks that comply with those constraints.

5.1.2 Adversarial Outcome-based Attacks

Depending on the outcome of the attack, there are two classifications of attacks called targeted and non-targeted attacks [19].

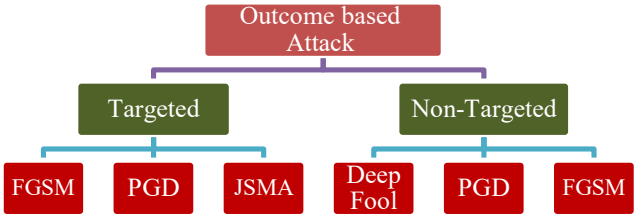

Figure 5. Outcome-based Adversary Attack Techniques.

1. **Targeted attacks:** These attacks cause a model to classify particular samples or a specific area of feature space to a particular (often incorrect) class. For illustration, assume that the class labels in a 3-class classification problem are A, B, and C. In a targeted attack, the input can be forced into class A, regardless of its actual class. Jacobian-Based Saliency Map (JSMA) is a famous technique for a targeted attack. Gradient approaches (FGSM, and PGD) can also be designed as targeted white-box attacks in case of the perturbation noise that has been added seeks to reduce the loss function of the image for a certain target class.

Jacobian-Based Saliency Map (JSMA) Attack

JSMA is a targeted white-box attack proposed by Papernot et al. [20]. It is based on the feature selection technique to decrease the number of features that need to be updated during adversarial sample construction. Perturbations are iteratively added to features in an increasingly prominent manner.

$$\nabla l(x) = \frac{\partial l(x)}{\partial x} = \left[\frac{\partial l_i(x)}{\partial x_y}\right]_{y\in 1,..m_{in}, i\in 1,...,m_{out}} \quad \text{Eq (7)}$$

Where x is the input, $l(x)$ is the Jacobian matrix, m_{in} and m_{out} is the neutrons' number in both the input and output layers sequentially. The Jacobian matrix shows how the logit outputs of various classes are affected by the components of the input x. The Jacobian matrix is used to build an adversarial salience map, which selects the features or pixels that need to be modified to produce the necessary changes in the logit outputs.

2. **Non-targeted attacks:** These attacks cause a DL model to incorrectly predict a subset of instances or parts of the feature space and don't specify which class they should be assigned to. From the previous example, a non-targeted attack doesn't stipulate that the input must be Class B or Class C. Exampled of the untargeted attack techniques are DeepFool (DF) Attack, FGSM and PGD attacks. While, FGSM and PGD attacks can be implemented as untargeted attacks when maximizing the loss function of the model.

DeepFool (DF) Attack

DeepFool is an untargeted white-box attack introduced by Moosavi-Dezfooli et al. [21]. DF is based on gradient iteration. This attack initially computes the perturbations, then continuously and iteratively moves the normal instances toward the decision boundary until adversarial examples are obtained after passing the decision boundary. Reducing the Euclidean distance between the real and perturbed samples is the basis of this method. Iteratively add perturbations while evaluating the decision boundary between classes.

$$\underset{\delta t}{\text{argmin}} \|\delta_t\| 2 \; subjected\; to\; f(x_t') + \nabla f(x_t')^t \delta_t = 0 \quad \text{Eq (8)}$$

Where x is the input and f is the classifier, $\|\delta_t\|$ is the minimal perturbation required to misclassify the input.

5.1.3 Adversarial Phase-based Attacks

According to the attack execution phase, there are two types of attacks: poisoning and evasion attacks.

1. **Poisoning attacks [22]:** This attack takes place during the training phase by constructing backdoors in the target model. These backdoors include poising adversarial examples in the training set. These attacks are usually white-box attacks or gray-box attacks, carried out by data modification or data injection. The Zoo, Biggio, and zugner's are examples of poising attacks.
2. **Evasion attacks [23]:** These attacks take place during the testing phase and involve adding small perturbations to benign test samples to fool the target model. This attack occurs when the model network has some perturbated inputs. These adversary inputs fool the model's classifier. These attacks may be black-box attacks or white-box attacks. The common techniques as mentioned before are the DF, FGSM, JSMA, PGD, and C&W attacks.

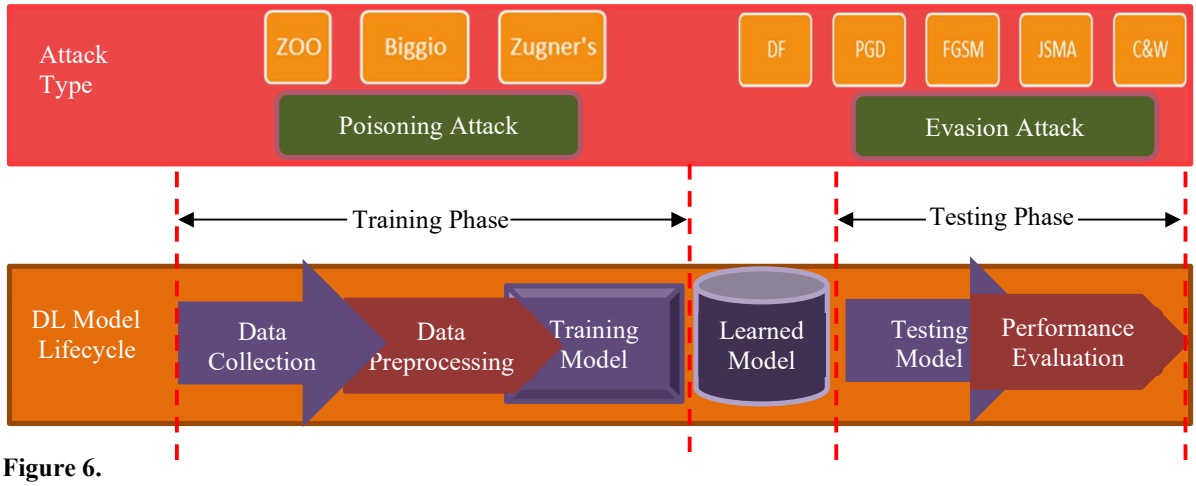

Figure 6.
Overview of Phase based Attacks. The current attack threat in DL is shown at the top of the image. The DL life cycle is shown in the second section.

5.2 Adversarial Attacks on Medical Images

The lack of robustness in DL models is demonstrated by the ability of adversarial samples to alter the behavior of classifiers and segmentation tasks. A weakness during segmentation could result in an inaccurate description of reality throughout the later phases of diagnosis. Another problem that creates a potential attack vector is the irregular and poorly understood boundaries of the classification. Although much research has been investigated on adversarial attacks for nature images, there is far less for medical imaging. According to many academics, adversarial scenarios are too challenging to occur in medical imaging. This section illustrates in detail the analysis of medical images, current adversarial attacks on medical images, the detection techniques, and the defenses model that have been applied in the medical domain using various imaging modalities.

5.2.1 Taxonomy of Medical Images

DL techniques are used in medical images to effectively extract information obtained from medical images captured utilizing a variety of imaging modalities. These modalities include ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), etc. These techniques are vital for the identification, location, and diagnosis processes of abnormalities because they provide Anatomical data regarding several human organs. Figure 7 shows a taxonomy of different medical imaging modalities and the existing applied adversarial attacks against each modality. The primary goal of medical image analysis is to help radiologists and physicians effectively diagnose and predict disease. The primary tasks of DL-based medical systems analysis, include image segmentation, retrieval, detection, classification, and reconstruction.

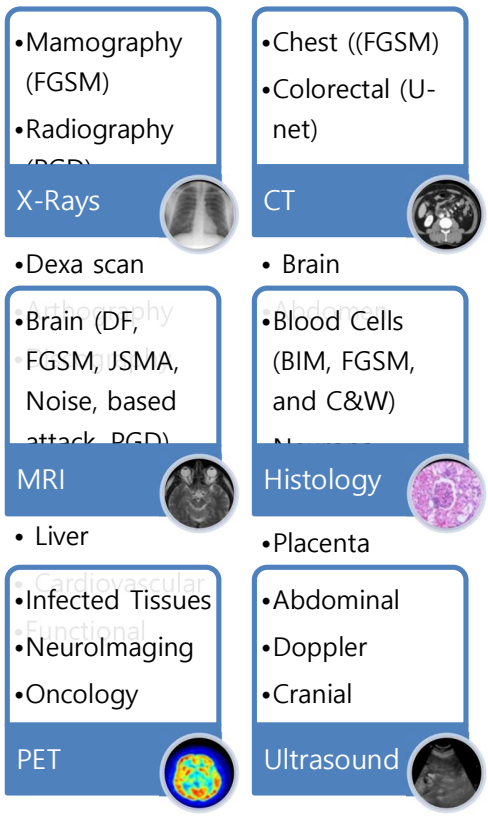

Figure 7. Sources of Medical Images Modalities with existing applied adversarial attacks.

5.2.2 Sources of Medical Vulnerabilities

As mentioned before, researchers have already shown that DL-based medical systems are susceptible to adversarial attacks. To ensure the security and integrity of DL based models and health data [24], it is crucial to address the following points as shown in Figure 8.

A. **Image Acquisition Vulnerabilities:** Some of the vulnerabilities that can affect the performance of the DL systems are discussed below, even if in practice medical images are carefully collected to ensure diagnostic accuracy.

- **Environmental And Instrumental Noise:** Medical imaging properties can be exploited by attackers. Camera direction can change benign retinal fundus pictures into adversarial samples. Similarities between the medical pictures' foreground and background make it challenging for the network to identify clean and perturbed samples. This is due to high-quality, consistent photo exposure. Noise can produce data artifacts. Due to the roughness of medical photos, DL models might focus on non-essential elements. Medical pictures are more vulnerable to assault than real-world images due to their nature.
- **Unskilled Staff:** healthcare systems include both technical and non-technical staff, and are so interdisciplinary. Therefore, they often struggle to find qualified individuals who can build and manage DL systems.

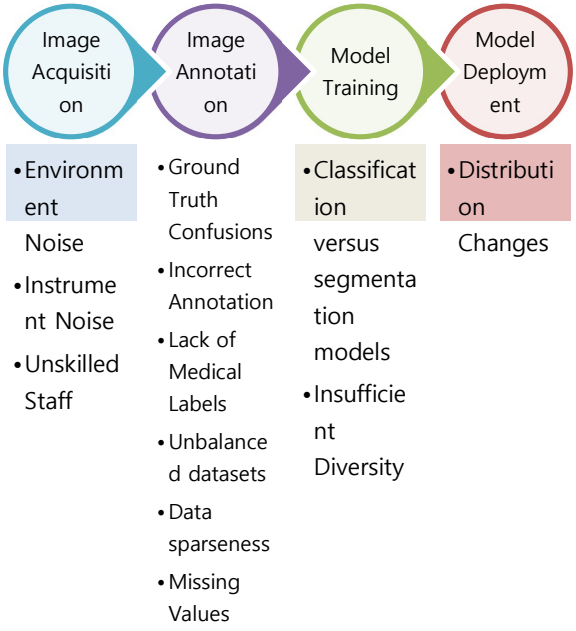

Figure 8. Vulnerabilities of Medical Deep Learning Pipeline.

B. **Medical Annotation Vulnerabilities:** Medical annotation is the process of labeling each sample. Ideally, qualified physicians should handle most of this duty (radiologists or physicians) to provide domain-enriched data sets, which are necessary for effective DL modeling in healthcare systems. According to research, it could be dangerous to train DL models without a firm understanding of the domain. This

section discusses particular vulnerabilities based on image annotations:

- **Ground Truth Confusions:** Malicious users who intend to manipulate the data worsen this problem by making its effects difficult to diagnose and impossible to recognize, even with expert human evaluation.
- **Incorrect Annotation:** Most commonly used medical images are tagged with coarse-grained labels. However, the true utility of DL is fine-grained, hidden layers in the clinical setting. Some efficiency issues can arise from improper labeling.
- **Lack of Medical Labels:** Medical photos are rare to train medical DL models due to a lack of legal regulations to communicate medical data, privacy issues, and time-consuming labeling. Even doctors and radiologists occasionally misinterpret medical pictures. All of these causes generate overfitting and memorization in medical DL models, which lowers generalizability and makes the network vulnerable to assaults.
- **Unbalanced and Limited datasets**: The number of datasets used to train DL models is insufficient. The lack of large-scale datasets is a major barrier to the effective implementation of DL techniques, as medical data is often modest in size.
- **Data sparseness (missing values)**: this is a typical occurrence in real-world data and can occur from several different causes. The effectiveness of DL models is substantially affected by the missing observations.

C. **Model Training Vulnerabilities:** model training vulnerabilities include privacy, ineffective or insufficient training, model tampering, and theft. When the DL models are trained with the wrong parameters, this is called incomplete or inappropriate training. Therefore, such critical applications must ensure the security and integrity of DL systems.

D.

- **Classification versus segmentation models:** Due to these factors, segmentation is more prone to attacks than classification models. Unlike classification, which assigns a label to the entire image, semantic segmentation assigns a label to each pixel. Second, it isn't easy to determine the success rate of attacks in segmentation models. Third, these models typically don't include high-quality image samples, which causes overfitting. Fourth, a segmentation image can easily be fooled by a minor change in image intensity.

- **Insufficient Diversity**: The architecture is uniform across nearly all medical machine vision tasks, although few learning models are similar. Since the design of transferable or global adversarial attacks has simplified attacking any medical system, the lack of a widespread variety of architectures and data sets makes it easier to compromise the security of neural networks.

A. **Deployment Vulnerabilities:** Ensuring the robustness of the system throughout the deployment phase is crucial.
Distribution Changes: occur when DL models that have been trained on images from one domain are subsequently operated on data from another domain. The performance of the model suffers greatly under these conditions. Additionally, these models used in healthcare are typically tested on new patients after being built using data from previous patients, raising concerns regarding the accuracy of DL predictions.

5.2.3 Existing Adversarial Attacks on Medical Images

Recent research studies have proved that the most sophisticated medical DL systems are considerably susceptible to adversarial attacks in three phases including the medical images classification phase, segmentation phase, and regression phase [25].

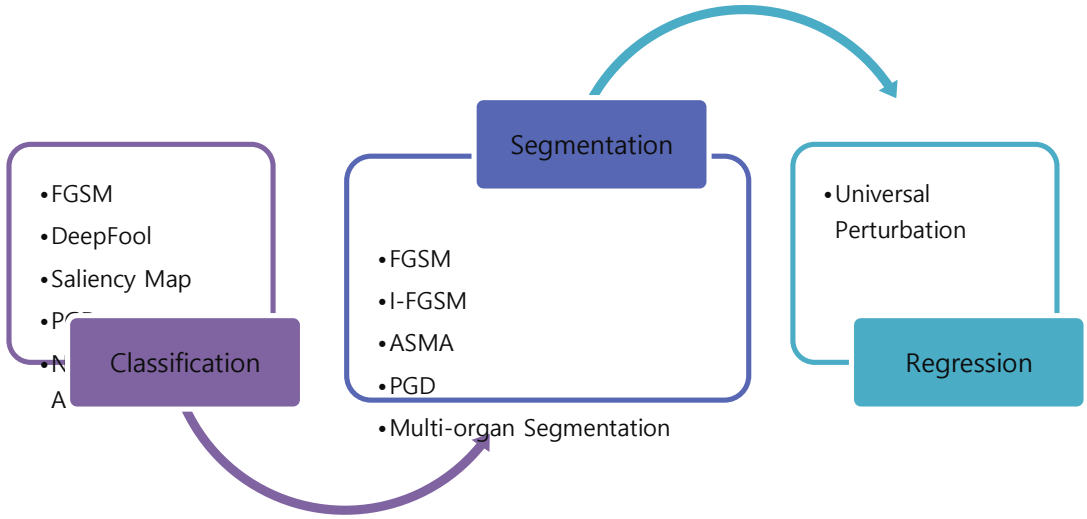

Figure 9. Review of Some Existing Attacks on Different DL Medical Images Phases.

A. Classification Phase Attacks

Classification is a significant use of DL in medical image analysis. Images are used as inputs, and deep learning models sort them into several categories. Models often indicate whether or not a patient has a disease. Many studies have been investigated in the area of DL medical classification attacks. For example, three adversarial attacks that are successfully used against many DL medical models (skin lesions classification and brain tumor segmentation) are first studied by Paschali, et al [26]. They used adversarial samples to investigate the vulnerabilities of these medical DL models. These adversarial samples are used to compare model performance on clean, crafted, and noisy data. The authors applied many DL models for skin lesions classifications including Inception (IV3), (IV4), and MobileNet (MN). While U-Net, SegNet, and DenseNet models have been utilized for the segmentation of the whole-brain. Many attack techniques have been used to create perturbated samples such as FGSM, DeepFool (DF) methods, and saliency map attacks (SMA). Finally, despite variations in accuracy when applying different attacks, the authors conclude that IV4 is recommended due to its equivalent generalization ability and resilience, among others. The scientists showed that for classification, the model's robustness is related to its depth, while for dense blocks, segmentation, and jump connections improve the model's effectiveness. As the adversary perturbation was from 0.97 to 0.99 was imperceptible. Finlayson et al. [27] demonstrated how the white and black-box attacks weaken the medical DL models. The pretrained ResNet50 model was used for diabetic retinopathy that is classified via retinal fundoscopy, dermoscopic, and pneumothorax from chest radiography. The black-and-white-box PGD attacks are applied against this model. The findings demonstrated that in both cases, the accuracy was significantly reduced. Kotia et al. [28] analyzed the impact of virtual adversarial training (VAT), FGSM, and noise-based adversarial attacks on brain tumor classification. The most successful attack was the FGSM, which reduced accuracy by 69%, At the same time, noise-based attacks andVAT only reduced accuracy by 34% and 24%, respectively. An adversarial training technique was used to increase the robustness of the model against noise-based and FGSM and produced excellent results. The training accuracy on the perturbated images was nearly identical to the test accuracy. In contrast, VAT adversary training was ineffective because it resulted in a 14% reduction in accuracy compared to training.

Huq et al. [29] examined the impact of adversary attacks on the classification of skin cancer. VGG16 and MobileNet models were used to classify seven types of skin cancer. Two white-box attacks were used to create perturbated samples including the FGSM and PGD. Accuracy dropped drastically after using those attacks. To increase the robustness, they recommended applying the adversarial training technique. Pal et al. [30] studied the impact of the

Manuscript received April 5, 2026
Manuscript revised April 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.4.17

FGSM attack on the CT and X-ray of COVID images. The VGG-16 and InceptionV3 models were used for the classification process. The results proved that these models are vulnerable to adversarial attack because the model accuracy decreased up to 63% in InceptionV3 and 90% in VGG-16. Yilmaz et al. [31] studied the impact of FGSM attacks on mammographic images. The Digital Database for Screening Mammograms (DDSM) has been used for experiments. DDSM includes two categories: malignant and normal. The structural similarity index (SSIM) is applied to correspond the original and perturbated images. They showed that when the SSIM dropped below 0.2, the accuracy dropped by as much as 30%.

B. Segmentation Phase Attacks

In medical imaging, segmentation refers to the extraction of particular components of the medical image, such as organs, cells, and tumors. In addition, clinical factors can be examined such as shape and volume. Li, et al [32]. Presented a custom attack against DL medical model called a query-based black-box (QBB). QBB within a limited query budget could change the foreground pixel classes. This attack used a more accurate gradient loss estimate and changed the fixed variance of the adaptive distribution to a learnable one. Experiments proved the efficiency of the suggested attack in deconstructing medical imaging segmentation models. Another contribution by Li et al. [33] applied FGSM and I-FGSM against pancreas segmentation. The pancreas is extremely hard to segment because of its small size, irregularities, texture, and location. They suggested a deep coarse-to-fine 3D model to address this issue. Datasets from JHMI and NIH were used in experiments. The framework is susceptible to the two types of attacks since they have significantly lower accuracy (85.83%). To address this issue, the authors suggested adversarial training for this model. Another segmentation attack called the adaptive segmentation mask attack (ASMA) was introduced by Ozbulak et al. [34].

ASMA is a targeted approach for medical image segmentation. U-Net model is used for the segmentation of two medical datasets for ISIC skin lesion segmentation and glaucoma optic disc. They proved that DL-based medical segmentation models are vulnerable to adversarial attacks. Cheng et al. [35] applied the universal random perturbation attack against the DL segmentation of brain tumors. The U-Net model has been used for segmentation. The MICCAI BraTS 2019 dataset was used in all experiments. The results indicated that concurrent perturbation significantly decreases model accuracy, while single-modal perturbation minimally reduces accuracy. Yao et al. [36] proposed an add-on to a white-box attack called Hierarchical Feature Constraint (HFC). HFC can be incorporated into any attack by hiding the adversarial perturbation in the original feature distribution. Adversary attacks are more easily identified on medical domain images than on natural domain images. The goal of HFC is to hide adversary features so that they are not easily detectable. The technique is applied to chest X-ray and fundoscopy datasets. ResNet50 and VGG16 models were used. The results proved that HFC can be applied to avoid adversarial samples from being detected. Another study by Anand et al. [37]. They applied the FGSM and PGD attacks against the segmentation of two medical datasets called MRI RVSC cardiac segmentation and chest radiography for pneumonia detection. Self-supervised learning (SSL), VGG11, and the U-Net segmentation model were used for experiments. The experimental results demonstrated that SSL produced better results as it learns more important features. Chen et al. [38] applied another adversarial attack against organ segmentation from CT scans. The intensity variations and geometrical deformations are used to create adversarial samples. The U-Net model was used for experiments. The results showed that the Dice score metric is significantly lowered in all organs.

C. Regression Phase Attacks

Regression models are used in the medical field. Examining the resistance of the regression models to attacks can be applicable for the reliability and security of medical systems. Li, Zhang, et al [39] published the first study on the effect of adversarial attacks against medical regression-based prediction models. They used 3D MRI brain samples to predict an individual's age. The experiments were performed using the hybrid DL and CNNs models. The adversarial samples were created with limited magnitude perturbation of l_0, l_2, and l_∞. A universal perturbation, as well as image-specific perturbation, are very successful in reducing the accuracy of the age prediction model. Since a single perturbation can inject considerable bias into the predictions, there is still serious concern about how resistant DL is to adversarial attacks. Table 1 lists the details and comparison between the above-mentioned attacks.

5.2.4 Defenses Against Medical Adversarial Attacks

This section discusses both the defense and detection techniques for DL-based medical imaging against adversarial attacks as shown in Figure 10.

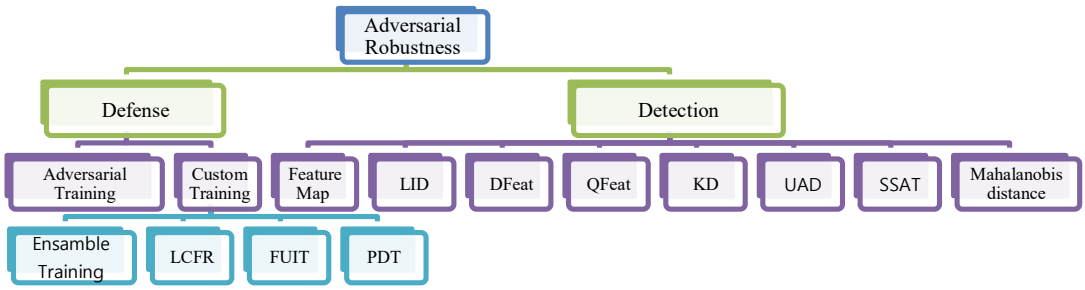

Figure 10. Medical DL-Based Adversarial Robustness Techniques.

A. Adversarial Training Defense Technique

Adversarial training (ADVT) is an adversarial defense technique in which DL models create adversarial examples and embed those examples in the training data. It is a natural defense mechanism against adversarial examples that increases the robustness of a neural network while the model is being trained. Biggio et al. [65], reported that a neural network can be trained on adversarial examples to be resistant to first-order attack strategies. For example, Huang et al. [67] advocated robust adversarial training techniques to solve the limited-angle tomography (LAT) problem. This problem can cause problems with CT reconstruction. Since poisonous noise is frequently found on CT scans, they used it as a disturbance in the images for training. The U-Net model was utilized for experiments with the dataset of Low-Dose CT scans. The poison noise retraining is quite substantial for angle-limited reconstruction but insufficient for non-local adversarial instances. Another contribution proposed by Vatian et al. [68] is on natural adversary attacks. They used the CNN model for brain MRI and CT scans for lung cancer classification. They showed that a noise that behaves like a "natural" adversarial example can exist in sophisticated medical imaging systems. To protect against these attacks, three defense strategies were applied. The ADVT with FGSM and JSMA achieved the best defense strategy compared with the other two strategies. While layer activation function switching with bounded ReLU and data augmentation with gaussian noise were the other two strategies. Liu et al. [69] studied the robustness of the ADVT defense technique in the classification of lung nodules using CT images. Three 3DResUNets models were used for model training and data were obtained from the NLST and LUNA datasets. They applied the PGD attack to confidently identify patterns that result in misclassification before training the network. The authors suggest using ADVT for nodule classification to increase model resistance to unexpected noise. Wu et al. [70] applied the adversarial training defense technique to increase the robustness of the diabetic retinopathy-based DL classification model. The PGD attack was combined with ResNet32 to create the perturbated samples. They proved that the effectiveness of the model under attack was greatly enhanced using the adversary training defense technique.

B. Custom Defense Techniques

Rao et al [71]. applied many attacks to diagnose X-ray thorax disease. As well, the authors investigate two different forms of protection techniques, the transformation of pixel deviation (PDT) and adversary PGD training. PDT chooses a random pixel from adverse instances and changes it with a different pixel chosen from a small square region. They concluded that their novel method solved the problem of poor performance of existing defense strategies for large disturbances. Mani et al. [72] proposed a custom adversarial defense model named a fuzzy unique image transformation (FUIT) to protect the COVID-19 classification model. Before feeding the training samples to the model, the image's pixels are modeled into an interval. The results proved that this technique is successful against six untargeted attacks. Although this approach doesn't change the design of the model, it requires additional time to transform the images during testing. Liu et al. [73] suggested another successful defense strategy for medical image segmentation using low-cost frequency refinement (LCFR). ASMA attack was applied against fundoscopy and skin photographs classification models. Since only frequency refinement of the images is required, this method doesn't affect on the model's complexity. Raul et al. [74] introduced an ensemble training defense technique on adversarial attacks using CT scans of the NLST dataset to predict the malignancy of lung nodules. The one-pixel black-box and FGSM white-box attacks were used to create the adversarial samples. The attacks lead to a 28% to 36% decrease in accuracy. Each CNN architecture of the

Manuscript received April 5, 2026
Manuscript revised April 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.4.17

proposed ensemble model was initialized with seven different parameters to drastically reduce the accuracy of the attack.

C. Detection Techniques

The detection of DL adversarial attacks is another essential technique to defend the medical classifiers. These techniques determine whether the input is benign or perturbated before feeding into the classification model input. The medical DL model refuses to predict the input sample if it can determine that it is perturbated. However, DL models used for medical imaging are more susceptible to adversary attacks than in the real world, instances of medical adversaries are easier to detect. Li et al. [75] implemented a detection technique for adversarial attack based on the feature map. Where adversarial samples have different feature maps compared with the original samples. The detection model rejects adversarial instances before the classification process. The PGD, FGSM, and BIM attacks were applied. The DenseNet121 model was used for the classification of chest X-rays. This technique doesn't require any change in the model architecture. Also, this technique doesn't need knowledge of the target model. Ma, et al. [76], studied the performance of four adversarial detection techniques for medical DL models. The four detection techniques include the local intrinsic dimensionality (LID), deep features (DFeat), kernel density (KD), and quantized features (QFeat) applied against four attacks. Four attack strategies including PGD, FGSM, C&W, and BIM are applied against the ResNet50 model. Three medical datasets were used for experiments including fundoscopy, chest X-ray and ISIC. The results demonstrated that the detection techniques against the four attacks offer highly robust performance. KD achieved the maximum AUC of 99% compared with other attacks for all three datasets.

Since these methods rely on guessing the density of the adversary's samples, even if the performance is quite good, the method's applicability is restricted to a few common attack types. Li, et al. [77] introduced a hybrid adversarial detection technique based on unsupervised adversarial detection (UAD) and semi-supervised adversarial training (SSAT). The system is proposed to deliver a robust DL medical imaging framework. The ResNet18 model was used to classify the OCT images. The results proved that the semi-supervised training greatly reduces the vulnerability of the model. This approach increases the complexity of the model by finding robust features. The authors also suggested a new risk metric for the adversarial attack that would depend on the classification scenario and the acceptance or rejection of unsupervised adversarial detection. Uwimana et al. [78] proposed another medical adversarial detection technique. The Mahalanobis distance is applied to detect the out-of-distribution (OOD) data and the adversarial samples of the malaria classification model. The DeepFool, BIM, FGSM, and C&W attacks were applied. This technique offers a detection accuracy of up to 99.95%. Table 2 shows a detailed comparison of these defense and detection techniques against DL-based medical adversarial attacks.

5.3 Adversarial Attacks on Medical Text (MQA Systems)

5.3.1 Nature of Text Attacks

Adversarial attacks on MQA systems typically involve subtle manipulations of text that exploit the vulnerabilities of natural language processing (NLP) models. These attacks can take various forms including word-level, character-level, sentence-level and multi-level manipulations.

- Word-level attacks: attackers replace or substitute key medical terms with semantically similar or dissimilar words to confuse the system. For example, a term like "malignant" could be replaced with "benign". This leades to incorrect or misleading medical advice.
- Character-level attacks: involve altering individual letters within a word such as changing "cancer" to "cqncer," which can fool the model while remaining unnoticed by human users.
- Sentence-based attacks: involve reordering words or altering sentence structure to distort the meaning without significantly changing the overall content. This disrupt the MQA system's ability to accurately process and respond to medical queries.
- Multi-level attacks: involve many types of attacks such as character, word, and sentence level attacks.

5.3.2 Impact on MQA

The consequences of these adversarial text attacks can be severe. Since MQA systems are often used to provide critical health information or assist in clinical decision-making, an adversarial perturbation could lead to dangerous misinterpretations of medical advice. For instance, an altered query might cause the system to recommend inappropriate treatments or incorrect diagnoses. This could result in real-world harm to patients. Misleading medical responses could undermine the trust between healthcare providers and patients. Also, this could lead to delays in diagnosis, inappropriate medication prescriptions, or even life-threatening decisions in clinical settings. The subtle nature of these attacks makes them particularly insidious as the erroneous output may not be immediately recognized as incorrect by users.

Adversarial attacks on medical text and MQA systems represent a critical and underexplored threat to the integrity of healthcare delivery. With the growing reliance on NLP-driven systems for clinical decision support and patient interaction, it is essential to develop robust defenses to mitigate the impact of these sophisticated attacks.

6. The Importance of Medical Adversarial Attacks in the Research Field

This section discusses the analysis of the importance of adversarial attacks based on the Scopus database [97] articles as a base with keywords " Adversarial attacks, defense of Medical images, healthcare" from 2018 to 2022. The following sections will illustrate how articles are distributed based on time, and funding institutions respectively. The overall evolution of the adversarial attack field can be understood intuitively and efficiently by gaining a broad perspective from field publications.

A. Distribution of Articles related to Medical Adversarial Attacks over Time

After 2018, medical adversarial attack research started a new era of development, as evidenced by a significant increase in research interest as shown in Figure 11. Research in this sector has grown rapidly and hasn't quite gained its peak yet. The published research in this field is expected to exceed two thousand by 2023. Research of medical adversarial attacks remains a major research focus at this time.

Top Funding Sponsors Entities related to Medical Adversarial Attacks

Figure 12 demonstrates how institutions have supported researchers over the past five years in the area of adversarial attacks against DL medical models. It also indicates that China's National Natural Science Foundation funded more than 700 research publications in the area of adversarial attacks, which is related to the field of computer science, according to the Scopus database.

The institutions are listed in order by the number of researches as follows: "National Natural Science Foundation of China, National Science Foundation, National Institutes of Health, National Key Research and Development Program of China, National Research Foundation of Korea, Fundamental Research Funds for the Central Universities, Horizon 2020 Framework Programme,

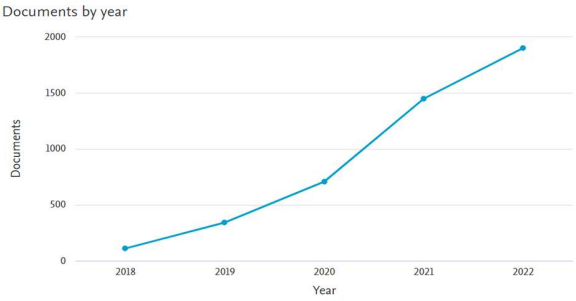

Figure 11. Medical adversarial attack research numbers from 2018 to 2022.

Engineering, and Physical Sciences Research Council, European Regional Development Fund, Ministry of Science, and ICT and Future Planning". Based on financing, China has the most publications in this discipline. Also, Adversarial attacks have been the subject of extensive research in all of the world's major nations and areas. This indicates that security research on adversarial attacks and related topics is highly regarded worldwide. As this topic is multidisciplinary, research in it hasn't become more difficult across national boundaries.

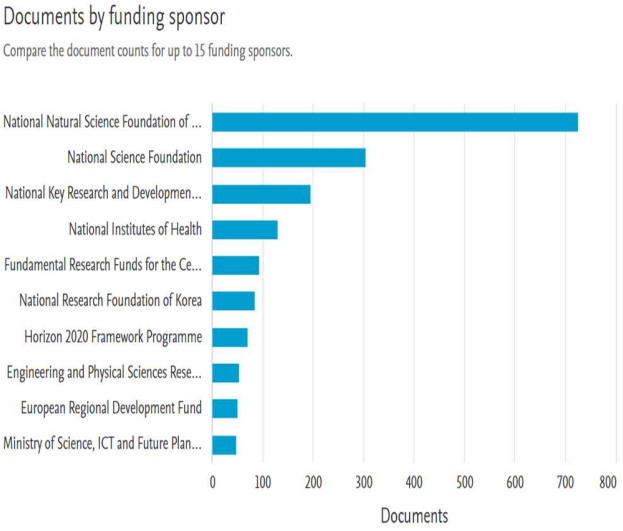

Figure 12. Medical Adversarial Attack research numbers funded by institutions from 2018 to 2022.

Table 1. Comparison of Existing Defense and Detection Techniques Against Adversarial Medical Attacks

Ref	Advs. Attacks	DL model	Defense Technique	Performance Improvement	Dataset	Advantages	Limitation
[67]	LAT	U-Net	ADVT	More resistance to poison noise	CT [79]	– The U-Net can reduce most artifacts	– The model is still vulnerable to non-local attack
[68]	FGSM, JSMA	CNN	ADVT, bounded ReLU and gaussian noise	Improves up to 2%	LIDCIDRI [80], BRATS 2015[81]	– Studied the impact of three defense techniques.	– Defend against only two types of white-box attacks.
[69]	PGD	3D ResNets	ADVT	Improves up to 10% and 35%	LUNA [82], NLST [83]	– Model robustness against three types of image noise – The model can be applied to other similar models if the data and the model are fully differentiable	– The model didn't synthesize all types of Nodules. – Motion distortion and metal artifacts noise not investigated
[70]	PGD	ResNet 32	ADVT	Accuracy enhanced by 40%	APTOS [84]	– Achieved a high defense rate	– Tested against a single model structure with a single medical modality against a single attack
[71]	FGSM, DII-FGSM, PGD, MIFGSM,	Dense Net, Res-50 and VGG16	PDT & ADVT	AUC up to 85%	CheXpert[85]	– Solve the limitation of excessive iterations – High success rate – Experiment against many attacks with different Dl model structures	– More enhancement is required to handle the large and different values of perturbations. – The performance of the PDT was tested against a single medical modality.
[72]	FGSM, C&W, BIM, DF, PGD, PGD-r	CNN	FUIT	The accuracy increased by up to 40%.	COVID-19 Chest X-Ray [86], CT scans [87]	– Six attacks are studied against binary and multiclass data – Achieved a high robustness rate	– The experiments are performed using small datasets with fewer images
[73]	ASMA	ResNet -50, U-Net, Dense Net	LCFR	The accuracy increased by up to 40%	Glaucoma optic [88], ISIC skin [89]	– Authors evaluate both segmentation and classification robustness – High defense rate	– Many attacks should be studied.
[74]	FGSM, One-pixel	CNN	Ensemble Training	Accuracy increased up to 34%.	NLST [90]	– High defense rate. – Seven CNN architectures with different seed points are trained to verify the stability.	– Defense against 3D slices should be studied. – Segmentation is performed based on a semi-automatic technique.
[75]	FGSM, BIM, PGD	Dense Net121	Detection-based Feature Map (FM)	F1 detection score up to 0.97	NIH ChestX-ray14[91]	– Defend against white and black-box attacks. – Can be embedded in any DL medical system.	– Multimodal distributions of complex datasets are required to be studied. – A new dimension reduction technique is required to be investigated.

Manuscript received April 5, 2026
Manuscript revised April 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.4.17

[76]	FGSM, BIM, PGD, C&W	ResNet 50	Detection Techniques (DFeat, LID, QFeat, KD)	Detection accuracy up to 98%	Diabetic retinopathy [92] Chest X-Ray [93] Melanoma [94]	– Four detection methods against four attacks are studied. – Binary and multi-class datasets are tested.	– Experiments ignored the effect of Black-box attacks.
[77]	FGSM, PGD, C&W	ResNet 18	Detection UAD, SSAT	Detection accuracy up to 0.97	OCT [95]	– Two detection methods against three white-box attacks are studied. – High-detection rate	– experiments ignored the binary classification results and the detection of black-box attacks.
[78]	BIM, FGSM, and C&W	VGG-19, ResNet -18	Detection using Mahalanobis distance	detection accuracy of up to 99.95%.	Malaria Cells [96]	– The system handles both the adversarial and out-of-distribution samples against three attacks and two DL models.	– Experiments ignored multi-class datasets and the detection of black-box attacks.

7. Discussion and Future Directions of Medical Adversarial Attacks

The integration of DL technologies in the healthcare sector is progressing rapidly. Below, we discuss the potential, limitations, and future directions for research on medical adversarial attacks.

7.1 Potential Vulnerabilities of Deep Learning in Healthcare

The applications of DL technologies in the healthcare industry are expanding rapidly. DL models are a branch of AI that has shown the potential to aid in diagnostic, analysis and constitute an inexpensive tool to support human decision-making. DL medical models are susceptible to cyberattacks which have the potential to be seriously damaging. DL systems can be tricked by adversarial attacks into making inaccurate predictions. However, it is crucial to ensure that these systems are resilient to artifacts interference throughout this future widespread deployment of AI systems. In this paper, many research contributions proved that DL models for medical imaging are more susceptible to adversary attacks than natural images. Furthermore, adversarial training techniques are often partially effective against adversarial attacks on medical imaging. The perturbation size of adversarial attacks against medical imaging modalities is substantially less than what is used for natural imaging in a different study.

7.2 Importance of Understanding Adversarial Attacks

A. **Mechanisms of Exploitation:** Understanding how adversarial attacks manipulate DL models is critical for developing robust solutions. The transformation of AI in healthcare isn't a simple task.

B. **Collaborative Efforts:** Multiple stakeholders must collaborate to address ethical and legal considerations in AI deployment. These ethics include:

- **Legal rules:** for data governance usage of patient data and regulations.
- **Data governance:** covers a variety of topics, such as how data is secured and stored, maintained, kept, and who has the privilege to access the medical database.
- **Policies and procedures:** to ensure the protection of patient data. Regulations were published to ensure that medical data is used correctly. It is commonly recognized that good labeling and image quality are essential for the development of medical systems.

7.3 Limitations of Current Research

It has become clear that adversary attacks are essential to assessing the vulnerability of DL networks. The improvement of adversarial attack techniques could help researchers address problems and develop more secure and effective medical learning systems. The main issues shortly will continue to be the visibility of the perturbation and the transferability of attack methods while having a significant impact on network predictions. Current research faces many limitations such as:

- **Complexity of Medical Data:** The variability and complexity of medical data can hinder the application of adversarial defense strategies.
- **Challenges in Data Annotation:** The high costs and privacy concerns associated with extensive

labeled data can be a barrier to training robust models.

- **Computational Overhead:** Adversarial training may introduce additional computational demands that are impractical in resource-constrained clinical settings.

More techniques of adversarial attacks aimed at weakening the robustness of targeting segmentation models will also be considered. another direction could be investigating new methods to maintain larger datasets and quickly produce images with medical annotations to create more reliable medical learning systems. Altering the neural network architecture which hasn't yet been examined is another method that can decrease network vulnerability.

7.4 Future Directions

The deployment of AI technologies in healthcare presents significant opportunities for improving clinical outcomes. However, addressing the vulnerabilities posed by adversarial attacks is essential for ensuring patient safety and maintaining the integrity of healthcare systems. Continued research in this domain is crucial for developing effective defenses and establishing a framework for the responsible and ethical use of AI in medical applications.

- **Advancing Adversarial Attack Techniques:** Researchers should explore more sophisticated techniques to better understand vulnerabilities in medical systems.
- **Focus on Key Issues:** Future investigations should prioritize the visibility of perturbations and the transferability of attack methods impacting network predictions.
- **Improving Medical Datasets:** Developing methods for maintaining larger datasets and rapidly generating annotated medical images will enhance the reliability of medical learning systems.
- **Exploring Neural Network Architectures:** Investigating alternative neural network architectures could potentially reduce vulnerabilities and enhance resilience against adversarial threats.
- **Developing Defense and Detection Techniques for Medical Text:** There is a need to explore robust defense mechanisms and detection techniques specifically for adversarial attacks on Medical Question Answering (MQA) systems and medical text. To ensure that text-based applications are equally protected against manipulation.

8. Conclusion

The adversarial attacks phenomenon against deep learning models has become a major barrier when DL-based applications are used in security-critical contexts. This presents serious security threats and challenges for DL-based medical models. Therefore, it has become urgently necessary to investigate robust techniques that are more resistant to these attacks. This paper provides an extensive and in-depth overview of adversarial attacks in the area of medical diagnosis-based DL models. It provides a taxonomy that summarized the different strategies of adversarial attacks. It also discusses and evaluates many published attack, detection, and defense techniques that have been published in the area of DL-based medical adversarial attacks field. This paper introduces an in-depth discussion for the root sources of medical vulnerabilities and the potential impact of these vulnerabilities on the robustness of healthcare systems. The state of art research direction is also visualized to better understand the development of medical adversarial attacks over the last five years. Based on the extensive field of research from multiple aspects such as model improvement, new research areas with significant and practical relevance are suggested. New proposed medical adversarial attacks in other areas such as medical visual question answering, clinical data, etc. haven't been reviewed, which would require further future work to accomplish.

References

1. Avanzo, Michele, Massimiliano Porzio, Leda Lorenzon, Lisa Milan, Roberto Sghedoni, Giorgio Russo, Raffaella Massafra et al. "Artificial intelligence applications in medical imaging: A review of the medical physics research in Italy." *Physica Medica* 83 (2021): 221-241.
2. Zare Harofte, Samaneh, Madjid Soltani, Saeed Siavashy, and Kaamran Raahemifar. "Recent Advances of Utilizing Artificial Intelligence in Lab on a Chip for Diagnosis and Treatment." *Small* 18, no. 42 (2022): 2203169.
3. Wang, Shanshan, Guohua Cao, Yan Wang, Shu Liao, Qian Wang, Jun Shi, Cheng Li, and Dinggang Shen. "Review and prospect: Artificial intelligence in advanced medical imaging." *Frontiers in Radiology* 1 (2021): 781868.
4. Arvanitis, Theodoros N. "Informatics Opportunities and Challenges in Medical Imaging: A Journey." *Achievements, Milestones and Challenges in Biomedical and Health Informatics* (2022): 19-29.

5. Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199* (2013). Available online: http://arxiv.org/abs/1312.6199 (accessed on 30 Nov 2022).
6. Ma, Xingjun, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. "Understanding adversarial attacks on deep learning based medical image analysis systems." Pattern Recognition 110 (2021): 107332.
7. Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." In *2017 ieee symposium on security and privacy (sp)*, pp. 39-57. Ieee, 2017.
8. Madry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." *arXiv preprint arXiv:1706.06083* (2017).
9. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
10. Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." In *Artificial intelligence safety and security*, pp. 99-112. Chapman and Hall/CRC, 2018.
11. Cearley 2019. Top 10 strategic technology trends for 2020: A gartner special report. https://www.gartner.com/en/doc/432920- top- 10- strategic- technology- trends- for- 2020, Accessed: 2022, December 5.
12. Boone, Lyndon, Mahdi Biparva, Parisa Mojiri Forooshani, Joel Ramirez, Mario Masellis, Robert Bartha, Sean Symons et al. "ROOD-MRI: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI." arXiv preprint arXiv:2203.06060 (2022).
13. Bhusal, Dipkamal, and Nidhi Rastogi. "Adversarial Patterns: Building Robust Android Malware Classifiers." arXiv preprint arXiv:2203.02121 (2022).
14. Miller, David J., Zhen Xiang, and George Kesidis. "Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks." Proceedings of the IEEE 108, no. 3 (2020): 402-433.
15. Zhang, Xiao, and Dongrui Wu. "On the vulnerability of CNN classifiers in EEG-based BCIs." IEEE transactions on neural systems and rehabilitation engineering 27, no. 5 (2019): 814-825.
16. Chen, Pin-Yu, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. "Ead: elastic-net attacks to deep neural networks via adversarial examples." In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.
17. Chen, Pin-Yu, 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, pp. 15-26. 2017.
18. Xiao, Chaowei, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. "Generating adversarial examples with adversarial networks." arXiv preprint arXiv:1801.02610 (2018).
19. Qiu, Shilin, Qihe Liu, Shijie Zhou, and Chunjiang Wu. "Review of artificial intelligence adversarial attack and defense technologies." Applied Sciences 9, no. 5 (2019): 909.
20. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proceedings of the2016 IEEE European Symposium on Security and Privacy; 2016 Mar 21–24; Saarbrucken, Germany; 2016. p. 372–87.
21. Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574-2582. 2016.
22. Hu, Yupeng, Wenxin Kuang, Zheng Qin, Kenli Li, Jiliang Zhang, Yansong Gao, Wenjia Li, and Keqin Li. "Artificial intelligence security: threats and countermeasures." ACM Computing Surveys (CSUR) 55, no. 1 (2021): 1-36.
23. Ahmed, Usman, Jerry Chun-Wei Lin, and Gautam Srivastava. "Mitigating adversarial evasion attacks of ransomware using ensemble learning." Computers and Electrical Engineering 100 (2022): 107903.
24. Kaissis, Georgios A., Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. "Secure,

privacy-preserving and federated machine learning in medical imaging." Nature Machine Intelligence 2, no. 6 (2020): 305-311.
25. Qu, Liangqiong, Niranjan Balachandar, Miao Zhang, and Daniel Rubin. "Handling data heterogeneity with generative replay in collaborative learning for medical imaging." Medical Image Analysis 78 (2022): 102424.
26. Paschali, Magdalini, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. "Generalizability vs. robustness: investigating medical imaging networks using adversarial examples." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 493-501. Springer, Cham, 2018.
27. Finlayson, Samuel G., Hyung Won Chung, Isaac S. Kohane, and Andrew L. Beam. "Adversarial attacks against medical deep learning systems." arXiv preprint arXiv:1804.05296 (2018).
28. Kotia, Jai, Adit Kotwal, and Rishika Bharti. "Risk susceptibility of brain tumor classification to adversarial attacks." In International Conference on Man–Machine Interactions, pp. 181-187. Springer, Cham, 2019.
29. Huq, Aminul, and Mst Tasnim Pervin. "Analysis of adversarial attacks on skin cancer recognition." In 2020 International Conference on Data Science and Its Applications (ICoDSA), pp. 1-4. IEEE, 2020.
30. Pal, Biprodip, Debashis Gupta, Md Rashed-Al-Mahfuz, Salem A. Alyami, and Mohammad Ali Moni. "Vulnerability in deep transfer learning models to adversarial fast gradient sign attack for covid-19 prediction from chest radiography images." Applied Sciences 11, no. 9 (2021): 4233.
31. Yilmaz, Ibrahim, Mohamed Baza, Ramy Amer, Amar Rasheed, Fathi Amsaad, and Rasha Morsi. "On the assessment of robustness of telemedicine applications against adversarial machine learning attacks." In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 519-529. Springer, Cham, 2021.
32. Li, Siyuan, Guangji Huang, Xing Xu, and Huimin Lu. "Query-based black-box attack against medical image segmentation model." Future Generation Computer Systems 133 (2022): 331-337.
33. Li, Yingwei, Zhuotun Zhu, Yuyin Zhou, Yingda Xia, Wei Shen, Elliot K. Fishman, and Alan L. Yuille. "Volumetric medical image segmentation: a 3D deep coarse-to-fine framework and its adversarial examples." In Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics, pp. 69-91. Springer, Cham, 2019.
34. Ozbulak, Utku, Arnout Van Messem, and Wesley De Neve. "Impact of adversarial examples on deep learning models for biomedical image segmentation." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 300-308. Springer, Cham, 2019.
35. Cheng, Guohua, and Hongli Ji. "Adversarial perturbation on MRI modalities in brain tumor segmentation." IEEE Access 8 (2020): 206009-206015.
36. Yao, Qingsong, Zecheng He, Yi Lin, Kai Ma, Yefeng Zheng, and S. Kevin Zhou. "A hierarchical feature constraint to camouflage medical adversarial attacks." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 36-47. Springer, Cham, 2021.
37. Anand, Deepak, Darshan Tank, Harshvardhan Tibrewal, and Amit Sethi. "Self-supervision vs. transfer learning: robust biomedical image analysis against adversarial attacks." In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1159-1163. IEEE, 2020.
38. Chen, Liang, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. "Intelligent image synthesis to attack a segmentation CNN using adversarial learning." In International Workshop on Simulation and Synthesis in Medical Imaging, pp. 90-99. Springer, Cham, 2019.
39. Li, Yi, Huahong Zhang, Camilo Bermudez, Yifan Chen, Bennett A. Landman, and Yevgeniy Vorobeychik. "Anatomical context protects deep learning from adversarial perturbations in medical imaging." Neurocomputing 379 (2020): 370-378.
40. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosc. 19(9), 1498–1507.(2007)
41. Ballerini, L., Fisher, R.B., Aldridge, R.B., Rees, J.: A color and texture based hierarchical K-NN approach to the classification of non-melanoma

skin lesions. In: Color Medical Image Analysis (2013).
42. Diabetic Retinopathy Detection (Available Online). https://www.kaggle.com/c/diabetic-retinopathy-detection/data , (last visited 22 Dec 2022).
43. Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 3462–3471.
44. Cheng, Jun. "Brain tumor dataset. figshare. dataset." (2017). https://doi.org/10.6084/m9.figshare.1512427.v5.
45. P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", Scientific Data, vol. 5, pp. 180161, 2018.
46. Cohen, J.P.; Morrison, P.; Dao, L.; Roth, K.; Duong, T.Q.; Ghassemi, M. Covid-19 Image Data Collection: Prospective Predictions Are the Future. arXiv 2020, arXiv:2006.11988. Available online: https://arxiv.org/pdf/2006.11988.pdf (accessed on 15 April 2021).
47. Yang, X.; He, X.; Zhao, J.; Zhang, Y.; Zhang, S.; Xie, P. COVID-CT-Dataset: A CT Scan Dataset about COVID-19. arXiv 2020, arXiv:2003.13865.
48. Breast Database. http://marathon.csee.usf.edu/Mammography/Database.html, Last visited 22 Dec 2022.
49. Shiraishi, Junji, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules." American Journal of Roentgenology 174, no. 1 (2000): 71-74.
50. Roth, H.R.; Lu, L.; Farag, A.; Shin, H.-C.; Liu, J.; Turkbey, E.; Summers, R.M. DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation. arXiv 2015, arXiv:1506.06448. Available online: http://arxiv.org/abs/1506.064(accessed on 4 June 2021).
51. Roth, H.R.; Lu, L.; Farag, A.; Shin, H.-C.; Liu, J.; Turkbey, E.; Summers, R.M. DeepOrgan: Multi-level Deep Convolutional Networks for Automated Pancreas Segmentation. arXiv 2015, arXiv:1506.06448. Available online: http://arxiv.org/abs/1506.064(accessed on 4 June 2021).
52. Pena-Betancor, C., et al.: Estimation of the relative amount of hemoglobin in the cup and neuroretinal rim using stereoscopic color fundus images. Invest. Ophthalmol. Vis. Sci. 56(3), 1562–1568 (2015).
53. Gutman, D., et al.: Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). CoRR. abs:1605.01397 (2016).
54. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans. Med. Imaging 2015, 34, 1993–2024.
55. Kaggle: Chest X-Ray Images (Pneumonia) (2019). https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia, last visited 25 Dec 2022.
56. Mendeley Data—Labeled Optical Coherence Tomography (OCT) and Chest X-ray Images for Classification. Available online: https://data.mendeley.com/datasets/rscbjbr9sj/2 (accessed on 5 June 2021).
57. Daniel Kermany, Kang Zhang, and Michael Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," Mendeley Data, vol. 2, 2018.
58. Tong, T., et al.: Discriminative dictionary learning for abdominal multi-organ segmentation. Med. Image Anal. 23(1), 92–104 (2015).
59. ABIDE 1: http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html, Last visited 22 Dec 2022.
60. IXI: http://brain-development.org/ixi-dataset/, Last visited 22 Dec 2022.
61. FCON-1000: http://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html, Last visited 22 Dec 2022.
62. OASIS1: https://www.oasis-brains.org/, Last visited 22 Dec 2022.
63. NDAR: https://ndar.nih.gov/study.html?id=318, Last visited 22 Dec 2022.

64. ADHD200: http://fcon_1000.projects.nitrc.org/indi/adhd200/, Last visited 22 Dec 2022.
65. NKI Rockland: http://fcon_1000.projects.nitrc.org/indi/enhanced/neurodata.html, Last visited 22 Dec 2022.
66. Biggio, Battista, Luca Didaci, Giorgio Fumera, and Fabio Roli. "Poisoning attacks to compromise face templates." In 2013 international conference on biometrics (ICB), pp. 1-7. IEEE, 2013.
67. Huang, Yixing, Tobias Würfl, Katharina Breininger, Ling Liu, Günter Lauritsch, and Andreas Maier. "Some investigations on robustness of deep learning in limited angle tomography." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 145-153. Springer, Cham, 2018.
68. Vatian, Aleksandra, Natalia Gusarova, Natalia Dobrenko, Sergey Dudorov, Niyaz Nigmatullin, Anatoly Shalyto, and Artem Lobantsev. "Impact of adversarial examples on the efficiency of interpretation and use of information from high-tech medical images." In 2019 24th Conference of Open Innovations Association (FRUCT), pp. 472-478. IEEE, 2019.
69. Liu, Siqi, Arnaud Arindra Adiyoso Setio, Florin C. Ghesu, Eli Gibson, Sasa Grbic, Bogdan Georgescu, and Dorin Comaniciu. "No surprises: Training robust lung nodule detection for low-dose CT scans by augmenting with adversarial attacks." IEEE Transactions on Medical Imaging 40, no. 1 (2020): 335-345.
70. Wu, Dawen, Shishi Liu, and Jian Ban. "Classification of Diabetic Retinopathy Using Adversarial Training." In IOP Conference Series: Materials Science and Engineering, vol. 806, no. 1, p. 012050. IOP Publishing, 2020.
71. Rao, Chendi, Jiezhang Cao, Runhao Zeng, Qi Chen, Huazhu Fu, Yanwu Xu, and Mingkui Tan. "A thorough comparison study on adversarial attacks and defenses for common thorax disease classification in chest x-rays." arXiv preprint arXiv:2003.13969 (2020).
72. Mani Tripathi, Achyut, and Ashish Mishra. "Fuzzy Unique Image Transformation: Defense Against Adversarial Attacks On Deep COVID-19 Models." arXiv e-prints (2020): arXiv-2009.
73. Liu, Qi, Han Jiang, Tao Liu, Zihao Liu, Sicheng Li, Wujie Wen, and Yiyu Shi. "Defending deep learning-based biomedical image segmentation from adversarial attacks: a low-cost frequency refinement approach." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 342-351. Springer, Cham, 2020.
74. Paul, Rahul, Matthew Schabath, Robert Gillies, Lawrence Hall, and Dmitry Goldgof. "Mitigating adversarial attacks on medical image understanding systems." In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1517-1521. IEEE, 2020.
75. Li, Xin, and Dongxiao Zhu. "Robust detection of adversarial attacks on medical images." In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1154-1158. IEEE, 2020.
76. Ma, Xingjun, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. "Understanding adversarial attacks on deep learning based medical image analysis systems." Pattern Recognition 110 (2021): 107332.
77. Li, Xin, Deng Pan, and Dongxiao Zhu. "Defending against adversarial attacks on medical imaging AI system, classification or detection?." In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1677-1681. IEEE, 2021.
78. Uwimana, Anisie, and Ransalu Senanayake. "Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis." arXiv preprint arXiv:2107.04882 (2021).
79. AAPM Low-Dose CT Grand Challenge, https://www.aapm.org/grandchallenge/lowdosect/, Last Visited 22 Dec 2022.
80. lidc-dataset, Web: https://github.com/topics/lidc-dataset, Last Visited 22 Dec 2022.
81. How to join BRATS 2015: Brain Tumor Image Segmentation Challenge. Web: https://www.virtualskeleton.ch/BRATS/Start2015, Last Visited 22 Dec 2022.
82. A. A. A. Setio et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," Med. Image Anal., vol. 42, pp. 1–13, Dec. 2017.
83. N. Lung Screening Trial Research Team, "The national lung screening trial: Overview and study design," Radiology, vol. 258, no. 1, pp. 243–253, Jan. 2011.
84. APTOS 2019 Blindness Detection,https://www.kaggle.com/c/aptos2019-blindness-detection, Last Visited 22 Dec 2022.

85. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In AAAI, 2019.
86. J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "Covid-19 image data collection: Prospective predictions are the future," arXiv preprint arXiv:2006.11988, 2020.
87. J. Zhao, Y. Zhang, X. He, and P. Xie, "Covid-ct-dataset: a ct scan dataset about covid-19," arXiv preprint arXiv:2003.13865, 2020
88. Sivaswamy, J., Krishnadas, S., Chakravarty, A., Joshi, G., Tabish, A.S., et al.: A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomed. Imaging Data Papers 2(1), 1004.(2015)
89. Codella, N.C., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 168–172. IEEE (2018).
90. N. L. S. T. R. Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," New England Journal of Medicine, vol. 365, no. 5, pp. 395–409, 2011.
91. Xiaosong Wang et al., "Chestx-ray8: Hospital-scale chest xray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in CVPR, 2017, pp. 2097–2106.
92. Kaggle, Kaggle diabetic retinopathy detection challenge, 2015, (https://www.kaggle.com/c/diabetic-retinopathy-detection), Last Visited 23 Dec 2022.
93. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. Summers, ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, IEEE Conference on Computer Vision and Pattern Recognition (2017), pp. 3462-3471.
94. ISIC, The international skin imaging collaboration, 2019, (https://www.isic archive.com/), Last Visited 23 Dec 2022.
95. Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122–1131, 2018.
96. NLM - Malaria Data, https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html, Last visited 23 Dec 2022.
97. Scopus Database, https://www.scopus.com/sources.uri, Last Visited 2 Jan 2023.