mathematics

MDPI

Review

A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense

Gladys W. Muoka [1], Ding Yi [1,*], Chiagoziem C. Ukwuoma [2], Albert Mutale [1], Chukwuebuka J. Ejiyi [1], Asha Khamis Mzee [1], Emmanuel S. A. Gyarteng [3], Ali Alqahtani [4] and Mugahed A. Al-antari [5,*]

[1] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; muokagladys@std.uestc.edu.cn (G.W.M.)
[2] College of Nuclear Technology and Automation Engineering, Chengdu University of Technology, Chengdu 610059, China
[3] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[4] Center for Artificial Intelligence and Computer Science Department, King Khalid University, Abha 61421, Saudi Arabia
[5] Department of Artificial Intelligence, College of Software & Convergence Technology, Daeyang AI Center, Sejong University, Seoul 05006, Republic of Korea
* Correspondence: yi.ding@uestc.edu.cn (D.Y.); en.mualshz@sejong.ac.kr (M.A.A.-a.)

Abstract: Deep learning approaches have demonstrated great achievements in the field of computer-aided medical image analysis, improving the precision of diagnosis across a range of medical disorders. These developments have not, however, been immune to the appearance of adversarial attacks, creating the possibility of incorrect diagnosis with substantial clinical implications. Concurrently, the field has seen notable advancements in defending against such targeted adversary intrusions in deep medical diagnostic systems. In the context of medical image analysis, this article provides a comprehensive survey of current advancements in adversarial attacks and their accompanying defensive strategies. In addition, a comprehensive conceptual analysis is presented, including several adversarial attacks and defensive strategies designed for the interpretation of medical images. This survey, which draws on qualitative and quantitative findings, concludes with a thorough discussion of the problems with adversarial attack and defensive mechanisms that are unique to medical image analysis systems, opening up new directions for future research. We identified that the main problems with adversarial attack and defense in medical imaging include dataset and labeling, computational resources, robustness against target attacks, evaluation of transferability and adaptability, interpretability and explainability, real-time detection and response, and adversarial attacks in multi-modal fusion. The area of medical imaging adversarial attack and defensive mechanisms might move toward more secure, dependable, and therapeutically useful deep learning systems by filling in these research gaps and following these future objectives.

Keywords: medical image analysis; deep learning; adversary attack; adversarial defense; deep neural networks

MSC: 68T01

Citation: Muoka, G.W.; Yi, D.; Ukwuoma, C.C.; Mutale, A.; Ejiyi, C.J.; Mzee, A.K.; Gyarteng, E.S.A.; Alqahtani, A.; Al-antari, M.A. A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense. *Mathematics* **2023**, *11*, 4272. https://doi.org/10.3390/math11204272

Received: 6 September 2023
Revised: 27 September 2023
Accepted: 10 October 2023
Published: 13 October 2023

1. Introduction

Deep neural networks (DNNs) have achieved remarkable success in natural image-processing tasks. The field of medical image analysis is not left out [1–3], including skin lesion diagnosis [4], diabetic retinopathy detection, and tumor segmentation [5,6]. Notably, an AI-based diabetic retinopathy detection system [7,8] has been approved by the Food and Drug Administration (FDA) of the United States [9]. In addition to enhancing efficiency and patient outcomes, medical diagnosis models driven by deep learning can reduce

clinical costs. Nevertheless, despite its promise, deep learning is vulnerable to adversarial attacks [10,11], which can severely disrupt DNNs (Figure 1). These adversarial attacks can be generated by introducing imperceptible perturbations into legitimate samples, making them difficult to detect manually. Adversarial vulnerability poses a significant challenge for the application of DNNs in safety-critical scenarios such as medical image analysis [12–14], as it can result in misdiagnosis, insurance fraud, and a loss of confidence in AI-based medical technology.

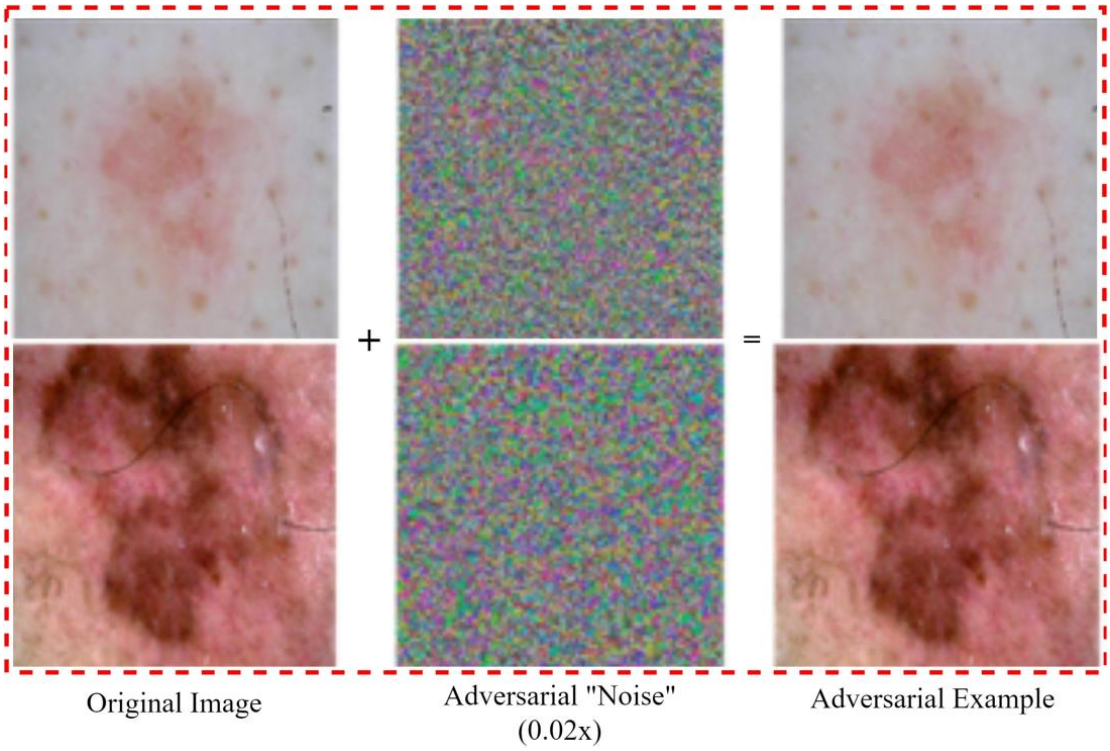

Figure 1. Examples of medical image adversarial attack.

The significance of robustness against adversarial attacks in medical image analysis was put forward by researchers who have analyzed the adversarial exposure of computer-aided diagnosis models from different perspectives [13,15,16]. As a result, much research has focused on defending against these adversarial attacks. Previous research focuses predominantly on adversarial training to improve network resilience and adversarial detection to identify adversarial attacks [17–19]. Some methods include image-level pre-processing [20] or feature enhancement techniques [21] within the adversarial defense context. These defense strategies have proven effective in establishing robustness against adversarial attacks for medical diagnosis's unified pattern recognition [22]. However, there is a substantial divide between the research-oriented settings and evaluations of various defense methods, making comparisons difficult. There exist several survey papers on adversarial attacks and defenses in medical image analysis. However, many focus on particular medical tasks or need a detailed taxonomy and exhaustive evaluation of existing attack and defense methods for computer-assisted diagnosis models. In addition, recent developments in adversarial attack and defense for medical image analysis systems still need to be adequately addressed.

Medical image analysis is essential to healthcare because it enables precise illness diagnosis, planning of treatments, and disease monitoring. However, as machine learning (ML) and artificial intelligence (AI) are used more often in the processing of medical images, concerns have been raised concerning the possible effects of adversarial assaults. The importance of minimizing false alarms through thorough image analysis was highlighted by Marinovich et al. in their article "Artificial Intelligence (AI) for breast cancer screening" [23].

They showed that an AI model designed for breast cancer diagnosis using mammograms had a false-positive rate of 3.5%. On the other side, a hospital in the US resolved a lawsuit in 2019 for USD 8.5 million after a radiologist failed to detect a tumor in a patient's CT scan [24]. A reliable second opinion may be provided via accurate AI-based image analysis, which lowers the likelihood of such mistakes. Zbrzezny et al.'s adversarial attack against a deep learning model used to diagnose diabetic retinopathy was proven in 2021 [25]. They deceived the model into incorrectly identifying the severity of the disease by introducing barely noticeable noise to retinal pictures. This demonstrates how susceptible AI healthcare systems are. A significant healthcare organization encountered a data breach in 2023 when nefarious individuals were able to obtain patient details, including medical photographs [26]. Personal health information abuse and identity theft may result from such breaches. Mobile applications with AI capabilities are being used to analyze medical images in several African nations, eliminating the need for radiology specialists and increasing diagnosis precision. For precise medical diagnosis and treatment choices, medical image analysis is essential. However, the vulnerability of AI systems to adversarial assaults and the possibility of human mistakes highlights the need for reliable and secure image analysis solutions to protect patient privacy and health.

The purpose of this research article is to address the aforementioned research problems by providing a systematic overview of recent advances in adversarial attack and defense for medical image analysis and discussing their benefits and limitations. In addition, we experimented with the widely known attack and defense adversarial strategy. The significant contribution of the manuscript is summarized as follows;

- This review article presents a, to date, comprehensive analysis of adversarial attack and defense strategies in the context of medical image analysis including the type of attack and defense tactics together with a cutting-edge medical analysis model for adversarial attack and defense.
- In addition, a comprehensive experiment was carried out to support the findings of this survey including classification and segmentation tasks.
- We conclude by identifying current issues and offering insightful recommendations for future research.

The introduction section is followed by the background of medical image analysis. This section is followed by Section 3, which talks about the overview of medical image adversarial attack and defense. In Section 3, more details of the medical adversarial attacks and defense are presented. Next, we have Section 4, where we carried out an extensive experiment to support our findings. The challenges and future works are presented in Section 5, while we conclude in Section 6. Figure 2 illustrates the PRISMA model strategy we used in this survey.

2. Review Background

2.1. Medical Image Analysis

Recently, deep learning has been applied in several domains extensively [27–29], of which the medical sector is not left out. It may also be applied to the development of new medications, medical decision-making, and novel approaches to the creation of various forms of medicine [30,31]. Computerized clinical results heavily depend on medical imaging such as X-rays, ultrasound, computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI) as they are further examined by professionals or radiologists [32,33]. The main purpose of the processing of medical images is to make the data represented more comprehensible [34]. Medical image analysis must be conducted precisely and quickly since any delay or incorrect diagnosis can be harmful to a patient's health [35]. Robotic deep learning is required to attain this high accuracy and quickness. Finding out which parts of the body are impacted by the illness is the main goal of medical image interpretation to help doctors understand how lesions grow. Four steps—image preprocessing, segmentation, feature extraction, and pattern detection and classification—make up most of the examination of a medical image [36–38]. Preprocessing

is used to fix undesired image defects or to enhance image data for later processing. The method of segregating areas, such as tumors and organs, for additional research, is referred to as segmentation. Feature extraction is a means of removing specific information from regions of interest (ROIs) to help in their identification [39,40]. The categorization of the ROI is aided by classification based on extracted characteristics.

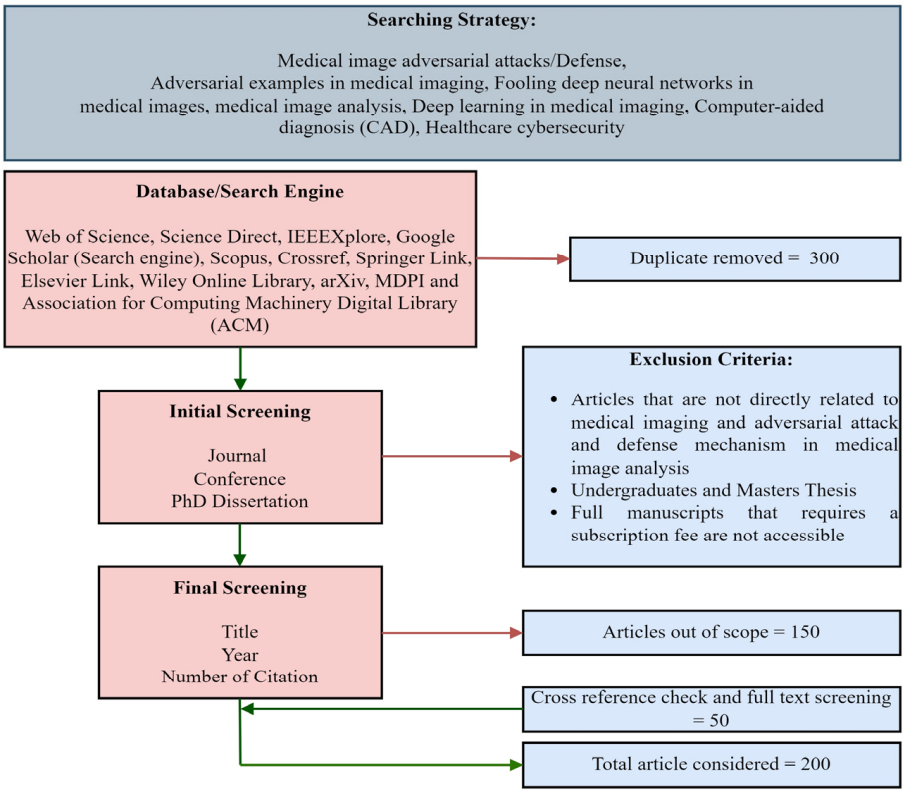

Figure 2. Paper selection using the PRISMA model strategy.

The deep learning concept has significantly advanced several AI disciplines, yet it is still open to severe security risks. The adversarial attack has received the most focus from the deep learning risk sector because it highlights several possible security issues for deep learning applications. Due to the adversarial attacks' optical similarity to its pure version, adversarial attacks can circumvent human inspection in addition to interfering with DNNs' interpretation process. Computer-aided diagnostic systems may significantly amplify this pernicious privacy risk, which might lead to fatal misinterpretation and possibly a crisis of societal credibility [13]. Taking into account a certain dataset $(x, y) \sim$ "D", where "D" depicts data distribution across pairs of provided samples X and their accompanying labels Y. We represent the medical analysis deep learning model as $f\theta(\cdot)$ with θ as the network parameter. An undetectable noise δ is frequently added to clean instances X to produce adversarial attacks $\hat{X}$, which are technically described as follows:

$$\hat{X} := X + \delta \text{ with } f_\theta(\hat{X}) \neq Y \text{ and } d(X, \hat{X}) \leq \epsilon \tag{1}$$

where δ is the highest permitted perturbation range for subtlety and is the dimension metric $d(.,.)$. By definition, adversarial samples $\hat{X}$ must be near their true equivalents X using an established measure, such as the $\updownarrow_p$ distance. The adversarial perturbation is as $\updownarrow_p$ norm bound as $\| \delta \|_p \leq \epsilon$. The sub-infinity and adversarial $\updownarrow_\infty$ norm threat model is represented in Equation (2):

$$\max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}(f_\theta(X + \delta), Y), \tag{2}$$

where the $\mathcal{L}$ largely depends on the actual task (segmentation, detection, or classification). Equation (2) can be integrated using Newton-like approaches [14,41] or algorithms based on gradient descent. Other white-box adversarial attack techniques include the Limited-memory BFGS method [14], the Fast Gradient Sign Method (FGSM) [42], etc. There are also additional threat models, such as the black-box attack, which in real-life situations presents a bigger security risk to computer-aided diagnostic models [12,43].

Several adversarial defense mechanisms have been established to defend deep learning models from adversarial attacks [38,39,44]. The most popular among them is adversarial training [42,45], which can increase inherent network resilience by supplementing adversarial cases as training data. At the inference step, the adversarial learned model is anticipated to foresee both adversarial samples accurately. Based on Equation (2), the conventional adversarial training [21] may be expanded to become the following min–max optimization formation:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|\infty\leq\epsilon} \mathcal{L}\left(f_{\theta(X+\delta),\, y}\right) \right] \tag{3}$$

To interfere with the target network, internal optimization seeks out the most harmful and threatening occurrences. Improving the realistic adversarial threat spanning network parameters is the primary focus of external mitigation. In general, adversarial training improves deep neural networks' (DNNs') intrinsic robustness without adding any extraneous parts, while preserving its ability to make accurate inferences from valid data. Different protection strategies focus on pre-processing data (both clean and hostile instances) without influencing later computer-aided analysis networks, strengthening the intrinsic network resistance against adversarial components [20,46]. In simple terms, the data pre-processing aims to retain the original form of clean inputs while converting hostile samples into benign equivalents for future inference. The resulting optimization problem allows us to create a pre-processing-based protection.

$$\min_{\psi} \mathbb{E}_{(X,Y)\sim\mathcal{D}} [\mathcal{L}(f_{\theta}(\psi(X+\delta)), Y) + \lambda \cdot \mathcal{L}(f_{\theta}(\psi(X)), Y)] \tag{4}$$

The pre-processing module ψ, which may include a modal or irregular operator intended to mitigate the effects of adversarial perturbations, is indicated by the weight coefficient λ. Nevertheless, distinguishable areas in medical imaging typically cover a small number of pixels. Relative to their equivalents in environmental images, biased characteristics still run a higher risk of being lost through pre-processing methods for medical images. Many adversarial attack and defensive strategies have demonstrated outstanding results with real-world images [47]. The tasks belonging to normal vision alongside those relevant to medical imaging nevertheless vary fundamentally in several ways, including data aspects, features, and task features. As a result, it is difficult to directly transfer adversarial attack and defensive strategies from the realm of natural imagery to the field of medicine. In addition, several research studies have shown that compared to natural images, medical images may be much more vulnerable to serious adversarial attacks [13,17,43]. The safety and reliability of computer-aided diagnostic models have to be considered carefully consideration, considering the sizeable healthcare industry and the significant effect of computer-aided diagnosis. This paper, therefore, provides a thorough overview of current developments in adversarial attack and response strategies in the field of medical image analysis.

Contemporary healthcare systems are built around pattern detection and classification in medical image analysis. Using cutting-edge technology to glean priceless information from medical imaging like X-rays, MRIs, CT scans, and ultrasounds is the central idea behind this multidimensional approach. The detection, categorization, and management of numerous medical diseases are made possible by these images, which are frequently complicated and filled with minute details. They provide essential insights into the inner workings of the human body. Fundamentally, pattern recognition is carefully examining

these images to find abnormalities, irregularities, or certain traits that might point to a condition, an injury, or other health-related issues. On the contrary, classification goes beyond pattern detection by classifying the discovered patterns into distinct groups or diagnoses. Artificial intelligence and machine learning algorithms are crucial in this regard because they can analyze enormous volumes of picture data and spot patterns that the human eye would miss [48,49]. Effective pattern recognition and classification have significant effects on medical image analysis. They make it possible to identify diseases early, arrange effective treatments, and track a patient's development over time. They also help to improve overall patient satisfaction and lower the margin of diagnosing error. Researchers and healthcare professionals are constantly working to improve and grow the methods and tools used for pattern detection and classification in this dynamic field, pushing the limits of the analysis of medical images and redefining how diseases are identified and treated. This overview just touches the surface of this important field, which is at the vanguard of contemporary medicine and provides patients everywhere with hope and innovation.

2.2. Adversarial Attack and Defense

Deep neural network (DNN) vulnerability has not been addressed or justified technically. DNN uses a substantial quantity of ambient input while training and derives conclusions from its internal framework and algorithmic process via its outcome. Szegedy et al. [9] first initiated the susceptibility of deep neural network models in image classification. Human eyes cannot tell the difference between the adversarial samples produced after applying a perturbation to the initial image, yet it was wrongly predicted by the employed deep learning network. Given the possible risk, real-world adversarial attacks on deep learning models raise serious concerns. In a noteworthy investigation, Eykholt et al. [50] perturbed an actual traffic signal with black and white graphics, leading to deliberate misclassification, to show the viability of strong physical-world attacks. These attacks are especially dangerous because they may resist a wide range of physical circumstances, such as shifting perspectives, miles, and qualities. This demonstrates how adversaries might potentially modify real-world items like traffic signs using deep learning algorithms, resulting in misunderstanding. Such attacks can trick machines that are autonomous, like self-driving vehicles, and have disastrous effects on the road, which has serious ramifications for security and privacy.

Real-world adversarial attacks on deep learning algorithms raise serious issues and can even be fatal, especially in the medical industry. Attack-related medical image misclassification might result in improper or late therapy, which could put people's lives in peril. Recognizing the need to protect ML models in the medical field and creating strong defensive measures to reduce the dangers brought on by adversarial attacks are essential. Ma et al.'s [17] work concentrated on adversarial attacks on deep learning-based algorithms for analyzing medical images. Due to the special properties of medical image information and deep neural network models, their research on baseline medical imagery databases demonstrated that malicious attacks on medical images are simpler to design. They additionally showed how medical adversarial scenarios frequently target tissues outside of diseased ones, leading to deeply distinctive and distinguishable traits. In the same vein, Paul et al. [47] built an ensemble-based defense mechanism and examined the effect of adversarial attacks on the precision of forecasting lung nodule cancer. To increase resilience over adversarial attacks, they also looked into including hostile images in the training sample. The accuracy of CNN prediction was considerably impacted by adversarial attacks, notably the Fast Gradient Sign Method (FGSM) and one-pixel attacks. They outline that training adversarial images and multi-initialization ensembles can boost classification accuracy. Ozbulak et al. [51] combined the skin lesions and glaucoma optic disc in a segmentation task while examining the effect of adversarial attacks on deep learning models. Furthermore, they developed an Adaptive Mask Segmentation Attack, an innovative algorithm that yields an adversarial attack with accurate prediction masks based

on perturbations, which are largely invisible to the human eye yet cause misclassification. This study shows the necessity for strong defenses to guarantee the dependability and accuracy of segmentation models used in healthcare settings.

3. Overview of Medical Image Adversarial Attack and Defense

This survey presents a comprehensive exploration into the realm of medical image adversarial attack and defense. Within this section, a detailed dissection of medical adversarial attacks is offered, alongside an in-depth investigation of defense strategies. This discourse will delve into the intricacies of both attack and defense methodologies within the context of medical image analysis. The accompanying Figure 3 provides a succinct visual representation of the yearly publication trend about adversarial attack and defense within the domain of medical image analysis.

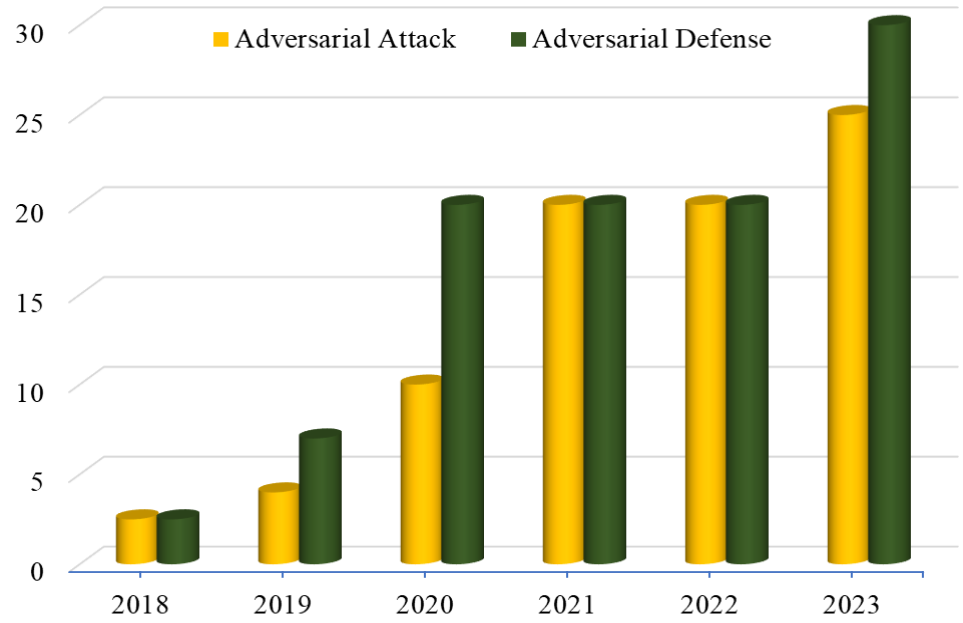

Figure 3. The number of papers published per year related to adversarial attack and defense for medical image analysis.

A. Medical Image Adversarial Attack and Defense Classification Task

The classification process simply explains the partition of medical images into discrete groups, typically involving the identification of various conditions or ailments, as shown in Figure 4. The objective of adversarial attacks targeting classification tasks is to strategically modify input images to elicit inaccurate classification. The decision-making mechanism of the model can be misled by perpetrators through the introduction of varying modifications that are difficult to detect. The consequences of effective classification attacks are significant, as they can result in erroneous treatment approaches, postponed interventions, and compromised well-being of patients. Paschalis et al. [12] have used an innovative methodology to assess the resilience of deep learning networks in the context of medical imaging. The researchers examine weaknesses in these state-of-the-art networks through the utilization of adversarial instances. Different techniques were utilized for classification, such as FGSM, Deep Fool (DF) [52], and saliency map attacks (SMA) [53]. On the other hand, dense adversarial generation (DAG) [54] was applied for semantic segmentation, with varying levels of perturbation and complexity. The results of the study indicate that images affected by noise are categorized in a manner comparable to clean images during classification tasks.

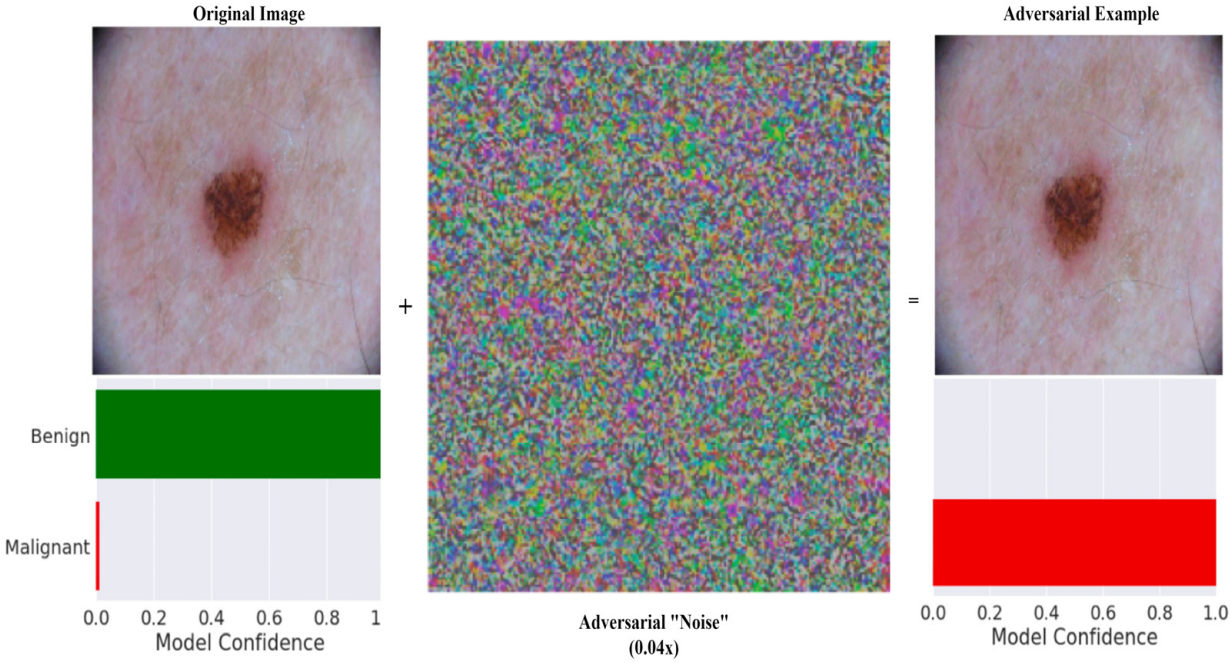

Figure 4. Example of adversary attack in a classification task scenario.

However, adversarial instances exhibit a unique tendency to be classified into other categories. The presence of Gaussian noise resulted in a decrease in the level of certainty in classification, while the majority of adversarial attacks exhibited a high level of certainty in their misclassification. Therefore, it may be argued that adversarial instances are better suitable for evaluating the resilience of models in comparison to test images that contain noise. Finlayson et al. [55] conducted a comprehensive investigation aimed at identifying vulnerabilities within deep neural network models used in the medical field. Both Pretrained Gradient Descent (PGD) and a basic patch attack were employed on three foundational models for the classification of medical disorders such as diabetic retinopathy, pneumothorax, and melanoma. The findings suggest that both forms of attacks have a high probability of success and are undetectable by human observers. Furthermore, these attacks seem to be effective even against advanced medical classifiers, especially the ResNet-50 model. It is important to note that the attacker's degree of network exposure does not significantly impact the efficacy of these attacks.

Adversarial attacks such as the Fast Gradient Sign Method (FGSM), the Basic Iterative Method (BIM) [56], and C&W attacks, were utilized in the identical medical setting as described in [55], with a specific emphasis on fundoscopy, chest X-ray, and dermoscopy. The aforementioned attacks were implemented on datasets that encompassed both two-class and multi-class classifications. Taghanaki et al. [57] conducted a thorough investigation to evaluate the susceptibilities of deep learning techniques in categorizing chest X-ray images across different illness categories. The researchers thoroughly analyzed the performance of two deep neural networks when subjected to 10 diverse adversarial attacks. In contrast to previous methodologies that employed a singular gradient-based attack, the researchers examined several gradient-based, score-based, and decision-based attack models. These models were examined on Inception-ResNetv2 and NasNet-large architectures, with their performance evaluated using chest X-ray images. To perform a more comprehensive examination of the vulnerabilities of convolutional neural networks (CNNs), Yilmaz et al. [58] undertook a groundbreaking experiment to assess the sensitivity of a classifier designed for mammographic images to adversarial attacks. The researchers analyzed the similarity between benign and malicious images utilizing the structural similarity index technique (SSIM), which is a perceptual model employed for assessing image similarity. Furthermore, Fast Gradient Sign Method (FGSM) attacks were implemented on convolutional neural

networks (CNNs) that have undergone training. In recent times, a category of sophisticated attacks referred to as universal adversarial perturbations (UAP) has been presented [52]. These attacks encompass perturbations that are agnostic to visual content, hence providing enhanced realism and effectiveness. The proposed approach utilizes an iterative algorithm to apply slight perturbations to input images.

The use of chest X-ray images to gain an understanding of different ailments kinds has generated significant attention among physicians and radiologists due to the potential for automated assessment enabled by deep learning networks [59,60]. As a result, the issue of safeguarding the integrity of these models has emerged as a matter of utmost importance. Rao et al. [61] conducted a comprehensive investigation of various attack and defense strategies employed in the classification of Thorax disorders by the analysis of chest X-rays. The scholars conducted a comparison assessment whereby they examined five distinct attack types, specifically DAA, DII-FGSM, MIFGSM, FGSM, and PG. Prior research has demonstrated that deep learning networks utilized in the prediction of vulnerability to COVID-19 are vulnerable to adversarial attacks. In their seminal work, Rahman et al. [62] conducted a comprehensive analysis of the effects of adversarial perturbations on deep learning networks, establishing themselves as pioneers in this field. The scope of their analysis included six discrete deep-learning applications specifically designed to diagnose COVID-19. Additionally, the researchers incorporated multi-modal adversarial instances into several diagnostic algorithms for COVID-19. The decline in image quality observed in these instances frequently arises from the presence of irregular illumination. In light of this concern, Cheng et al. [63] addressed the matter by using an adversarial attack technique. The researchers presented a new type of attack known as the "adversarial exposure attack". The methodology employed by the researchers entails the creation of hostile images by the manipulation of image exposure to radiation, intending to mislead the deep neural networks that underlie the image recognition system.

B. Medical Image Adversarial Attack and Defense Segmentation Task

Segmentation attacks can be summarized as attacks involving the deliberate division or fragmentation of data or information to compromise its integrity, confidentiality, or availability, as shown in Figure 5. Segmentation plays a crucial function in delineating distinct regions of interest within medical images, facilitating the detection and characterization of abnormalities, organs, or malignancies. The focus of adversarial attacks in the context of segmentation tasks revolves around the manipulation of pixel values or gradients to disturb the accurate delineation of boundaries.

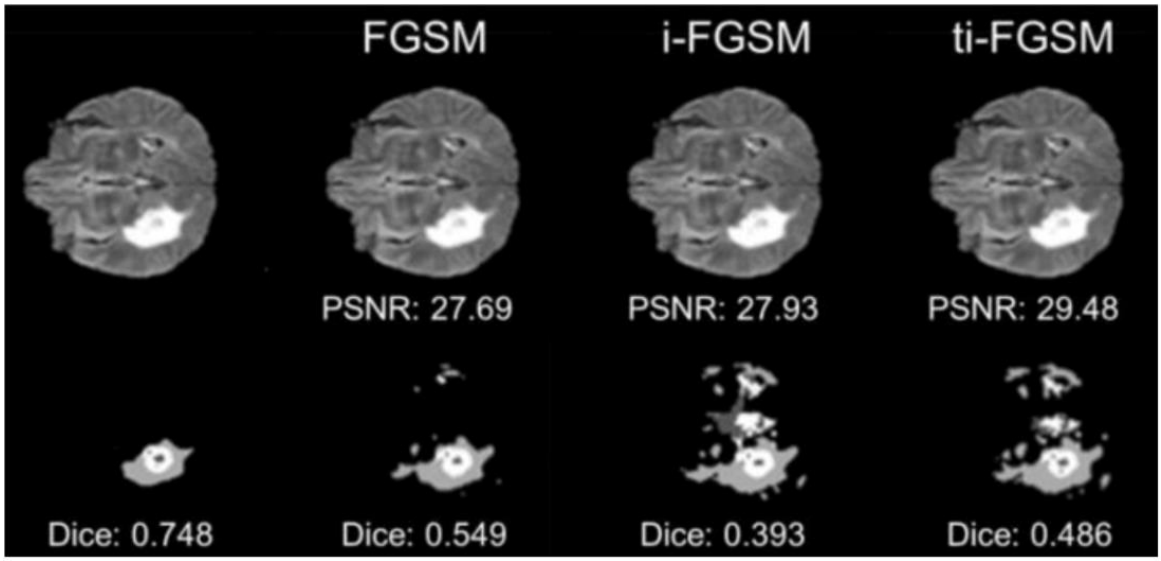

Figure 5. Example of adversary attack in a segmentation task scenario.

The vulnerability of segmentation models to adversarial attacks highlights the necessity of implementing robust protections to guarantee the accurate representation of anatomical structures. Therefore, it is crucial to develop various ways to create adversarial cases that can specifically target segmentation models. To fulfill this aim, Chen et al. [64] proposed a novel approach to attack segmentation convolutional neural networks (CNNs) through the utilization of adversarial learning. The methodology employed by the researchers entails integrating a variational auto-encoder (VAE) with Generative Adversarial Networks (GAN) to produce images that demonstrate deviations and alterations in visual characteristics. The purpose of these images is to subvert the effectiveness of medical segmentation models. The quantification of the attack effect is determined by a significant decrease in the Dice score [65], a commonly employed metric for evaluating the effectiveness of segmentation in medical imaging, as matched with actual truth segmentation. To prioritize the protection of medical neural networks for the benefit of patients, Cheng and Ji [63] conducted a study to examine the effects of universal adversarial perturbations on brain tumor segmentation models across four different modalities. The researchers utilized the MICCAI BraTS dataset, which is recognized as the most extensive publicly available compilation of MRI brain tumor images, and implemented them in a U-Net model. The perturbations were created using a Gaussian distribution.

C. Medical Image Adversarial Attack and Defense Detection Task

Detection attacks refer to a type of cyber-attack where an adversary attempts to evade detection of security systems or protocols. Detection tasks involve the process of identifying particular objects or anomalies within medical images, such as accurately determining the existence of cancers or lesions, as shown in Figure 6. The objective of adversarial attacks in detection tasks is to intentionally alter or confuse the visual characteristics of specific objects to circumvent the detection capabilities of the model. By making subtle alterations to the characteristics of an anomaly, individuals with malicious intent can avoid being detected, which may result in occurrences of false negatives. In line with the above, Ma et al. [17] investigated this particular issue by using four detection approaches on medical deep neural networks. These methods include Kernel density (KD) [66], Deep features (DFeat), local intrinsic dimensionality (LID) [67], and quantized features (QFeat) [68]. Li and Zhu [69] proposed an unsupervised learning method as a means to detect adversarial attacks on medical imaging. The authors posited that their unique methodology has the potential to operate as a self-contained component within any medical imaging system based on deep learning, hence augmenting the system's resilience. Li et al. [70] developed a hybrid approach to create a robust artificial intelligence framework for medical imaging in their research.

The architecture presented in this study is founded on the principles of semi-supervised adversarial training, unsupervised adversarial detection, and the introduction of a novel metric for evaluating the susceptibility of the system to adversarial risks (SSAT and UAD). The technique proposed by the authors effectively tackles two primary obstacles encountered in the identification of adversarial samples. These challenges are the scarcity of labeled images in medical applications and the ineffectiveness of current detection approaches when faced with fresh and previously undetected attacks.

3.1. Medical Imaging Adversarial Attacks

In the field of medical image analysis, the integration of machine learning techniques, particularly deep neural networks, has resulted in a significant improvement in both diagnostic accuracy and the development of therapeutic approaches [10,22]. However, this progress has also brought about new challenges, evident in adversarial attacks. These attacks involve deliberately injecting carefully engineered distortions into input data to deceive the models. Within the field of medical image analysis, adversarial attacks pose a significant threat to the reliability and integrity of tasks related to classification, segmentation, and detection. These tasks are crucial for accurate diagnosis and effective patient care.

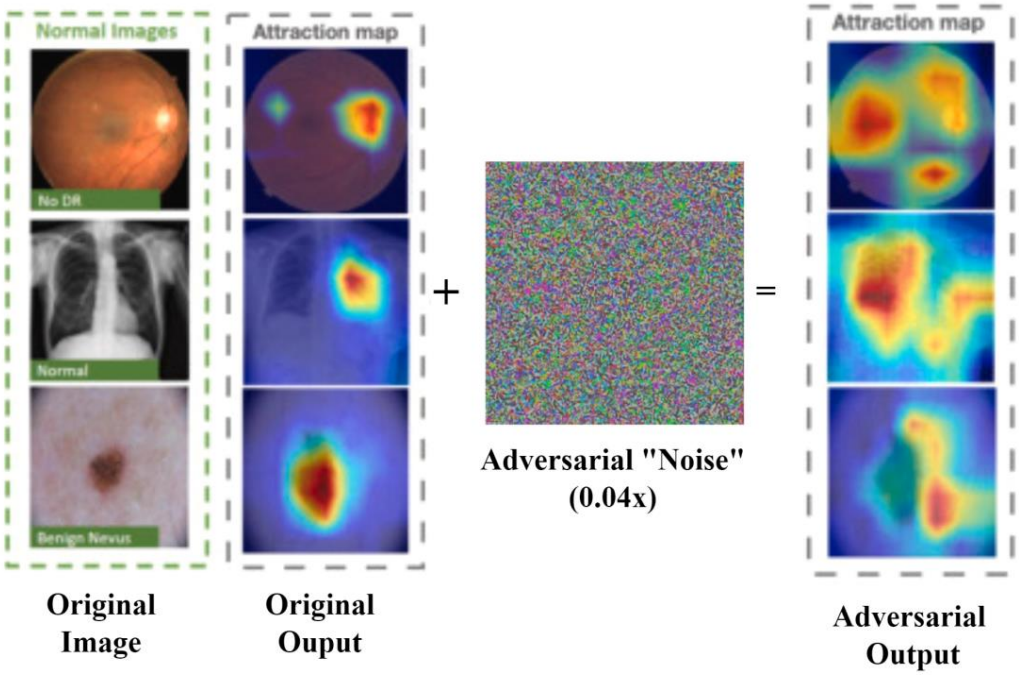

Figure 6. Example of adversary attack in a detection task scenario.

3.1.1. White Box Attacks

Within the domain of adversarial machine learning, a notable category of attacks is sometimes referred to as "white box attacks". The aforementioned attacks exhibit a distinguishing feature wherein the perpetrator possesses an in-depth understanding of the architecture, parameters, and training data of the specific machine learning model being targeted. The thorough comprehension of the underlying workings of the model allows adversaries to create intricate and tailored adversarial attacks with the specific goal of exploiting weaknesses and inducing misclassification or erroneous outputs. White-box attacks have the most potential to cause disruption. As a result, they are frequently used to evaluate the effectiveness of associated defenses as well as the resilience of machine learning and deep learning models. Zhang et al.'s key work [71] pioneered the proof that even the smallest unnoticeable perturbations placed into an image might cause a DNN to misclassify an image. The authors set out to solve the equation governing the smallest perturbation required to cause a neural network misclassification. The C&W attack technique [72] utilizes L0, L2, and L∞ norms to generate adversarial samples while adhering to specified perturbation restrictions. This algorithm is supported by an improved version of the L-BFGS optimization methodology (one of the strongest existing methods for targeted attacks). The authors of reference [73] developed FGSM, a technique for quickly calculating adversarial perturbations. An adjustment to the adversarial perturbation magnitude in the FGSM aligns with the gradient of the loss function of the model. This method makes it easier for untargeted attacks to produce misclassification by manipulating the gradient of the model's loss function with respect to the input. The I-FGSM method was improved by [74] by focusing on the category with the lowest confidence level. This method of targeting creates adversarial samples that trick the model into categorizing segments that are vastly different from the correct one, increasing the attack's potential for disruption. Likewise, in [53], a technique was put out to limit the L0 norm perturbation by changing only a small subset of the image pixels as opposed to the complete image. This method uses a saliency map and is greedy, repeatedly changing one pixel at a time while computing the derivative of the output bias for each input feature at the model's final layer. With this technique, a subtle and targeted perturbation strategy is introduced (Table 1).

Table 1. Summary of adversarial attack works in the context of medical image analysis.

Ref./Year	Adversarial Attack Type	Task	Image Modality
[12]/2018	FGSM, Deep pool, JSMA	Classification/segmentation	MRI, Dermoscopy
[75]/2018	FGSM	Classification	Fundoscopy
[13]/2018	PGD, AdvPatch	Classification	Fundoscopy, X-ray, Dermoscopy
[76]/2019	FGSM, I-FGSM, TI-FGSM	Segmentation	MRI
[77]/2019	Multi-task VAE	Segmentation	CT
[51]/2019	Adaptive segmentation mask Attack	Segmentation	Fundoscopy, Dermoscopy
[78]/2019	PDG	Classification	X-ray, Histology
[61]/2020	FGSM, PGD, MI-FGSM, DAA, DII-FGSM	Classification	X-Ray
[79]/2020	Adversarial exposure attack	Classification	Fundoscopy
[62]/2020	BIM, L-BFGS, PGD, JSMA	Classification, Object detection	CT, X-ray
[80]/2020	Zoo	Classification	Ultrasound
[81]/2020	Adaptive targeted I-FGSM	Landmark detection	MRI, X-ray
[82]/2020	FGSM	Classification	Fundoscopy
[83]/2021	UAP	Classification	OCT, X-ray, Dermoscopy
[84]/2021	FGSM, BIM, PGD	Classification	CT, MRI, X-ray
[85]/2021	IND and OOD Attacks	Segmentation	MRI
[86]/2021	Stabilized medical image attack	Classification, Segmentation	CT, Endoscopy, Fundoscopy
[87]/2021	PGD	Segmentation	X-Ray
[88]/2021	CW	Classification	CT, X-ray, Microscopy
[89]/2021	FGSM	Classification	CT, X-ray
[90]/2021	Multi-scale attack	Segmentation	Fundoscopy, Dermoscopy
[91]/2022	AmdGAN	Classification	CT, OCT, X-ray, Fundoscopy, Dermoscopy, Ultrasound, Microscopy
[92]/2022	UAP	Classification	X-ray, Fundoscopy, Dermoscopy
[93]/2022	Attention-based I-FGSM	Classification	CT
[94]/2022	Modified FGSM with with tricks to break defenses	Classification, Segmentation	CT, MRI, X-ray, Dermoscopy, Fundoscopy
[95]/2022	FGSM	Segmentation	Microscopy
[96]/2022	FGSM	Classification	X-ray
[97]/2022	Digital watermarking	Classification	CT, MRI, X-ray
[98]/2022	FGSM, BIM, PGD, No-sign operation	Classification	X-Ray
[99]/2022	FGSM	Classification	Fundoscopy
[100]/2022	FGSM, PGD, CW	Classification	CT, Dermoscopy, Microscopy
[101]/2022	Improved adaptive square attack	Segmentation	X-Ray
[43]/2022	FGSM, BIM, PGD, MI-FGSM	Classification	CT, Fundoscopy
[102]/2022	Adversarial k-space noise, Adversarial rotation	Reconstruction	MRI
[103]/2022	FGSM, L-BFGS	Classification	Fundoscopy
[15]/2022	Feature space-restricted attention attack	Classification	X-ray, Fundoscopy, Dermoscopy
[104]/2023	FGSM, PGD	Classification	Classification
[105]/2023	FGSM	Classification	CT
[106]/2023	PGD, FGSM, BIM, GN	Segmentation	CT, MRI
[107]/2023	PGD, BIM, FGSM	Classification, Detection	CT, MRI
[108]/2023	PDG, CW, BIM	Classification	X-ray, Fundoscopy, Dermoscopy
[109]/2023	FGSM, MI-FGSM, PDG, CW	Reconstruction	X-ray, Fundoscopy
[110]/2023	L-BFGS, FGSM, PDG, CW	Classification, Reconstruction	MNIST
[111]/2023	FGSM, PGD, MIM, CW	Classification, Segmentation	X-Ray, CT, Ultrasound

3.1.2. Black-Box Attacks

Black-box attacks are instances in which the perpetrator of the attack has minimal or no understanding of the internal operations of the specific machine learning model

being targeted. As opposed to existing white-box adversarial attacks that primarily rely on obtaining multiple backward gradients of target models, the attacker in this approach considers the target DNN as the locally installed model for generating related adversarial samples. When considering the merits of different scenarios, it might be argued that the general black-box scenario offers a more appropriate environment for simulating practical adversarial attacks. In essence, the attacker regards the model as an opaque entity, with solely the input–output characteristics of the model being accessible. Their only knowledge is the capacity to see the system's outputs in response to particular inputs. However, there are occasions when this capability is limited, frequently as a result of restrictions placed on the number of requests to avoid raising suspicion, which makes their efforts more difficult. The works of [14] show that, even if two models are trained on distinct datasets, adversarial attacks designed to cause misclassification in one model may equally fool another machine, which explains the transferability of adversarial attacks. Making use of this idea, the authors of [112] launched black-box attacks against a DNN intending to develop a replacement model that is trained on artificial inputs that the target DNN labeled. This alternative model was used to provide adversarial instances that might damage the intended DNN. Ref. [113] also provided three realistic threat models that were in line with actual world circumstances within the scope of black-box scenarios. These consist of label-only, incomplete information, and query-limited options. Query-efficient methods are required by the query-limited model, which uses Natural Evolutionary Strategies to predict gradients for implementing the PGD attack. The method alternates between merging the original image and maximizing the probability of the intended target class in cases where only the top-k label probabilities are given. Similar to this, the attack makes use of noise robustness to create targeted attacks when the attacker only has access to the top-k predicted labels. The Invariant Feature Transform is used in the FGBB approach [114], which makes use of image-extracted features to direct the development of adversarial perturbations. Higher probabilities are given by the algorithm to pixels that significantly affect how the human visual system perceives objects. As a two-person game, the process of creating an adversarial example is conceptualized, with one player limiting the distance to an adversarial example and the other taking on different roles that produce the fewest possible antagonistic examples. Square Attack [115] is a novel method that functions without relying on local gradient information and is hence resistant to gradient masking. This attack chooses localized square-shaped updates in arbitrary positions using an algorithmic search mechanism. The perturbation strategically lines up with the decision boundaries as a result. Figure 7 depicts the summary of adversarial attack taxonomy in medical imaging.

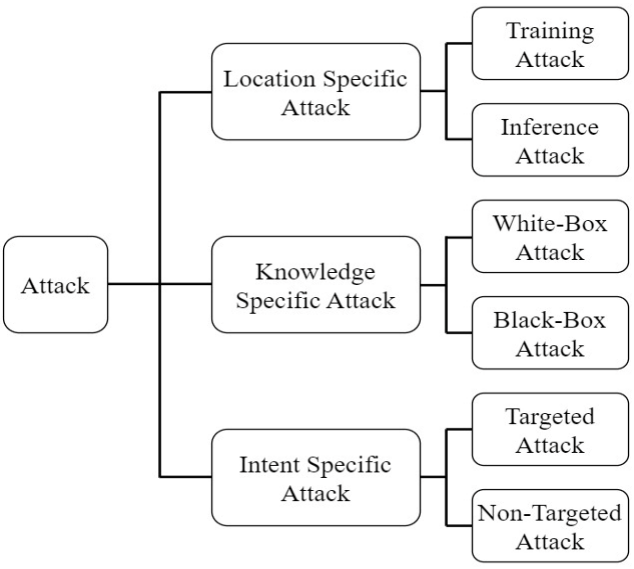

Figure 7. Taxonomy of medical image adversary attack.

3.2. Medical Adversarial Defense

In light of the potential risks, numerous defense approaches have been proposed to counteract medical adversarial attacks. This research analyzed a range of tactics and methodologies to reduce the adverse effects of adversarial attacks on the processing of medical images. Specifically, we offer a comprehensive examination of each category of adversarial defense technique, encompassing its prerequisites, limitations, and results. The mitigation of adversarial attacks in medical image analysis necessitates adopting a comprehensive strategy that integrates sophisticated methodologies derived from machine learning and medical imaging. The prioritization of the creation and integration of effective adversarial defenses is crucial within the healthcare industry as it increasingly adopts AI-powered solutions. This is necessary to protect patient safety, to uphold the accuracy of diagnoses, and to guarantee AI's ethical and secure utilization in medical applications. The advancement of the area of medical adversarial defense and the establishment of a robust basis for AI-driven healthcare will heavily rely on the collaborative endeavors of researchers, practitioners, and policymakers on Adversarial defenses. Figure 8 depicts the summary of adversarial defense in medical imaging.

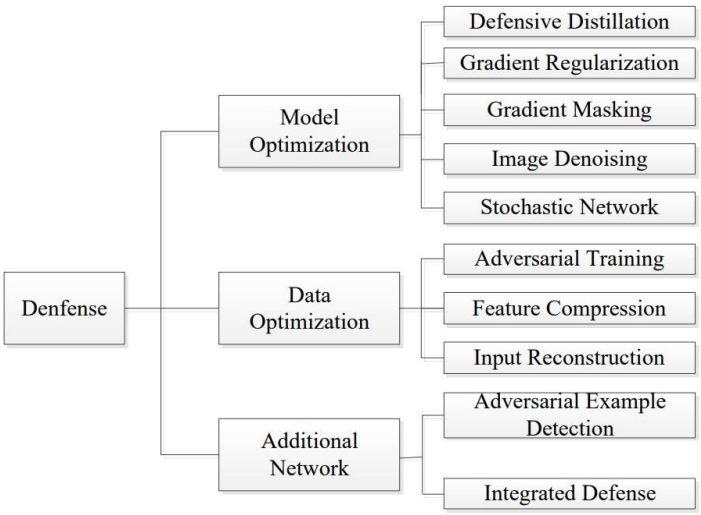

Figure 8. Taxonomy of medical image adversary defense.

3.2.1. Image Level Preprocessing

Images need to be processed before they can be utilized for model training and inference. This encompasses but is not restricted to changes in color, size, and direction, as shown in Figure 9. Pre-processing is carried out to improve the image's quality so we can analyze it with greater accuracy. Through preprocessing, we may eliminate undesired deformities and enhance certain properties crucial for the application we are developing. Those qualities might change based on the application.

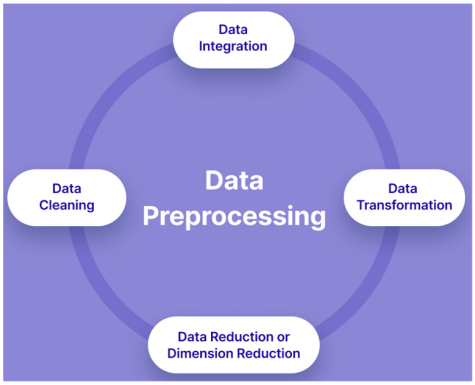

Figure 9. Basic steps of image preprocessing.

Typically, an adversarial image is composed of a pristine image and its matching adversarial alteration. Meanwhile, deep neural networks (DNNs) have been shown to attain impressive performance for clean images, but they are still vulnerable to adversarial attacks [14,116]. Therefore, the process of reducing the perturbation component from the adversarial example can enhance the subsequent network diagnosis. Moreover, there is no necessity to undergo retraining or make alterations to medical models when implementing image-level preprocessing techniques. This approach proves to be advantageous and secure in the domain of biomedical image analysis. Most pre-processing-based defense approaches primarily focus on medical classification tasks, as evidenced by the majority of existing research [61,117]. Therefore, it is necessary to do image-level preprocessing to preserve the identifiable components for further diagnostic purposes. The authors suggest the Medical Retrain-less Diagnostic Framework (MedRDF) [20] as a method to enhance the robustness of a pre-trained diagnosis model during the inference step. In the initial step, MedRDF generates several replicas of the input images, each of which is subjected to isotropic noise perturbations. The generation of these duplicates is anticipated by the utilization of majority voting after the application of a bespoke denoising algorithm. Furthermore, a comprehensive metric is proposed to determine the confidence level of MedRDF diagnosis, aiding healthcare professionals in their clinical practice. Kansal et al. [46] expanded upon the High-level representation Guided Denoiser (HGD) [118] to defend medical picture applications against hostile examples in both white-box and black-box scenarios instead of focusing just on pixel-level denoising. The incorporation of high-level information can enhance the process of eliminating the adversarial effect at the image level, leading to a more accurate final diagnosis without causing visual disruption.

3.2.2. Feature Enhancement

Feature enhancement, a powerful strategy in the field of machine learning, holds substantial potential when utilized as a method to mitigate adversarial risks. Adversarial attacks, which are defined by their ability to introduce subtle modifications to input data, present significant difficulties to the resilience of machine learning models. The utilization of feature augmentation approaches within the context of adversarial defense involves transforming and enhancing data representations to strengthen the model's ability to resist these attacks. By incorporating decision-influencing features, models are strengthened with enhanced capacities to differentiate genuine patterns from adversary interference, hence bolstering their effectiveness and reliability. Numerous approaches for enhancing features have been developed to improve the resilience of medical classification models [119–121]. Pooling layers are frequently employed in neural network modeling as a means to decrease the size of feature maps. In their study, Taghanaki et al. [21] made modifications to the medical classification networks by substituting max-pooling layers with average-pooling layers. One possible explanation for the observed improvement in resilience is that the utilization of average-pooling allows for the acquisition of a more incredible amount of contextual information at a global level, as opposed to the selection of only the largest value in max-pooling. This increased contextual understanding presents a greater challenge for adversarial attacks. Additionally, the Auto Encoder (AE) can be integrated into computer-aided diagnosis models to perform feature-level denoising [122]. This denoising process is separate from the image-level preprocessing technique. Meanwhile, the incorporation of feature invariance guidance is employed to mitigate the model's susceptibility to adversarial attacks. In line with the findings of reference [123], Han et al. [120] also incorporated dual-batch normalization into adversarial training. This modification resulted in a notable enhancement in the resilience of diagnostic models without compromising their accuracy under clean conditions.

In addition to the conventional medical classification problem, feature-enhanced methods have been utilized in various medical imaging tasks, such as segmentation [124], object identification [21], and low-level vision [125]. The Non-Local Context Encoder (NLCE) [126] is a proposed module that aims to enhance the resilience of biomedical image

segmentation models. It functions as a plug-and-play component. The NLCE module, as noted by [21], is designed to capture spatial dependencies at a global level and enhance features by including contextual information. This module may be seamlessly integrated into different deep neural network (DNN) models for medical picture segmentation. In their study, Stimpel et al. [125] employ the guided filter technique along with a guidance map that is learned to enhance the resolution and reduce noise in medical images. The guided filter demonstrates a robust capability in mitigating the impact of adversarial attacks on the produced outputs. The exploration of feature improvement has emerged as a promising approach within the field of adversarial defense. This approach aims to tackle the significant issue of adversarial attacks by promoting the development of models that are more robust and dependable. Through the utilization of enhanced data representation, models can acquire improved abilities to navigate the complex domain of adversarial perturbations, leading to increased levels of resilience, precision, and reliability. The incorporation of feature enhancement approaches into adversarial defense tactics is anticipated to have a significant impact on fostering machine learning systems against the constantly emerging adversarial attacks as research and innovation advance.

3.2.3. Adversary Training

The fragility of modern AI systems has been brought to light by adversarial attacks, which entail deliberate modification of input data to deceive machine learning algorithms. Adversarial training, a proactive defense mechanism, has been recognized as an effective technique for improving the resilience of these models against such attacks. Through the integration of adversarial attacks during the training phase, adversarial training enhances the capability of models to tolerate perturbations and to generate dependable predictions, even when exposed to adversarial input. Notably, a significant number of research studies have diversified upon existing adversarial training techniques that were developed initially for conventional images and applied to the field of medical classification tasks [18,22,112–114]. The study conducted by [18,127] focused on investigating adversarial cases in medical imaging. The researchers developed multiple strategies to mitigate the impact of these adverse occurrences. During the adversarial training phase, integrating both FGSM [42] and JSMA [53] approaches is employed to generate adversarial instances. To strengthen the resilience of the system, the researchers incorporated Gaussian noise into the data used for adversarial training. Additionally, they substituted the original Rectified Linear Units (ReLU) activation function with the Bounded ReLU variant.

Xu et al. conducted a study where they not only evaluated the robustness of several computer-aided diagnosis models but also implemented PGD-based adversarial training [45] and the Misclassification Aware adveRsarial Training (MART) approach [128,129] to improve the resilience of these models. To enhance the evaluation of resilience to typical disturbances, the researchers also established a fresh medical dataset known as the Robust-Benchmark. The aforementioned efforts primarily focus on improving the ability of individual models to withstand hostile attacks. Moreover, there are various adversarial defense methods specifically designed for medical image analysis, along with the transfer of natural defense techniques to diagnostic models [130–132]. In their study, Liu et al. [130] examined three distinct types of adversarial augmentation cases that might be incorporated into the training dataset to enhance robustness. The authors employ Projected Gradient Descent (PGD) [45,69] as an iterative method to search for the most challenging latent code to generate adversarial nodules that the target diagnosis model cannot detect. The concept of adversarial training is a significant development in the domain of adversarial machine learning, as it offers a means to enhance models' resilience against adversarial perturbations. Adversary training enhances the robustness and reliability of AI systems by endowing models with the capability to discern authentic inputs from hostile ones. The ongoing evolution of the adversarial landscape necessitates additional study and innovation in adversarial training approaches. These advancements have the potential to establish a

more secure and reliable basis for the deployment of machine-learning models in many applications (Table 2).

Table 2. Summary of adversarial defense works in the context of medical image analysis.

Ref./Year	Evaluation Metrics	Defense Model	Task	Image Modality
[57]/2018	FGSM, PGD, BIM, L-BFGS, DeepFool	Feature enhancement	Classification	X-ray
[133]/2018	FGSM, BIM	Adversarial training	Reconstruction	CT
[134]/2019	FGSM	Adversarial training	Segmentation	MRI
[135]/2019	FGSM, I-FGSM, CW	Feature enhancement	Classification	X-ray, Dermoscopy
[21]/2019	FGSM, CW, PGD, BIM, GN, SPSA, MI-FGSM	Feature enhancement	Classification, Segmentation, Object Detection	X-ray, Dermoscopy
[18]/2019	GN	Adversarial training	Classification	CT, MRI
[126]/2019	I-FGSM	Feature enhancement	Segmentation	X-ray, Dermoscopy
[136]/2019	FGSM, I-FGSM	Adversarial training	Segmentation	CT
	FGSM	Adversarial training	Classification	MRI
[125]/2019	Optimization-based attack	Feature enhancement	Low-level vision	X-ray, MRI
[137]/2020	DAG	Adversarial detection	Segmentation	MRI
[138]/2020	FGSM, I-FGSM	Feature enhancement	Regression	MRI
[61]/2020	FGSM, PGD, DAA, MI-FGSM, DII-FGSM	Adversarial Training, Pre-processing	Classification	X-ray
[47]/2020	OPA, FGSM	Adversarial training	Classification	CT
[139]/2020	PGD	Adversarial training	Classification	Fundoscopy
[140]/2020	PGD, FGSM	Adversarial training	Classification, Segmentation	X-ray, MRI
[141]/2020	PGD, I-FGSM	Adversarial training	Classification, Segmentation	CT, MRI
[142]/2020	False-negative adversarial feature	Adversarial training	Reconstruction	MRI
[143]/2020	GAN-based attack	Adversarial training	Reconstruction	CT, X-ray
[144]/2020	PGD, FGSM	Adversarial training	Classification	Dermoscopy
[44]/2020	PGD, BIM, CW, FGSM, DeepFool,	Pre-processing	Classification	CT, X-ray
[130]/2020	PGD	Adversarial training	Classification	CT
[64]/2020	Adversarial bias attack	Adversarial training	Segmentation	MRI
[145]/2020	ASMA	Pre-processing	Segmentation	Fundoscopy, Dermoscopy
[131]/2020	FGSM, Deep Fool, Speckle noise attack	Adversarial training, feature enhancement	Classification	X-ray, Fundoscopy
[17]/2021	FGSM, BIM, PGD, CW	Adversarial detection	Classification	X-ray, Fundoscopy, Dermoscopy
[146]/2021	PGD, CW	Adversarial detection	Classification	X-ray
[124]/2021	PGD, FGSM	Feature enhancement	Segmentation	CT, MRI
[83]/2021	UAP	Adversarial training	Classification	OCT, X-ray, Dermoscopy
[70]/2021	FGSM, PGD, CW	Adversarial training, adversarial detection	Classification	OCT
[70]/2021	PGD, GAP	Adversarial training	Classification	X-ray, Fundoscopy, Dermoscopy
[131]/2021	FGSM, Deep fool, Speckle noise attack	Adversarial training, feature enhancement	Classification	X-ray, Fundoscopy
[147]/2021	FGSM	Adversarial training	Segmentation	CT
[148]/2021	FGSM, BIM, CW, Deep Fool	Adversarial detection	Classification	Microscopy
[120]/2021	PGD	Feature enhancement	Classification	CT, MRI, X-ray
[19]/2021	FGSM	Adversarial training	Object Detection	CT, Microscopy
[76]/2021	FGSM, I-FGSM, TI-FGSM	Distillation	Segmentation	MRI
[149]/2021	PGD, AA	Feature enhancement, adversarial training	Segmentation	CT, MRI
[150]/2022	FGSM	Adversarial training	Classification	MRI
[46]/2022	FGSM, PGD	Pre-processing	Classification	X-ray

Table 2. *Cont.*

Ref./Year	Evaluation Metrics	Defense Model	Task	Image Modality
[151]/2022	Hop skip jump attack	Adversarial detection	Classification	MRI, X-ray, Microscopy
[20]/2022	I-FGSM, PGD, CW	Pre-processing	Classification	X-ray, Dermoscopy
[127]/2022	FGSM, PGD, BIM	Adversarial training	Classification	CT, MRI, X-ray
[152]/2022	OPA	Adversarial detection	Classification	Microscopy
[128]/2022	DDN	Adversarial training	Classification	MRI
[153]/2022	FGSM, PGD	Adversarial training	Classification	OCT, X-ray, Dermoscopy
[154]/2022	PGD, I-FGSM	Adversarial training	Segmentation, Object Detection, Landmark Detection	MRI, X-ray, Microscopy
[155]/2022	FGSM, PGD, BIM	Adversarial training	Classification	CT, MRI, X-ray
[156]/2022	FGSM, BIM, CW, PGD, AA, DI-FGSM	Pre-processing	Classification	Dermoscopy
[157]/2022	FGSM, PGD, FAB, Square attack	Feature enhancement	Classification	Microscopy
[158]/2022	FGSM, PGD, Square attack, Moment-based adversarial attack	Adversarial training	Classification, Segmentation	X-ray, Microscopy
[132]/2022	FGSM, PGD, BIM, Auto PGD	Adversarial detection, feature enhancement	Classification	X-ray, Fundoscopy
[159]/2022	FGSM, PGD, CW	Adversarial training, feature enhancement	Classification	Ultrasound
[160]/2022	FGSM, PGD, SMA	Feature enhancement	Segmentation	CT
[103]/2022	L-BFGS, FGSM	Adversarial training, distillation	Classification	Fundoscopy
[161]/2022	FGSM	Adversarial training	Classification	Microscopy
[119]/2022	FGSM	Feature enhancement	Classification	Fundoscopy
[162]/2022	DAG, I-FGSM	Pre-processing	Segmentation	MRI, X-ray, Fundoscopy
[163]/2022	PGD	Adversarial training	Segmentation	MRI
[164]/2022	FGSM, PGD	Feature enhancement	Classification	X-ray, Fundoscopy
[104]/2023	FGSM, PGD	Adversarial training	Classification	Dermoscopy
[105]/2023	FGSM	Adversarial training	Classification	CT
[106]/2023	PDG, FGSM, BIM, GN	Adversarial training	Segmentation	MRI
[107]/2023	PGD, BIM, FGSM	Adversarial training	Classification, Detection	CT, MRI
[108]/2023	PDG, CW, BIM	Feature enhancement	Classification	X-ray, Fundoscopy, Dermoscopy
[109]/2023	FGSM, MI-FGSM, PDG, CW	Adversarial training, feature distillation	Reconstruction	X-ray, Fundoscopy
[110]/2023	L-BFGS, FGSM, PDG, CW	Adversarial training	Classification, Reconstruction	MNIST
[111]/2023	FGSM, PGD, MIM, CW	Adversarial training	Classification, Segmentation	X-Ray, CT, Ultrasound

4. Experiment

This section illustrates the exhibition of the practical experimentation and evaluation of medical image adversarial attack and defense especially in the area of segmentation and classification. First, the employed datasets and the type of data preprocessing we used are introduced, followed by the evaluation metrics and the proposed approach. The qualitative and quantitative results are presented next, and we conclude the section with the results discussion.

4.1. Implemented Attack and Defense Model

In this review article, we used the ResNet-18 [165] (Figure 10) and Auto Encoder-Block Switching [166] (Figure 11) for medical image attack and defense classification while using the U-Net [3] architecture (Figure 12) for medical image segmentation task. In this study, we primarily focus on attacks under the "norm threat model," which is the most typical

case. We provide accurate results on both clean and adversarial instances, achieved using five powerful adversarial attack techniques: FGSM [42], PGD [45] with 20 steps and a step size of 1/255, CW [41], and Auto Attack (AA) [167]. The norm perturbation range includes 2/255, 4/255/ 6/255, and 8/255. On a PC running Windows and the Python environment, we carried out our experiment. The machine had a 2.30 GHz Intel(R) Core(TM) i5-8300H processor and a 4 GB NVIDIA GeForce GTX 1050 Ti graphics card. We used open-source TensorFlow.Keras deep learning framework to build the network, which we found to be a useful resource. We used distributed processing and relied on the CUDA 8.0 and CUDNN 5.1 requirements to increase the effectiveness of our training. Table 3 summarizes the implementation hyperparameters of the models.

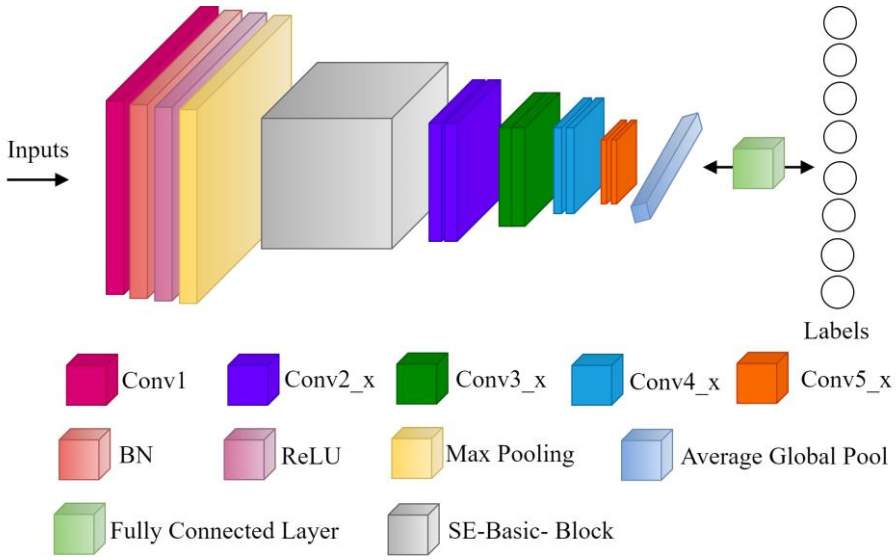

Figure 10. ResNet-18 architecture used for adversarial attack implementation.

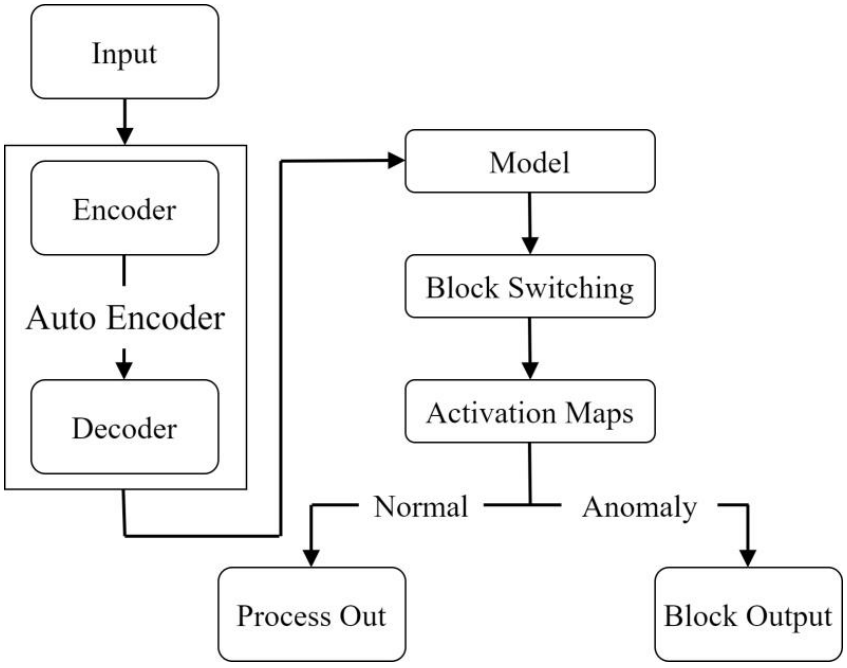

Figure 11. Auto Encoder-block switching architecture used for adversarial defense implementation.

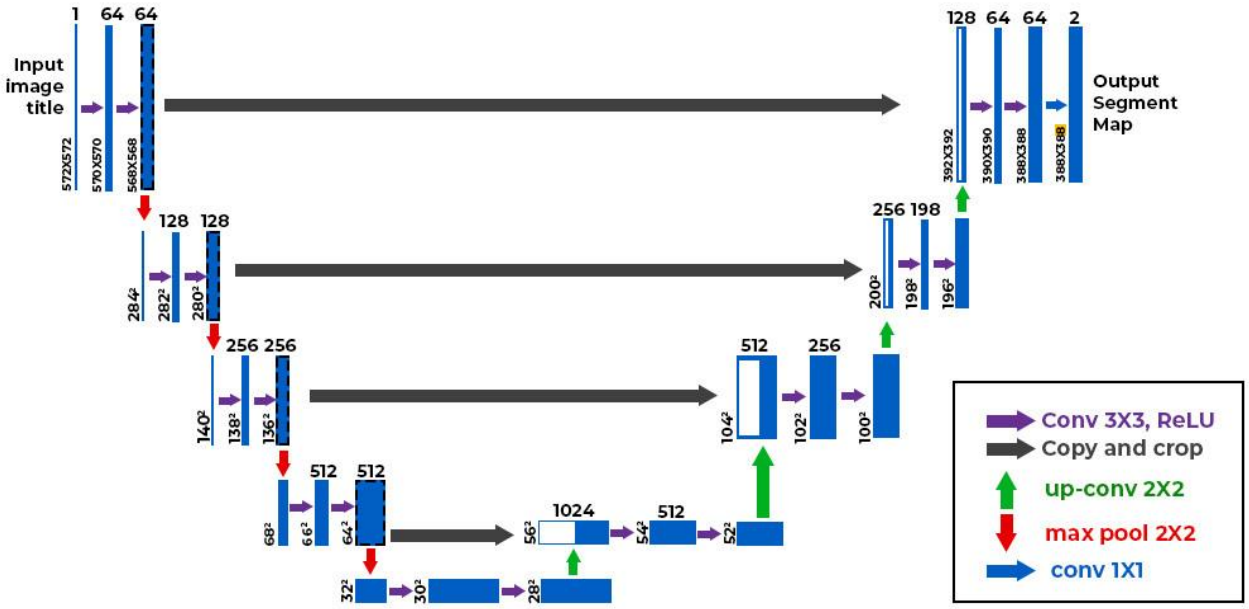

Figure 12. UNET architecture used for adversarial segmentation attack and defense implementation.

Table 3. Implementation hyperparameters of the implemented models.

Resnet-18 Architecture	
Input size	224 × 224
Weight decay	1×10^{-4}
momentum	0.9
Mini batch	256
Optimizer	Adams optimizer
Initial learning rate	0.1
Reduction in learning rate	10 per 30 epochs
Iterations—number of epoch	100
Loss function	Categorical Cross-Entropy
Batch size	4
UNET Architecture for Segmentation	
Input size	256 × 256 × 3
Filters per convolutional layer	64, 128, 256, 512, 1024
Optimizer	Adams Optimizer
Training loss	Binary cross entropy (BCE) and Dice loss
Cosine annealing learning rate scheduler/learning rate	1×10^{-4}
momentum	0.9
Epoch	400
Batch size	8
Adversarial attack perturbations	1, 2, 4, 6, 8
Auto Encoder-Block Switching Architecture	
Input size	224 × 224 × 3
Weight decay	1×10^{-4}
Optimizer	Adams optimizer
learning rate	0.002
Number of epoch	150
Loss function	Mean Square Error
Batch size	4
Input size	224 × 224
Adversarial attack perturbations	1, 2, 4, 6, 8

4.2. Dataset and Data Preprocessing

The datasets we used include (1) Messidor dataset [168], which has four classes with s total number of 1200 RGB samples of the eye fundus for classifying diabetic retinopathy based on retinopathy grade; (2) the ISIC 2017 dataset [169], which has three classes with 2750 samples and is collected by the International Skin Imaging Collaboration for the classification and segmentation of skin lesions; (3) the Chest-ray 14 dataset [170], which has 112,120 frontal view X-ray samples from 14 thorax disorders; and lastly (4) the COVID-19 database [171], which contains 21,165 chest X-ray samples with segmentation-ready lung masks. Figure 13 shows the pictorial representation of the employed datasets.

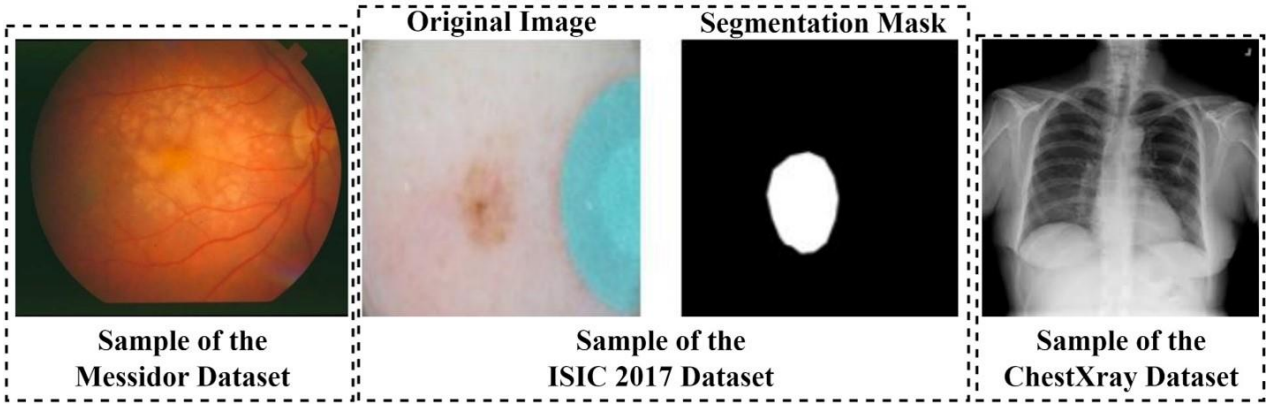

Figure 13. Illustration of the employed dataset.

Noting that this experiment is based on classification and segmentation scenarios, we carried out both binary and multiclassification cases for the classification task. For the Messidor dataset, we used the original partition i.e., 960 fundus samples for training with several data augmentation techniques, including random rotations and flips, and the rest for testing. The preprocessing strategy that was carried out on the ISIC 2017 dataset includes resizing and center-cropping. Due to the unbalanced nature of the Chest-ray 14 dataset, we randomly selected 10,000 X-ray samples while using 8000 as training samples and 2000 as testing samples while randomly flipping and normalizing them. For the segmentation task, we used a total of 2750 X-ray images (2000 for the training set) as originally designed by the ISIC 2017 dataset.

4.3. Evaluation Metrics

We used different metrics for the different tasks carried out in this manuscript. For the classification task, we used the accuracy evaluation metrics since adversarial attacks may mislead a target model into incorrect predictions, which is measured mathematically as follows:

$$\text{Accuracy} = \frac{(\text{True positives} + \text{True Negatives})}{\text{True positives} + \text{True negatives} + \text{False positives} + \text{False negatives})} \tag{5}$$

For the segmentation task, noting that adversarial attacks lead to incorrect segmentation results, we employed the use of mean Intersection over Union (mIoU) and Dice coefficient (F1_score) to measure the generated segmentation mask and target mask as mathematically calculated as follows:

$$\text{mIOU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{6}$$

$$\text{Dice} = \frac{2 \times \text{Area of Overlap}}{\text{Total Area}} \tag{7}$$

The idea of adversarial defense tries to make neural network outputs comparable for both adversarial and clean cases. This defensive tactic strengthens the network's ability to withstand hostile examples.

4.4. Results and Analysis

This subsection first introduces several examples of adversarial attacks on various medical diagnostic classification tasks, as shown in Figure 14, illustrating varying levels of attack potency. It is obvious that, even when subjected to small perturbations, AI models are extremely vulnerable to hostile situations. The disruption of clinical evaluation is exacerbated by the attacker's ability to mislead the medical categorization models into making confidently false diagnoses. Furthermore, the incorrect partitioning results brought on by this attack may also result in incorrect treatment suggestions.

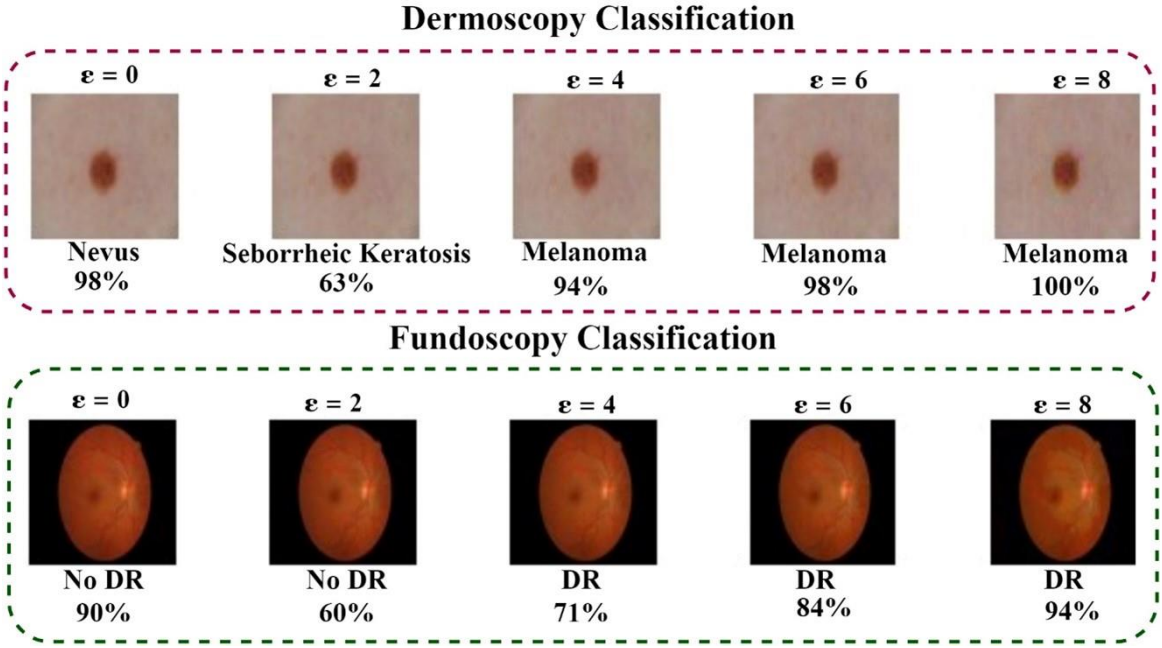

Figure 14. Medical adversarial instances (classification tasks) along with predictions across a range of perturbation rates.

4.4.1. Adversary Attack Experimental Results

Table 4 summarizes the level of accuracy of the employed deep learning model against various attacks in both binary and multiclass classification situations in order to thoroughly assess the effect of adversarial attacks on medical classification algorithms. It is vital to remember that we create adversaries using the cross-entropy loss function. Based on the results presented in Table 4 and Figure 14, the utilized deep learning model accuracy noticeably decreases as the size of the attack perturbations grows. Notably, compared to its binary classification, multiclass classification models suffer from a more dramatic decline in accuracy. However, since the majority of medical adversarial defensive strategies proposed by researchers recently focus on binary tasks, we argue that achieving adversarial robustness for multiclass medical classification is a more difficult task, with ramifications for many clinical contexts.

Furthermore, we assess the effectiveness of medical segmentation models against PGD-based adversarial cases in addition to adversarial attacks against medical classification, as shown in Table 5 and Figure 15. The Dice loss and Binary Cross-Entropy (BCE) loss were used as the adversarial losses whilst creating the adversary attacks. Adversarial instances produced by decreasing Dice loss might enhance attack effectiveness against medical segmentation methods.

Table 4. Adversarial (white-box) classification result (%) using the ResNet-18 (binary and multi-class classification tasks) under various attack scenarios.

Adversarial Attack Type	ε	Dermoscopy		Fundoscopy	
		Binary (%)	Multi-Class (%)	Binary (%)	Multi-Class (%)
Clean Image	0	71.3	50.0	64.9	60.0
AA [167]	2/255	22.1	2.4	3.8	2.5
	4/255	8.4	0.8	0.0	0.0
	6/255	4.5	0.3	0.0	0.0
	8/255	1.6	0.1	0.0	0.0
CW [41]	2/255	22.5	2.5	7.1	6.7
	4/255	9.3	0.9	1.3	2.1
	6/255	5.2	0.7	0.0	0.0
	8/255	2.1	0.4	0.0	0.0
FGSM [42]	2/255	37.6	4.8	38.8	13.8
	4/255	35.7	4.5	27.9	9.2
	6/255	36.2	6.1	25.7	15.3
	8/255	34.5	8.0	24.2	20.4
PGD [45]	2/255	22.8	2.6	8.8	4.6
	4/255	12.5	0.8	1.7	0.4
	6/255	11.9	0.6	0.2	0.0
	8/255	12.8	0.4	0.0	0.0

Table 5. Adversarial segmentation result (white box)using the u-net model based on PGD adversarial attack.

Adversarial Loss	ε	COVID-19		Dermoscopy	
		mIOU	Dice	mIOU	Dice
None	0	0.976	0.982	0.801	0.875
BCE	2/255	0.559	0.690	0.405	0.517
	4/255	0.355	0.492	0.248	0.354
	6/255	0.200	0.412	0.198	0.301
	8/255	0.230	0.350	0.167	0.255
Dice	2/255	0.473	0.610	0.340	0.450
	4/255	0.265	0.391	0.158	0.243
	6/255	0.213	0.332	0.102	0.198
	8/255	0.152	0.246	0.082	0.140

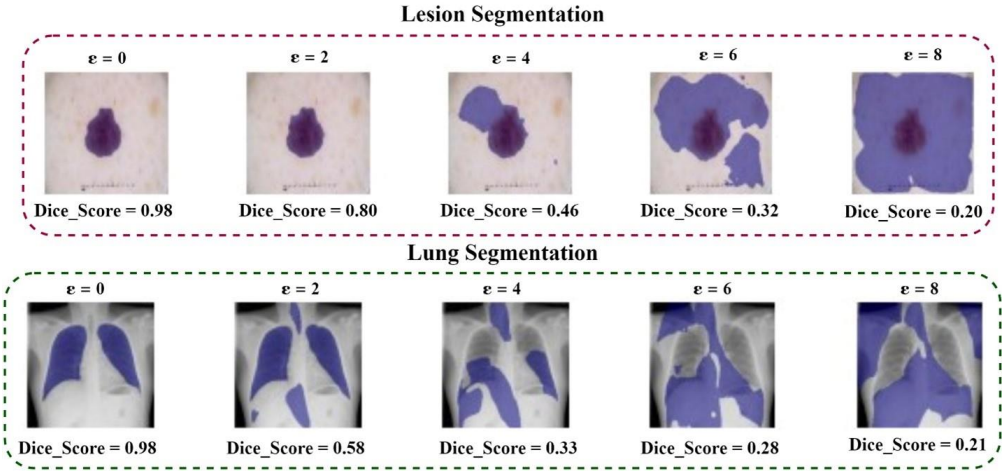

Figure 15. Medical adversarial instances (segmentation tasks) along with predictions across a range of perturbation rate.

4.4.2. Adversary Defense Experimental Results

Modern adversarial training methods mostly focus on improving adversarial images while training. This improves the model's capacity for precise judgments in both routine and adversarial cases. This work applies PGD-AT [45] adversarial training approaches to the field of biomedical imaging to further these methodologies. In both binary and multi-class contexts, the research illustrates the effectiveness of this strategy in obtaining adversarial resilience for medical classification (see Table 6). The findings show that adversarial-trained models preserve their resilience in the face of different attack configurations. It is important to note that PGD-AT [45] differs from other medical classification models in terms of robustness since it focuses on various methods for creating internal enemies.

Table 6. White-box accuracy (%) of adversarial trained (PGD-AT [45]) medical classification models for binary and multi-class classification in various scenarios.

Adversarial Attack Type	ε	Dermoscopy		Fundoscopy	
		Binary (%)	Multi-Class (%)	Binary(%)	Multi-Class (%)
Clean Image	0	61.2	52	64.9	60
AA [167]	2/255	55.6	43.7	55.2	41.3
	4/255	48.4	35.5	52.3	35
	6/255	41.6	26	36.2	31.9
	8/255	34.8	20.9	42.1	28.3
CW [41]	2/255	55.6	43.6	56.4	42.3
	4/255	48.5	36.2	53.6	34.9
	6/255	42.5	29.4	48	31.6
	8/255	36.8	23.3	43.5	28.8
FGSM [42]	2/255	55.6	44.5	55.6	44.5
	4/255	49.2	38.2	53.6	40
	6/255	44.3	32.3	49.3	38.2
	8/255	39.5	25.5	45.5	36.4
PGD [45]	2/255	57.9	48	56.4	41.6
	4/255	48.2	36.2	53.2	36.4
	6/255	42.4	29.7	48.2	33.8
	8/255	35.1	22.4	43.6	31.2

We expand our method to adversarial training in the context of medical segmentation problems in addition to demonstrating adversarial resilience for single-label classification models. By analyzing how well these models function under various attack scenarios, as shown in Table 7, we evaluate the efficacy of this adversarial training for medical segmentation. To increase the natural robustness of medical segmentation models, we use the widely used PGD-AT [45] adversarial training technique. Our findings show that compared to their natively trained counterparts, adversarial-trained segmentation models are more resilient to various types of adversarial attacks. Notably, our research shows that attacking the Dice loss still has a greater success rate than attacking the BCE loss.

Table 7. Results of white-box resilience against PGD attack in various attack configurations utilizing u-net for biomedical segmentation.

Adversarial Loss	ε	COVID-19		Dermoscopy	
		mIOU (%)	Dice (%)	mIOU (%)	Dice (%)
None	0	0.976	0.982	0.801	0.875
BCE	2/255	0.931	0.962	0.773	0.842
	4/255	0.890	0.940	0.733	0.810
	6/255	0.854	0.912	0.684	0.705
	8/255	0.812	0.887	0.601	0.700
Dice	2/255	0.923	0.958	0.767	0.838
	4/255	0.880	0.932	0.723	0.803
	6/255	0.832	0.894	0.669	0.750
	8/255	0.774	0.860	0.594	0.694

5. Discussion

Adversarial phenomena have demonstrated substantial prospective applicability, even within the context of cutting-edge deep neural networks (DNNs), irrespective of the extent of access granted to the attacker in relation to the model, as well as their potential to remain imperceptible to the human visual system. In comparison to various other sectors within the field of computer vision, it has been established that medical-oriented DNNs exhibit heightened vulnerability when subjected to adversarial attacks. Pertinent to this notion, it has been shown that adversarial samples with limited disturbance can trick advanced medical systems, which otherwise demonstrate excellent performance when handling clean data. The aforementioned occurrence highlights the vulnerability of medical deep neural networks (DNNs) when exposed to adversarial inputs.

A. Dataset and Labeling

The availability of labeled datasets for training models in medical imaging is significantly more limited compared to datasets for general computer vision tasks. Typically, datasets for general computer vision encompass a wide range of several hundred thousand to millions of annotated photographs [158]. This scarcity can be attributed to several factors, including concerns over patient privacy and the absence of widely adopted procedures for the exchange of medical data. Another issue is that the process of assigning labels to medical images is a labor-intensive and time-consuming activity, and it is worth noting that the true nature of images in medical databases sometimes presents ambiguity and controversy, even among physicians and radiologists. Hence, this disparity in dataset size has a direct impact on the performance of deep neural networks (DNNs) when used in medical imaging analysis. In light of the scarcity of accessible annotated medical datasets, several scholars have lately put forth several approaches aimed at addressing this issue employing straightforward augmentation techniques. These phenomena have a substantial impact on the generalizability of the network and render it susceptible to adversarial attacks. The effective use of basic augmentation methods, including cropping, rotating, and flipping, has shown to be successful in the creation of unique and unconventional imagery. To address the challenges of vanishing gradient and overfitting, researchers have employed several techniques such as improved activation functions, modified cost function architecture, and drop-out methods [159]. The problem of high computational load has been effectively mitigated by the utilization of highly parallel technology, such as graphics processing units (GPUs), coupled with the implementation of batch normalization techniques.

On the other hand, a synergistic methodology involves the amalgamation of convolutional neural networks (CNNs) with transfer learning approaches [160]. The essence of this approach is in the use of parameters obtained by convolutional neural networks (CNNs) in the context of the primary application to enable the training of the modal. The incorporation of transfer learning into convolutional neural network (CNN) frameworks

is a significant direction for future study. This approach shows promise in addressing the challenge of limited labeled medical data. Additionally, a potential approach for increasing the quantity of the dataset involves the utilization of a crowdsourcing technique [103]. The notion of crowdsourcing in the context of health concerns involves the distribution of solutions, facilitated by skilled entities, from a specific research group to a wider population, comprising the general public. This channel offers a compelling direction for future study, facilitating a shift from individual duties to collective endeavors, resulting in societal benefits.

B. Computational Resources

Deep neural networks (DNNs) have become a fundamental component in the domain of machine learning, significantly transforming several disciplines like computer vision, natural language processing, and medical diagnostics. Nonetheless, this has not come without challenges, the process of training DNNs requires a significant amount of computer power, data processing skills, and memory resources. The significance of computational resources lies in their ability to support the sophisticated architecture of deep neural networks characterized by several layers and intricate connection patterns, which enable them to learn hierarchical representations from raw input. However, the effectiveness of training is inherently dependent on the presence of enough computational resources. These resources facilitate the implementation of several rounds through the training dataset, wherein the network adapts its internal parameters to reduce the disparity between anticipated outputs and actual labels. To effectively train deep neural networks (DNNs), it is necessary to have hardware components that possess significant computational capabilities, such as graphics processing units (GPUs) or specialized hardware like tensor processing units (TPUs). These devices have been specifically designed to maximize performance in the extensive parallel calculations required for training neural networks.

AlexNet is widely recognized as one of the most prominent and influential deep neural network (DNN) architectures in the field. The first implementation of AlexNet [172] involved the utilization of dual-GPU training. However, further advancements in GPU computation allowed for the transition to a single GPU configuration, which facilitated the integration of eight deep layers into the network architecture. The AlexNet architecture is widely regarded as the foundational model for many DNN topologies. VGGNet is a convolutional neural network that has been enhanced and improved by Simonyan and Zisserman [173]. The proposed approach employs a series of stacked convolutional layers and a maximum pooling layer in a repetitive manner. The network employed in this study is a commonly utilized architecture that utilizes a range of 16 to 19 convolutional neural network (CNN) layers to extract picture data. The VGGNet's breakthrough in extracting visual features may be attributed to its utilization of a 3×3 convolution and 2×2 pooling kernels. It is worth noting that memory constraints are also a challenge, which arises from the presence of several layers, each containing learnable parameters. As a result, a significant amount of memory is needed to record gradients, activations, and intermediate findings during the training procedure. The limitation of memory might impede the training process of big networks or require a decrease in batch sizes, thereby affecting the pace at which convergence is achieved. The training time of deep convolutional neural networks (CNNs) can be rather extensive, resulting in a significant time investment for research and development efforts. Achieving faster training can be facilitated by the utilization of distributed computing setups or by using cloud-based resources. Finally, energy consumption associated with training DNNs is substantial due to the computing requirements involved. The aforementioned element has significant importance within the field of artificial intelligence, as it aligns with the overarching goals of enhancing energy efficiency and promoting environmental sustainability.

C. Robustness against Target Attacks

Within the field of medical imaging, the convergence of artificial intelligence and healthcare has significant potential. However, it is crucial to address the important frontier

of achieving resilience against targeted assaults. This requires committed attention and focus. Significant progress has been achieved in enhancing the security of medical imaging systems against non-targeted or random assaults, as shown by the study conducted by Bortsova et al. [174]. However, the emergence of adversarial threats in the form of focused attacks poses a novel and severe obstacle. In the context of targeted assaults, malevolent entities purposefully aim to influence the model to induce certain diagnostic mistakes, which might pose a threat to the well-being of patients. The resolution of this dilemma necessitates a fundamental adjustment in the approach to research and development, accompanied by a significant need for tactics that possess the capability to not only identify but also successfully counteract these specific types of assaults. The study conducted by Han et al. [175] highlights the need to address this research gap to maintain the reliability and credibility of medical imaging systems. Nevertheless, the inherent features that made AI-powered medical imaging systems very promising also expose them to potential vulnerabilities in the form of adversarial assaults. Adversarial assaults include the deliberate manipulation of input data to mislead the AI model into generating inaccurate or detrimental predictions. In instances of non-targeted or random assaults, the primary objective of the attacker is to impede the operational capabilities of the model without a predetermined objective or intended recipient. These assaults may elicit worry, but they may not always result in precise diagnostic mistakes.

In contrast, targeted assaults are purposefully planned with the explicit intention of causing the AI model to produce predetermined diagnostic mistakes. Within the domain of medical imaging, this may include inducing the model to erroneously classify a medically fit individual as suffering from a grave pathological illness or to disregard a crucial sickness. The ramifications of these specific assaults in the healthcare sector are significant, possibly leading to the postponement or inaccuracy of medical interventions, superfluous medical procedures, or even endangerment of people. The current framework of defensive mechanisms in medical imaging mostly focuses on enhancing the system's ability to withstand non-targeted or stochastic assaults. The purpose of these defensive mechanisms is to enhance the resilience of the AI model against little perturbations in input data, hence preventing the generation of inaccurate outcomes caused by modest changes in picture quality or structure. Although the aforementioned defenses provide unquestionable value, their efficacy in countering targeted assaults may be limited due to their potential inadequacy in addressing the complexity and purposefulness of such attacks. The study conducted by Bortsova et al. [174] serves as a pertinent reminder of the current emphasis on non-targeted assaults in the field of medical imaging. The research highlights the need to enhance the overall resilience of AI models to protect against typical hostile perturbations that may arise during regular functioning.

To bolster the resilience of medical imaging systems against deliberate assaults, future research endeavors must investigate novel methodologies and approaches. One potential avenue of research is the exploration and refinement of adversarial training approaches that are expressly designed to address the weaknesses that are exploited by targeted assaults. The process of adversarial training entails the incorporation of adversarial instances into the training dataset, which allows the model to acquire the ability to identify and withstand manipulations that are often found in real-world scenarios. It is crucial to note that the advancement of countermeasures against targeted assaults should not only prioritize detection but should also emphasize active mitigation. This entails enhancing medical imaging systems with the capacity to promptly detect and address any dubious input data. If a targeted attack is identified or presumed, the system needs to possess established processes that can promptly identify and notify healthcare professionals of the problem. Furthermore, the system should be equipped to possibly implement countermeasures to avert or mitigate the adverse effects of the assault on patient care. The research conducted by Han et al. [175] highlights the urgent need to tackle the issue of targeted assaults within the field of medical imaging. The research community must acknowledge the significance

of this research gap and give precedence to the creation of strong defenses specifically designed for the intricacies of healthcare applications.

D. Evaluation of Transferability and Adaptability

The concepts of applicability have become crucial factors that contribute to the increased effectiveness of attacks, resulting in a higher vulnerability of deep neural networks (DNNs) to misclassification. Previous research has mostly focused on examining the vulnerability of the network to basic, non-targeted attacks. Nevertheless, these basic attacks have limited effectiveness in some deep neural network (DNN) models and may be easily detected using newly developed security techniques. Similarly, there has been a significant effort to create adversarial attacks that surpass the limitations imposed by certain models and images. This has emerged as a crucial obstacle in current research in the field of computer vision. These sophisticated attacks, which are not limited by specific models or images, have the potential to be applied in many medical learning tasks. there is a lack of well-developed security mechanisms and detection systems specifically designed to mitigate universal attacks. The lack of effective defenses in the field is further emphasized by the increased vulnerability of deep neural networks (DNNs) to universal attacks, which may fool with greater ease and use fewer resources. The metric of adaptability is a significant measure that arises within the contextual framework of previous research efforts. The relevance of this phenomenon is in its ability to measure the degree to which an attack might spread its effects across several models, especially when these models are limited to a black-box state. According to [63], it has been hypothesized that attacks with high adaptability would need a sacrifice in the quality of the adversarial image while maintaining a high success rate (100%) of the attack. The application of unsupervised learning has gained popularity as a means to enhance adaptability. It functions as a feasible approach to improve the ability of networks to effectively respond to the transmission of adversarial perturbations. Hence, the primary issues and obstacles that are prevalent in the current study domain focus on the fundamental concepts of transferability and universality within the framework of adversarial attacks on deep neural networks (DNNs).

E. Interpretability and Explainability

The notions of interpretability and explainability have become fundamental principles in the domain of artificial intelligence and machine learning, especially in the context of healthcare applications. The confidence that doctors and healthcare practitioners have in defensive AI systems is significantly influenced by these two interrelated characteristics. This subject has been extensively investigated in the research conducted by Saeed et ai. [176]. The trust gap has far-reaching ramifications within the healthcare sector, particularly with the decision-making process of physicians, which directly affects the well-being of patients. Busnatu et al. [177] argue convincingly for the need to emphasize interpretability and provide clear explanations for the decision-making processes of future defenses to address these problems. The use of interpretability in AI for healthcare not only improves its overall efficacy but also fosters increased acceptability among clinical professionals. To comprehensively examine the importance of interpretability and explainability, it is necessary to first acknowledge their inherent interrelation. The concept of interpretability pertains to the capacity of an artificial intelligence (AI) model to be grasped and understood by individuals, specifically those who possess expertise in the relevant field. On the other hand, explainability focuses on the capability of these models to provide explicit and logical arguments for the judgments they make. Collectively, these characteristics provide a level of transparency into the internal mechanisms of artificial intelligence (AI) systems, a trait that is progressively essential within the healthcare domain.

The absence of interpretability inside artificial intelligence (AI) systems has significant implications. When artificial intelligence (AI) models function as opaque entities sometimes referred to as "black boxes" (a phrase used by Saeed et al. [176]) healthcare professionals face a basic quandary. Patients are required to rely on these systems for crucial healthcare decisions, sometimes without possessing a comprehensive understanding of the specific

processes or rationales behind a given suggestion or diagnosis. The lack of transparency undermines the confidence that healthcare professionals have in AI-driven solutions, leading to hesitancy and doubt. To tackle this difficulty, it is crucial to establish artificial intelligence (AI) defenses that possess both efficacy and interpretability. An interpretable AI system offers doctors a transparent comprehension of its decision-making process, enabling them to place more faith in and seamlessly incorporate AI suggestions into their workflow. The concept of interpretability may be seen in several forms. For instance, an artificial intelligence (AI) system can provide a comprehensive analysis of the aspects it took into account throughout the process of generating a suggestion. This includes emphasizing the most significant variables and their corresponding weights. In addition, it has the potential to provide doctors with visual representations or explanations in natural language, enhancing transparency and facilitating accessibility in the decision-making process.

Furthermore, the need for interpretability goes beyond singular suggestions. Artificial intelligence (AI) systems must include the capability to provide explanations for the rejection or disregard of certain inputs. In the field of healthcare, where each data point can influence a patient's diagnosis or treatment plan, it is essential to comprehend the reasons behind the exclusion of a certain piece of information. These data equip healthcare professionals with the necessary knowledge to make well-informed judgments and guarantee that crucial particulars are not unintentionally disregarded. The study conducted by Busnatu et al. [177] highlights the importance of this particular technique. Their research connects with the wider industry trend towards transparent and trustworthy AI in healthcare by stressing the significance of interpretable AI defenses that provide a clear justification for their judgments. The advantages of interpretable artificial intelligence (AI) extend beyond enhancing trust and acceptability in healthcare settings. Additionally, they assist in continuous endeavors aimed at tackling concerns related to bias, justice, and accountability within artificial intelligence (AI) systems. The identification and rectification of biases or mistakes in data or algorithms is facilitated when AI judgments are clear and intelligible. This enhances the ability of AI systems to provide equitable and dependable assistance to physicians serving different patient groups. In their 2023 research, Saeed et al. [176] appropriately highlighted the understandable hesitance of practitioners to use "black box" artificial intelligence (AI) solutions. To address this hesitancy and effectively use artificial intelligence (AI) in the field of healthcare, it is imperative to prioritize the development of interpretable AI defensive mechanisms. The aforementioned defenses must not alone give precise suggestions but also provide physicians with lucid insights into the process of decision-making. By using this approach, it is possible to establish a connection between parties, enhance the capabilities of healthcare professionals to make well-informed choices and to guarantee that artificial intelligence (AI) becomes a beneficial partner in the endeavor to achieve improved healthcare results. The proposition made by Saeed et al. [176] for the implementation of interpretable AI defenses has significant importance in shaping the future of healthcare.

F. Real-time Detection and Response

Within the dynamic realm of healthcare, characterized by continuous change, the use of data-driven technologies has assumed a progressively significant position. Consequently, the notion of real-time detection and reaction has arisen as a crucial focal point. The aforementioned change in paradigm calls for a comprehensive reassessment of current defensive mechanisms since a significant portion of them lack the necessary capabilities to function optimally in real-time situations. This inadequacy has been highlighted by Paul et al. [178] in their recent research. The presence of this insufficiency not only reveals susceptibilities but also prompts substantial inquiries about the safety of patients, the security of data, and the general effectiveness of healthcare systems. The significance of real-time capabilities within the healthcare domain cannot be emphasized. In contrast to several other fields where the ramifications of delayed reaction may be comparatively less significant, the healthcare sector is distinguished by a continual stream of data, contact with patients, and sometimes life-threatening occurrences. Hence, the incapacity of traditional

defensive mechanisms to function in real-time presents a direct peril to the welfare of patients, healthcare practitioners, and the overall integrity of the healthcare system. Real-time detection and reaction play a crucial role in maintaining patient safety, protecting confidential medical data, and upholding the reliability of medical equipment within a healthcare environment. Nevertheless, achieving this ambition is a considerable challenge. The implementation necessitates a comprehensive strategy that encompasses advanced technology while also accounting for ethical, legal, and practical factors. Future endeavors in research and development should prioritize the pursuit of real-time detection and reaction to hostile inputs inside healthcare systems. This objective was well conveyed by Wang et al. [179] in their scholarly work published in 2023. One of the foremost obstacles in facilitating instantaneous identification and reaction is the ever-changing and uncertain characteristics of healthcare settings. In contrast to controlled laboratory conditions or typical computer systems, healthcare systems exhibit a state of continual change. Patients are admitted and discharged from the system, their medical conditions undergo modifications, and a constant influx of fresh data is received. The current dynamic environment necessitates a degree of agility and response that is lacking in most present protection mechanisms.

In addition, the emergence of telemedicine and remote patient monitoring has brought out a completely novel aspect of the field of healthcare. The advent of remote healthcare services has enabled patients to obtain medical care inside the confines of their residences; nevertheless, this development also introduces novel susceptibilities. The implementation of real-time detection and reaction has become crucial in safeguarding the confidentiality and privacy of patients' data, as well as assuring the dependability of remote medical equipment. In addition to the aforementioned immediate challenges, the use of real-time capabilities in healthcare procedures has significant potential for enhancing efficiency and precision. In the field of diagnostic imaging, the integration of AI algorithms to aid radiologists in anomaly detection has been seen. The provision of real-time feedback has shown the potential to enhance and expedite the diagnostic process. By promptly notifying users of potentially vital discoveries, the technology may guarantee that no significant time is wasted in commencing suitable interventions. The integration of real-time capabilities in the healthcare sector necessitates a thorough analysis of hostile inputs' characteristics. Within this particular context, hostile inputs comprise a broader range of occurrences, including intentional assaults as well as unintentional mistakes, system failures, and unanticipated oddities. To effectively identify and respond to events in real-time, it is essential to adopt a comprehensive strategy that considers a broad range of possible risks and anomalies, regardless of whether they are deliberate or accidental. Machine learning and artificial intelligence technologies are crucial in attaining this objective. These technologies can undergo training to identify patterns that are suggestive of hostile inputs. These inputs may range from deliberate efforts to breach a system to abnormalities that have the potential to harm medical care. Machine learning models provide the capability to perform ongoing analysis of data streams, detect deviations from expected patterns, and initiate appropriate actions promptly, therefore operating in real-time. Furthermore, the successful deployment of real-time detection and response systems requires the establishment of a complete framework that encompasses factors beyond technical aspects. The involvement of ethical and legal considerations arises, namely with patient permission, data protection, and liability. Achieving a harmonious equilibrium between the need for real-time patient monitoring to ensure safety and the imperative of upholding individual rights and adhering to regulatory frameworks is a nuanced and dynamic problem.

G. Adversarial Attacks in Multi-modal Fusion

The emergence of multi-modal fusion models has initiated a paradigm shift in the field of artificial intelligence, enabling the development of systems that possess the ability to analyze and incorporate data from diverse sources, including text, pictures, and audio. This advancement empowers these systems to make judgments that are more thorough and well-informed. Nevertheless, the advancement in this field has encountered obstacles,

and a significant issue that has arisen is the advent of adversarial assaults that have the potential to undermine the reliability of these multi-modal systems. The study conducted by Ding et al. [180] emphasizes the presence of distinct vulnerabilities in multi-modal fusion models, which may be exploited by hostile entities to affect the process of combining and interpreting input from various modalities. To ensure the dependability and credibility of multi-modal fusion methodologies, it is crucial to allocate resources towards the enhancement of resilient protective measures, as recommended by Cao et al. [181]. The notion of multi-modal fusion exemplifies the increasing intricacy and refinement of artificial intelligence. In contrast to models that function within a single modality, multi-modal models include input from several sources to get a comprehensive understanding of the data. In the domain of natural language processing, a multi-modal model can integrate many modalities such as text, pictures, and audio to enhance its understanding of the context and semantics of a certain topic. This methodology has facilitated the exploration of several prospects across diverse fields, including healthcare, driverless cars, and content recommendation systems. In the field of healthcare, the use of multi-modal fusion techniques facilitates enhanced diagnostic precision by the incorporation of diverse patient data originating from many sources, including but not limited to medical pictures, electronic health records, and patient interviews. Autonomous cars boost safety by enabling the vehicle to sense its surroundings via the use of a diverse array of sensors, cameras, and radar systems. Content recommendation systems provide personalized recommendations by monitoring a user's interactions with text, images, and videos.

Nevertheless, as the complexity and capabilities of multi-modal fusion models continue to advance, there is a concomitant rise in susceptibility. The phenomenon of adversarial assaults, which entails the manipulation of input data to deceive artificial intelligence (AI) systems, is not a recent development. However, its implications on multi-modal fusion systems pose distinct and specific issues. The aforementioned attacks can use the interplay between several modalities, such as text and picture, to make discreet but harmful modifications that may not be recognized by defenses particular to each modality. To exemplify the possible ramifications, let us contemplate a hypothetical situation whereby a multi-modal model is used within a healthcare environment to aid in the diagnostic process of patients by leveraging a fusion of medical imagery and textual patient information. The potential for an adversary to influence the textual patient data by making subtle alterations to symptoms or medical history exists, aiming to induce a mistake in the AI model. The intrinsic complexity of multi-modal fusion poses significant challenges in the detection of hostile operations. The study conducted by Ding et al. [180] underscores the imperative nature of addressing these vulnerabilities and devising targeted defenses for the protection of multi-modal fusion models. The research conducted by the authors sheds insight into the many methods via which adversarial assaults might manipulate the complex interconnections across modalities. This underscores the need to implement complete preventive measures. The development of defenses to ensure the security of multi-modal fusion techniques and the preservation of information integrity is a relevant and critical concern, as emphasized by Song et al. The implementation of defensive measures should include a comprehensive approach that takes into account the distinct characteristics of multi-modal fusion models.

The development of strong detection techniques is a crucial factor in effectively countering adversarial assaults in the context of multi-modal fusion. The used techniques should possess the ability to detect inconsistencies or irregularities within the data across several modalities. For example, when there is a lack of consistency between the information documented in a patient's medical record and the observations made in corresponding medical imaging, it should prompt a heightened level of attention for closer examination. These detectors must be calibrated to account for the complexities of intermodal interactions and should be engineered to resist advanced hostile manipulations. In addition, it is essential for the advancement of defensive measures to prioritize the reduction in adverse effects resulting from hostile assaults. This encompasses the exploration of methods for the restoration and rectification of modality-specific impairments. In the healthcare scenario, in the

event of an assault that alters textual patient records, the system needs to include processes that can detect and correct false information. This may be achieved by cross-referencing the manipulated data with other trustworthy sources or by executing supplementary diagnostic tests. The proposition made by Song et al. to establish defensive measures is a crucial undertaking to protect the dependability and credibility of multi-modal fusion methodologies. The advancement of AI systems necessitates the prioritization of protecting their multi-modal fusion capabilities. This is not just a technical obstacle but also a crucial need to guarantee the reliability and safety of applications in many fields. The whole potential of multi-modal fusion models in boosting our lives depends on our capacity to comprehend, adjust, and safeguard against hostile onslaught.

H. Future Work

Adversarial attacks have demonstrated significant importance in assessing the vulnerability of deep learning networks. Consequently, enhancing those attacks might assist researchers in addressing the limitations and devising strategies for more effective and secure medical educational systems. The primary hurdles soon are the transferability of attack techniques and the visibility of perturbations, both of which have a significant impact on the network's predictions. These challenges have been seen as a recent trend. Furthermore, additional attack tactics targeting models and undermining their resistance will be taken into account. To enhance the dependability of medical learning systems, it is vital to establish various methods for generating medical labeled images and expanding datasets in a more streamlined manner. Another approach to mitigate the network's susceptibility involved modifying the neural network's design, an area that has yet to be thoroughly investigated.

The utilization of pre-trained models, which involve training sophisticated neural networks using large-scale labeled imagery from a particular source, is anticipated to play a significant role in the future of deep neural networks (DNNs) when the availability of annotated images is limited to a small number. Reducing the number of learning parameters in deep neural networks (DNNs) by freezing certain network layers at constant parameter values can be of great value. These parameter values can be learned directly from other networks that have been trained on comparable tasks. The remaining portion of the network, which has been reduced in terms of parameters, can afterward undergo standard training for the intended purpose [163]. The topic of multimodal deep machine learning is anticipated to experience significant growth and interdisciplinary collaboration in the coming decade. This approach holds great potential as it enables the integration of diverse image sources to inform decision-making processes

6. Conclusions

The domain of adversarial attacks and defense is rapidly evolving within AI and machine learning. Addressing existing challenges and delving into future directions is imperative to constructing resilient, reliable, and secure AI systems capable of withstanding the intricacies of adversarial manipulation. The article provides an in-depth analysis of the current techniques employed for producing adversarial attacks to attack deep learning networks utilized in medical imaging and examines the many defense strategies that have been developed to identify and alleviate these perturbations. The examination of attacks and defenses on classification and segmentation models has been undertaken, with a particular focus on investigating the impact of neural network parameters on resistance and vulnerabilities. Collaborative efforts among researchers, practitioners, and policymakers will play a pivotal role in shaping the trajectory of adversarial attacks and defense, ensuring AI technologies' ethical and responsible deployment.

Author Contributions: Conceptualization, G.W.M.; methodology, C.C.U.; software, C.J.E.; validation, C.C.U., E.S.A.G. and A.M.; formal analysis, C.J.E.; investigation, A.K.M. and M.A.A.-a.; data curation, A.M.; writing—original draft preparation, G.W.M.; writing—review and editing, C.C.U.; visualization, A.K.M.; supervision, D.Y.; project administration, D.Y.; resources, and funding acquisition, A.A.; resources, writing—review and editing, and funding acquisition, M.A.A.-a. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Deanship of Scientific Research, King Khalid University, Saudi Arabia, under Grant number (RGP2/332/44).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets we used are publicly available and can be found in the references: Messidor dataset [168], The ISIC 2017 dataset [169], the Chest-ray 14 dataset [170], and the COVID-19 database [171].

Acknowledgments: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2022-00166402 and RS-2023-00256517). This work was supported by the Deanship of Scientific Research, King Khalid University, Saudi Arabia, under Grant number (RGP2/332/44).

Conflicts of Interest: The authors declare no conflict of interest. The sponsors had no role in the design, execution, interpretation, or writing of the study.

References

1. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition; Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
2. Zemskova, V.E.; He, T.L.; Wan, Z.; Grisouard, N. A deep-learning estimate of the decadal trends in the Southern Ocean carbon storage. *Nat. Commun.* **2022**, *13*, 4056. [CrossRef] [PubMed]
3. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3523–3542. [CrossRef] [PubMed]
4. Anaam, A.; Al-antari, M.A.; Hussain, J.; Abdel Samee, N.; Alabdulhafith, M.; Gofuku, A. Deep Active Learning for Automatic Mitotic Cell Detection on HEp-2 Specimen Medical Images. *Diagnostics* **2023**, *13*, 1416. [CrossRef]
5. Ge, Z.; Demyanov, S.; Chakravorty, R.; Bowling, A.; Garnavi, R. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In Proceedings of the Medical Image Computing and Computer Assisted Intervention−MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, 11–13 September 2017; pp. 250–258. [CrossRef]
6. Zhang, J.; Xie, Y.; Xia, Y.; Shen, C. Attention Residual Learning for Skin Lesion Classification. *IEEE Trans. Med. Imaging* **2019**, *38*, 2092–2103. [CrossRef] [PubMed]
7. Pereira, S.; Pinto, A.; Alves, V.; Silva, C.A. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* **2016**, *35*, 1240–1251. [CrossRef] [PubMed]
8. Muhammad, K.; Khan, S.; Del Ser, J.; Albuquerque, V.H.C.D. Deep Learning for Multigrade Brain Tumor Classification in Smart Healthcare Systems: A Prospective Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 507–522. [CrossRef] [PubMed]
9. Retinal Physician—Artificial Intelligence for the Screening of Diabetic Retinopathy. Available online: https://www.retinalphysician.com/issues/2022/november-december-2022/artificial-intelligence-for-the-screening-of-diabe (accessed on 20 August 2023).
10. Koohi-Moghadam, M.; Wang, H.; Wang, Y.; Yang, X.; Li, H.; Wang, J.; Sun, H. Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. *Nat. Mach. Intell.* **2019**, *1*, 561–567. [CrossRef]
11. Piloto, L.S.; Weinstein, A.; Battaglia, P.; Botvinick, M. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nat. Hum. Behav.* **2022**, *6*, 1257–1267. [CrossRef]
12. Paschali, M.; Conjeti, S.; Navarro, F.; Navab, N. Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*; Springer: Cham, Switzerland, 2018; Volume 11070, pp. 493–501. [CrossRef]
13. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial attacks on medical machine learning. *Science* **2019**, *363*, 1287–1289. [CrossRef]
14. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199v4.

15. Wang, Z.; Shu, X.; Wang, Y.; Feng, Y.; Zhang, L.; Yi, Z. A Feature Space-Restricted Attention Attack on Medical Deep Learning Systems. *IEEE Trans. Cybern.* **2022**, *53*, 5323–5335. [CrossRef] [PubMed]
16. Tian, B.; Guo, Q.; Juefei-Xu, F.; Le Chan, W.; Cheng, Y.; Li, X.; Xie, X.; Qin, S. Bias Field Poses a Threat to Dnn-Based X-Ray Recognition. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021. [CrossRef]
17. Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; Lu, F. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognit.* **2021**, *110*, 107332. [CrossRef]
18. Vatian, A.; Gusarova, N.; Dobrenko, N.; Dudorov, S.; Nigmatullin, N.; Shalyto, A.; Lobantsev, A. Impact of Adversarial Examples on the Efficiency of Interpretation and Use of Information from High-Tech Medical Images. In Proceedings of the 24th Conference of Open Innovations Association, Moscow, Russia, 8–12 April 2019. [CrossRef]
19. Zhou, H.-Y.; Wang, C.; Li, H.; Wang, G.; Zhang, S.; Li, W.; Yu, Y. SSMD: Semi-Supervised Medical Image Detection with Adaptive Consistency and Heterogeneous Perturbation. 2021. Available online: http://arxiv.org/abs/2106.01544 (accessed on 20 August 2023).
20. Xu, M.; Zhang, T.; Zhang, D. MedRDF: A Robust and Retrain-Less Diagnostic Framework for Medical Pretrained Models Against Adversarial Attack. *IEEE Trans. Med. Imaging* **2022**, *41*, 2130–2143. [CrossRef] [PubMed]
21. Taghanaki, S.A.; Abhishek, K.; Azizi, S.; Hamarneh, G. A Kernelized Manifold Mapping to Diminish the Effect of Adversarial Perturbations. 2019. Available online: http://arxiv.org/abs/1903.01015 (accessed on 20 August 2023).
22. Xu, M.; Zhang, T.; Li, Z.; Liu, M.; Zhang, D. Towards evaluating the robustness of deep diagnostic models by adversarial attack. *Med. Image Anal.* **2021**, *69*, 101977. [CrossRef]
23. Marinovich, M.L.; Wylie, E.; Lotter, W.; Lund, H.; Waddell, A.; Madeley, C.; Pereira, G.; Houssami, N. Artificial intelligence (AI) for breast cancer screening: BreastScreen population-based cohort study of cancer detection. *eBioMedicine* **2023**, *90*, 104498. [CrossRef]
24. Family Members Awarded $16.7 Million after Radiologist Missed. Available online: https://www.reliasmedia.com/articles/21632-family-members-awarded-16-7-million-after-radiologist-missed-evidence-of-lung-cancer (accessed on 27 September 2023).
25. Zbrzezny, A.M.; Grzybowski, A.E. Deceptive Tricks in Artificial Intelligence: Adversarial Attacks in Ophthalmology. *J. Clin. Med.* **2023**, *12*, 3266. [CrossRef]
26. Biggest Healthcare Data Breaches Reported This Year, So Far. Available online: https://healthitsecurity.com/features/biggest-healthcare-data-breaches-reported-this-year-so-far (accessed on 27 September 2023).
27. Kumar, A.; Kumar, D.; Kumar, P.; Dhawan, V. Optimization of Incremental Sheet Forming Process Using Artificial Intelligence-Based Techniques. *Nat.-Inspired Optim. Adv. Manuf. Process Syst.* **2020**, *8*, 113–130. [CrossRef]
28. Mukherjee, A.; Sumit; Deepmala; Dhiman, V.K.; Srivastava, P.; Kumar, A. Intellectual Tool to Compute Embodied Energy and Carbon Dioxide Emission for Building Construction Materials. *J. Phys. Conf. Ser.* **2021**, *1950*, 012025. [CrossRef]
29. Phogat, M.; Kumar, A.; Nandal, D.; Shokhanda, J. A Novel Automating Irrigation Techniques based on Artificial Neural Network and Fuzzy Logic. *J. Phys. Conf. Ser.* **2021**, *1950*, 012088. [CrossRef]
30. Ukwuoma, C.C.; Hossain, M.A.; Jackson, J.K.; Nneji, G.U.; Monday, H.N.; Qin, Z. Multi-Classification of Breast Cancer Lesions in Histopathological Images Using DEEP_Pachi: Multiple Self-Attention Head. *Diagnostics* **2022**, *12*, 1152. [CrossRef]
31. Ukwuoma, C.C.; Qin, Z.; Agbesi, V.K.; Ejiyi, C.J.; Bamisile, O.; Chikwendu, I.A.; Tienin, B.W.; Hossin, M.A. LCSB-inception: Reliable and effective light-chroma separated branches for Covid-19 detection from chest X-ray images. *Comput. Biol. Med.* **2022**, *150*, 106195. [CrossRef]
32. Ukwuoma, C.C.; Qin, Z.; Heyat, M.B.B.; Akhtar, F.; Smahi, A.; Jackson, J.K.; Furqan Qadri, S.; Muaad, A.Y.; Monday, H.N.; Nneji, G.U. Automated Lung-Related Pneumonia and COVID-19 Detection Based on Novel Feature Extraction Framework and Vision Transformer Approaches Using Chest X-ray Images. *Bioengineering* **2022**, *9*, 709. [CrossRef] [PubMed]
33. Ukwuoma, C.C.; Qin, Z.; Agbesi, V.K.; Cobbinah, B.M.; Yussif, S.B.; Abubakar, H.S.; Lemessa, B.D. Dual_Pachi: Attention-based dual path framework with intermediate second order-pooling for COVID-19 detection from chest X-ray images. *Comput. Biol. Med.* **2022**, *151*, 106324. [CrossRef] [PubMed]
34. Ritter, F.; Boskamp, T.; Homeyer, A.; Laue, H.; Schwier, M.; Link, F.; Peitgen, H.O. Medical image analysis. *IEEE Pulse* **2011**, *2*, 60–70. [CrossRef] [PubMed]
35. Phogat, M.; Kumar, D.; Phogat, M.; Kumar, D. Classification of Complex Diseases using an Improved Binary Cuckoo Search and Conditional Mutual Information Maximization. *Comput. Sist.* **2020**, *24*, 1121–1129. [CrossRef]
36. Ker, J.; Wang, L.; Rao, J.; Lim, T. Deep Learning Applications in Medical Image Analysis. *IEEE Access* **2017**, *6*, 9375–9379. [CrossRef]
37. Ukwuoma, C.C.; Cai, D.; Gati, E.S.; Agbesi, V.K.; Deribachew, G.; Yobsan Bayisa, L.; Abu, T. Attention-Based End-to-End Hybrid Ensemble Model for Breast Cancer Multi-Classification. *Off. Publ. Direct Res. J. Public Health Environ. Technol.* **2023**, *8*, 22–39.
38. Anaam, A.; Al-antari, M.A.; Gofuku, A. A deep learning self-attention cross residual network with Info-WGANGP for mitotic cell identification in HEp-2 medical microscopic images. *Biomed. Signal Process. Control* **2023**, *86*, 105191. [CrossRef]
39. Fraiwan, M.; Audat, Z.; Fraiwan, L.; Manasreh, T. Using deep transfer learning to detect scoliosis and spondylolisthesis from X-ray images. *PLoS ONE* **2022**, *17*, e0267851. [CrossRef]
40. Abdel-Monem, A.; Abouhawwash, M. A Machine Learning Solution for Securing the Internet of Things Infrastructures. *Sustain. Mach. Intell. J.* **2022**, *1*. [CrossRef]

41. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57. [CrossRef]
42. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. In Proceedings of the 3rd International Conference on Learning Representations ICLR 2015, San Diego, CA, USA, 7–9 May 2014; pp. 1–11.
43. Pranava Raman, B.M.S.; Anusree, V.; Sreeratcha, B.; Preeti Krishnaveni, R.A.; Dunston, S.D.; Rajam, M.A.V. Analysis of the Effect of Black Box Adversarial Attacks on Medical Image Classification Models. In Proceedings of the Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 11–12 August 2022; pp. 528–531. [CrossRef]
44. Tripathi, A.M.; Mishra, A. Fuzzy Unique Image Transformation: Defense Against Adversarial Attacks on Deep COVID-19 Models. *arXiv* **2020**, arXiv:2009.04004.
45. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 6th International Conference on Learning Representations ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2017. Available online: https://arxiv.org/abs/1706.06083v4 (accessed on 18 April 2023).
46. Kansal, K.; Krishna, P.S.; Jain, P.B.; Surya, R.; Honnavalli, P.; Eswaran, S. Defending against adversarial attacks on Covid-19 classifier: A denoiser-based approach. *Heliyon* **2022**, *8*, e11209. [CrossRef] [PubMed]
47. Paul, R.; Schabath, M.; Gillies, R.; Hall, L.; Goldgof, D. Mitigating Adversarial Attacks on Medical Image Understanding Systems. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging, Iowa City, IA, USA, 3–7 April 2020; pp. 1517–1521. [CrossRef]
48. Abdelhafeez, A.; Ali, A.M. DeepHAR-Net: A Novel Machine Intelligence Approach for Human Activity Recognition from Inertial Sensors. *Sustain. Mach. Intell. J.* **2022**, *1*. [CrossRef]
49. Abdelhafeez, A.; Aziz, A.; Khalil, N. Building a Sustainable Social Feedback Loop: A Machine Intelligence Approach for Twitter Opinion Mining. *Sustain. Mach. Intell. J.* **2022**, *1*. [CrossRef]
50. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust Physical-World Attacks on Deep Learning Visual Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitio, Seattle, WA, USA, 14–19 June 2018; pp. 1625–1634. [CrossRef]
51. Ozbulak, U.; Van Messem, A.; De Neve, W. Impact of Adversarial Examples on Deep Learning Models for Biomedical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*; Springer: Cham, Switzerland, 2019; Volume 11765, pp. 300–308. [CrossRef]
52. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582. [CrossRef]
53. Papernot, N.; Mcdaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the 2016 IEEE European Symposium on Security and Privacy, Saarbrucken, Germany, 21–26 March 2016; pp. 372–387. [CrossRef]
54. Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; Yuille, A. Adversarial Examples for Semantic Segmentation and Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
55. Finlayson, S.G.; Chung, H.W.; Kohane, I.S.; Beam, A.L. Adversarial Attacks Against Medical Deep Learning Systems. *arXiv* **2018**, arXiv:1804.05296.
56. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In Proceedings of the 5th International Conference on Learning Representations ICLR 2017, Toulon, France, 24–26 April 2017.
57. Asgari Taghanaki, S.; Das, A.; Hamarneh, G. Vulnerability analysis of chest X-ray image classification against adversarial attacks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*; Springer: Cham, Switzerland, 2018; Volume 11038, pp. 87–94. [CrossRef]
58. Yilmaz, I. Practical Fast Gradient Sign Attack against Mammographic Image Classifier. 2020. Available online: https://arxiv.org/abs/2001.09610v1 (accessed on 28 August 2023).
59. Ukwuoma, C.C.; Qin, Z.; Belal Bin Heyat, M.; Akhtar, F.; Bamisile, O.; Muaad, A.Y.; Addo, D.; Al-antari, M.A. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *J. Adv. Res.* **2023**, *48*, 191–211. [CrossRef]
60. Ukwuoma, C.C.; Cai, D.; Heyat, M.B.B.; Bamisile, O.; Adun, H.; Al-Huda, Z.; Al-antari, M.A. Deep learning framework for rapid and accurate respiratory COVID-19 prediction using chest X-ray images. *J. King Saud Univ.Comput. Inf. Sci.* **2023**, *35*, 101596. [CrossRef]
61. Rao, C.; Cao, J.; Zeng, R.; Chen, Q.; Fu, H.; Xu, Y.; Tan, M. A Thorough Comparison Study on Adversarial Attacks and Defenses for Common Thorax Disease Classification in Chest X-rays. 2020. Available online: https://arxiv.org/abs/2003.13969v1 (accessed on 20 August 2023).
62. Rahman, A.; Hossain, M.S.; Alrajeh, N.A.; Alsolami, F. Adversarial Examples—Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices. *IEEE Internet Things J.* **2021**, *8*, 9603. [CrossRef]
63. Cheng, G.; Ji, H. Adversarial Perturbation on MRI Modalities in Brain Tumor Segmentation. *IEEE Access* **2020**, *8*, 206009–206015. [CrossRef]

64. Chen, C.; Qin, C.; Qiu, H.; Ouyang, C.; Wang, S.; Chen, L.; Tarroni, G.; Bai, W.; Rueckert, D. Realistic adversarial data augmentation for mr image segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*; Springer: Cham, Switzerland, 2020; Volume 12261, pp. 667–677. [CrossRef]
65. Bertels, J.; Leuven, K.U.; Eelbode, T.; Vandermeulen, D.; Maes, F.; Berman, M.; Bisschops, R.; Blaschko, M.B. Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory & Practice Opportunistic Screening for Vertebral Compression Fractures in CT View Project Endoscopic and Transmural Evaluation of Healing in IBD and the Impact on Clinical. Available online: https://www.researchgate.net/publication/337048291 (accessed on 28 August 2023).
66. Feinman, R.; Curtin, R.R.; Shintre, S.; Gardner, A.B. Detecting Adversarial Samples from Artifacts. Available online: http://github.com/rfeinman/detecting-adversarial-samples (accessed on 28 August 2023).
67. Ma, X.; Li, B.; Wang, Y.; Erfani, S.M.; Wijewickrema, S.; Schoenebeck, G.; Song, D.; Houle, M.E.; Bailey, J. Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality. In Proceedings of the 6th International Conference on Learning Representations ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018. Available online: https://arxiv.org/abs/1801.02613v3 (accessed on 28 August 2023).
68. Lu, J.; Issaranon, T.; Lu, T.; Forsyth, D. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
69. Li, X.; Zhu, D. Robust Detection of Adversarial Attacks on Medical Images. Available online: https://github.com/xinli0928/MGM (accessed on 25 August 2023).
70. Li, X.; Pan, D.; Zhu, D. Defending against adversarial attacks on medical imaging ai system, classification or detection? In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging, Nice, France, 13–16 April 2021; pp. 1677–1681. [CrossRef]
71. Zhang, M.; Chen, Y.; Qian, C. Fooling Examples: Another Intriguing Property of Neural Networks. *Sensors* **2023**, *23*, 6378. [CrossRef]
72. Liu, C. Evaluating Robustness Against Adversarial Attacks: A Representational Similarity Analysis Approach. In Proceedings of the International Joint Conference on Neural Networks 2023, Gold Coast, Australia, 18–23 June 2023. [CrossRef]
73. Ren, K.; Zheng, T.; Qin, Z.; Liu, X. Adversarial Attacks and Defenses in Deep Learning. *Engineering* **2020**, *6*, 346–360. [CrossRef]
74. Sen, J.; Dasgupta, S. Adversarial Attacks on Image Classification Models: FGSM and Patch Attacks and their Impact. *arXiv* **2023**, arXiv:2307.02055. [CrossRef]
75. Shah, A.; Lynch, S.; Niemeijer, M.; Amelon, R.; Clarida, W.; Folk, J.; Russell, S.; Wu, X.; Abramoff, M.D. Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 1454–1457. [CrossRef]
76. Liu, Z.; Zhang, J.; Jog, V.; Loh, P.L.; McMillan, A.B. Robustifying Deep Networks for Medical Image Segmentation. *J. Digit. Imaging* **2021**, *34*, 1279. [CrossRef]
77. Chen, L.; Bentley, P.; Mori, K.; Misawa, K.; Fujiwara, M.; Rueckert, D. Intelligent image synthesis to attack a segmentation CNN using adversarial learning. In *Simulation and Synthesis in Medical Imaging: 4th International Workshop, SASHIMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 13 October 2019, Proceedings 4*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11827, pp. 90–99. [CrossRef]
78. Kovalev, V.; Voynov, D. Influence of Control Parameters and the Size of Biomedical Image Datasets on the Success of Adversarial Attacks. 2019. Available online: https://arxiv.org/abs/1904.06964v1 (accessed on 25 August 2023).
79. Cheng, Y.; Juefei-Xu, F.; Guo, Q.; Fu, H.; Xie, X.; Lin, S.-W.; Lin, W.; Liu, Y. Adversarial Exposure Attack on Diabetic Retinopathy Imagery. 2020. Available online: https://arxiv.org/abs/2009.09231v1 (accessed on 25 August 2023).
80. Byra, M.; Styczynski, G.; Szmigielski, C.; Kalinowski, P.; Michalowski, L.; Paluszkiewicz, R.; Ziarkiewicz-Wroblewska, B.; Zieniewicz, K.; Nowicki, A. Adversarial attacks on deep learning models for fatty liver disease classification by modification of ultrasound image reconstruction method. In Proceedings of the 2020 IEEE International Ultrasonics Symposium 2020, Las Vegas, NV, USA, 7–11 September 2020. [CrossRef]
81. Yao, Q.; He, Z.; Han, H.; Zhou, S.K. Miss the Point: Targeted Adversarial Attack on Multiple Landmark Detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020, Proceedings, Part IV 23*; Springer International Publishing: Cham, Switzerland, 2020; Volume 12264, pp. 692–702. [CrossRef]
82. Yoo, T.K.; Choi, J.Y. Outcomes of Adversarial Attacks on Deep Learning Models for Ophthalmology Imaging Domains. *JAMA Ophthalmol.* **2020**, *138*, 1213. [CrossRef] [PubMed]
83. Hirano, H.; Minagi, A.; Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med. Imaging* **2021**, *21*, 1–13. [CrossRef]
84. Joel, M.Z.; Umrao, S.; Chang, E.; Choi, R.; Yang, D.; Omuro, A.; Herbst, R.; Krumholz, H.; Aneja, S. Adversarial Attack Vulnerability of Deep Learning Models for Oncologic Images. *medRxiv* **2021**. [CrossRef]
85. Chen, J.; Qian, L.; Urakov, T.; Gu, W.; Liang, L. Adversarial Robustness Study of Convolutional Neural Network for Lumbar Disk Shape Reconstruction from MR Images. 2021. Available online: https://arxiv.org/abs/2102.02885v1 (accessed on 25 August 2023).
86. Qi, G.; Gong, L.; Song, Y.; Ma, K.; Zheng, Y. Stabilized Medical Image Attacks. *arXiv* **2021**, arXiv:2103.05232.

87. Bortsova, G.; Dubost, F.; Hogeweg, L.; Katramados, I.; de Bruijne, M. Adversarial Heart Attack: Neural Networks Fooled to Segment Heart Symbols in Chest X-ray Images. 2021. Available online: https://arxiv.org/abs/2104.00139v2 (accessed on 25 August 2023).
88. Kovalev, V.A.; Liauchuk, V.A.; Voynov, D.M.; Tuzikov, A.V. Biomedical Image Recognition in Pulmonology and Oncology with the Use of Deep Learning. *Pattern Recognit. Image Anal.* **2021**, *31*, 144–162. [CrossRef]
89. Pal, B.; Gupta, D.; Rashed-Al-mahfuz, M.; Alyami, S.A.; Moni, M.A. Vulnerability in Deep Transfer Learning Models to Adversarial Fast Gradient Sign Attack for COVID-19 Prediction from Chest Radiography Images. *Appl. Sci.* **2021**, *11*, 4233. [CrossRef]
90. Shao, M.; Zhang, G.; Zuo, W.; Meng, D. Target attack on biomedical image segmentation model based on multi-scale gradients. *Inf. Sci.* **2021**, *554*, 33–46. [CrossRef]
91. Wang, X.; Lv, S.; Sun, J.; Wang, S. Adversarial Attacks Medical Diagnosis Model with Generative Adversarial Networks. *Lect. Notes Data Eng. Commun. Technol.* **2022**, *89*, 678–685. [CrossRef]
92. Minagi, A.; Hirano, H.; Takemoto, K. Natural Images Allow Universal Adversarial Attacks on Medical Image Classification Using Deep Neural Networks with Transfer Learning. *J. Imaging* **2022**, *8*, 38. [CrossRef] [PubMed]
93. Patel, P.; Bhadla, M.; Upadhyay, J.; Suthar, D.; Darji, D. Predictive COVID-19 Risk and Virus Mutation isolation with CNN based Machine learning Technique. In Proceedings of the 2022 2nd International Conference on Innovative Practices in Technology and Management, Pradesh, India, 23–25 February 2022; pp. 424–428. [CrossRef]
94. Levy, M.; Amit, G.; Elovici, Y.; Mirsky, Y. The Security of Deep Learning Defences for Medical Imaging. *arXiv* **2022**, arXiv:2201.08661.
95. Kwon, H.; Jeong, J. AdvU-Net: Generating Adversarial Example Based on Medical Image and Targeting U-Net Model. *J. Sensors* **2022**, *2022*, 4390413. [CrossRef]
96. Júlio de Aguiar, E.; Marcomini, K.D.; Antunes Quirino, F.; Gutierrez, M.A.; Traina, C.; Traina, A.J.M. Evaluation of the impact of physical adversarial attacks on deep learning models for classifying COVID cases. In *Medical Imaging 2022: Computer-Aided Diagnosis*; SPIE: Bellingham, WA, USA, 2022; p. 122. [CrossRef]
97. Apostolidis, K.D.; Papakostas, G.A. Digital Watermarking as an Adversarial Attack on Medical Image Analysis with Deep Learning. *J. Imaging* **2022**, *8*, 155. [CrossRef]
98. Wei, C.; Sun, R.; Li, P.; Wei, J. Analysis of the No-sign Adversarial Attack on the COVID Chest X-ray Classification. In Proceedings of the 2022 International Conference on Image Processing and Media Computing (ICIPMC 2022), Xi'an, China, 27–29 May 2022; pp. 73–79. [CrossRef]
99. Selvakkumar, A.; Pal, S.; Jadidi, Z. Addressing Adversarial Machine Learning Attacks in Smart Healthcare Perspectives. *Lect. Notes Electr. Eng.* **2022**, *886*, 269–282. [CrossRef]
100. Ahmed, S.; Dera, D.; Hassan, S.U.; Bouaynaya, N.; Rasool, G. Failure Detection in Deep Neural Networks for Medical Imaging. *Front. Med. Technol.* **2022**, *4*, 919046. [CrossRef]
101. Li, S.; Huang, G.; Xu, X.; Lu, H. Query-based black-box attack against medical image segmentation model. *Futur. Gener. Comput. Syst.* **2022**, *133*, 331–337. [CrossRef]
102. Morshuis, J.N.; Gatidis, S.; Hein, M.; Baumgartner, C.F. Adversarial Robustness of MR Image Reconstruction Under Realistic Perturbations. In *International Workshop on Machine Learning for Medical Image Reconstruction*; Springer: Cham, Switzerland, 2022; Volume 13587, pp. 24–33. [CrossRef]
103. Bharath Kumar, D.P.; Kumar, N.; Dunston, S.D.; Rajam, V.M.A. Analysis of the Impact of White Box Adversarial Attacks in ResNet While Classifying Retinal Fundus Images. In *International Conference on Computational Intelligence in Data Science*; Springer: Cham, Switzerland, 2022; Volume 654, pp. 162–175.
104. Purohit, J.; Attari, S.; Shivhare, I.; Surtkar, S.; Jogani, V. Adversarial Attacks and Defences for Skin Cancer Classification. *arXiv* **2022**, arXiv:2212.06822. [CrossRef]
105. Li, Y.; Liu, S. The Threat of Adversarial Attack on a COVID-19 CT Image-Based Deep Learning System. *Bioengineering* **2023**, *10*, 194. [CrossRef]
106. Dai, Y.; Qian, Y.; Lu, F.; Wang, B.; Gu, Z.; Wang, W.; Wan, J.; Zhang, Y. Improving adversarial robustness of medical imaging systems via adding global attention noise. *Comput. Biol. Med.* **2023**, *164*, 107251. [CrossRef] [PubMed]
107. Joel, M.Z.; Avesta, A.; Yang, D.X.; Zhou, J.G.; Omuro, A.; Herbst, R.S.; Krumholz, H.M.; Aneja, S. Comparing Detection Schemes for Adversarial Images against Deep Learning Models for Cancer Imaging. *Cancers* **2023**, *15*, 1548. [CrossRef] [PubMed]
108. Niu, Z.H.; Yang, Y. Bin Defense Against Adversarial Attacks with Efficient Frequency-Adaptive Compression and Reconstruction. *Pattern Recognit.* **2023**, *138*, 109382. [CrossRef]
109. Bountakas, P.; Zarras, A.; Lekidis, A.; Xenakis, C. Defense strategies for Adversarial Machine Learning: A survey. *Comput. Sci. Rev.* **2023**, *49*, 100573. [CrossRef]
110. Laykaviriyakul, P.; Phaisangittisagul, E. Collaborative Defense-GAN for protecting adversarial attacks on classification system. *Expert Syst. Appl.* **2023**, *214*, 118957. [CrossRef]
111. Chen, F.; Wang, J.; Liu, H.; Kong, W.; Zhao, Z.; Ma, L.; Liao, H.; Zhang, D. Frequency constraint-based adversarial attack on deep neural networks for medical image classification. *Comput. Biol. Med.* **2023**, *164*, 107248. [CrossRef]

112. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security 2017, Abu Dhabi, United Arab Emirates, 2–6 April 2017; pp. 506–519. [CrossRef]
113. Ilyas, A.; Engstrom, L.; Athalye, A.; Lin, J. Black-Box Adversarial Attacks with Limited Queries and Information. PMLR; pp. 2137–2146. Available online: https://proceedings.mlr.press/v80/ilyas18a.html (accessed on 26 August 2023).
114. Wicker, M.; Huang, X.; Kwiatkowska, M. Feature-Guided Black-Box Safety Testing of Deep Neural Networks. In *Tools and Algorithms for the Construction and Analysis of Systems: 24th International Conference, TACAS 2018, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2018, Thessaloniki, Greece, 14–20 April 2018, Proceedings, Part I 24*; Springer: Cham, Switzerland, 2017; Volume 10805, pp. 408–426. [CrossRef]
115. Andriushchenko, M.; Croce, F.; Flammarion, N.; Hein, M. Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; Volume 12368, pp. 484–501. [CrossRef]
116. Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; Madry, A. Adversarial Examples Are Not Bugs, They Are Features. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 125–136.
117. Yao, Q.; He, Z.; Zhou, S.K. Medical Aegis: Robust Adversarial Protectors for Medical Images. November 2021. Available online: https://arxiv.org/abs/2111.10969v4 (accessed on 20 August 2023).
118. Liao, F.; Liang, M.; Dong, Y.; Pang, T.; Hu, X.; Zhu, J. Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser. Available online: https://github.com/lfz/Guided-Denoise (accessed on 20 August 2023).
119. Daanouni, O.; Cherradi, B.; Tmiri, A. NSL-MHA-CNN: A Novel CNN Architecture for Robust Diabetic Retinopathy Prediction Against Adversarial Attacks. *IEEE Access* **2022**, *10*, 103987–103999. [CrossRef]
120. Han, T.; Nebelung, S.; Pedersoli, F.; Zimmermann, M.; Schulze-Hagen, M.; Ho, M.; Haarburger, C.; Kiessling, F.; Kuhl, C.; Schulz, V.; et al. Advancing diagnostic performance and clinical usability of neural networks via adversarial training and dual batch normalization. *Nat. Commun.* **2021**, *12*, 1–11. [CrossRef]
121. Chen, L.; Zhao, L.; Chen, C.Y.C. Enhancing adversarial defense for medical image analysis systems with pruning and attention mechanism. *Med. Phys.* **2021**, *48*, 6198–6212. [CrossRef]
122. Xue, F.F.; Peng, J.; Wang, R.; Zhang, Q.; Zheng, W.S. Improving Robustness of Medical Image Diagnosis with Denoising Convolutional Neural Networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019, Proceedings, Part VI 22*; Springer: Cham, Switzerland, 2019; Volume 11769, pp. 846–854. [CrossRef]
123. Xie, C.; Tan, M.; Gong, B.; Wang, J.; Yuille, A.; Le, Q. V Adversarial Examples Improve Image Recognition. Available online: https://github.com/tensorflow/tpu/tree/ (accessed on 25 August 2023).
124. Carannante, G.; Dera, D.; Bouaynaya, N.C.; Fathallah-Shaykh, H.M.; Rasool, G. SUPER-Net: Trustworthy Medical Image Segmentation with Uncertainty Propagation in Encoder-Decoder Networks. *arXiv* **2021**, arXiv:2111.05978.
125. Stimpel, B.; Syben, C.; Schirrmacher, F.; Hoelter, P.; Dörfler, A.; Maier, A. Multi-modal Deep Guided Filtering for Comprehensible Medical Image Processing. *IEEE Trans. Med. Imaging* **2019**, *39*, 1703–1711. [CrossRef] [PubMed]
126. He, X.; Yang, S.; Li, G.; Li, H.; Chang, H.; Yu, Y. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8417–8424. [CrossRef]
127. Joel, M.Z.; Umrao, S.; Chang, E.; Choi, R.; Yang, D.X.; Duncan, J.S.; Omuro, A.; Herbst, R.; Krumholz, H.M.; Aneja, S. Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology. *JCO Clin. Cancer Inform.* **2022**, *6*, e2100170. [CrossRef] [PubMed]
128. Hu, L.; Zhou, D.W.; Guo, X.Y.; Xu, W.H.; Wei, L.M.; Zhao, J.G. Adversarial training for prostate cancer classification using magnetic resonance imaging. *Quant. Imaging Med. Surg.* **2022**, *12*, 3276–3287. [CrossRef] [PubMed]
129. Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; Gu, Q. Improving Adversarial Robustness Requires Revisiting Misclassified Examples. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020; pp. 1–14.
130. Liu, S.; Setio, A.A.A.; Ghesu, F.C.; Gibson, E.; Grbic, S.; Georgescu, B.; Comaniciu, D. No Surprises: Training Robust Lung Nodule Detection for Low-Dose CT Scans by Augmenting with Adversarial Attacks. *IEEE Trans. Med. Imaging* **2021**, *40*, 335–345. [CrossRef] [PubMed]
131. Lal, S.; Rehman, S.U.; Shah, J.H.; Meraj, T.; Rauf, H.T.; Damaševičius, R.; Mohammed, M.A.; Abdulkareem, K.H. Adversarial Attack and Defence through Adversarial Training and Feature Fusion for Diabetic Retinopathy Recognition. *Sensors* **2021**, *21*, 3922. [CrossRef]
132. Almalik, F.; Yaqub, M.; Nandakumar, K. Self-Ensembling Vision Transformer (SEViT) for Robust Medical Image Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Cham, Switzerland, 2022; Volume 13433, pp. 376–386. [CrossRef]
133. Huang, Y.; Würfl, T.; Breininger, K.; Liu, L.; Lauritsch, G.; Maier, A. Some Investigations on Robustness of Deep Learning in Limited Angle Tomography. *Inform. Aktuell* **2019**, *17*, 21. [CrossRef]

134. Ren, X.; Zhang, L.; Wei, D.; Shen, D.; Wang, Q. Brain MR Image Segmentation in Small Dataset with Adversarial Defense and Task Reorganization. In *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, 13 October 2019, Proceedings 10*; Springer Nature: Cham, Switzerland, 2019; Volume 11861, pp. 1–8. [CrossRef]
135. Thite, M.; Kavanagh, M.J.; Johnson, R.D. Evolution of human resource management & human resource information systems: The role of information technology. In *Human Resource Information Systems: Basics*; Kavanagh, M.J., Thite, M., Johnson, R.D., Eds.; Applications & Directions: Thousand Oaks, CA, USA, 2012; pp. 2–34. Available online: https://www.researchgate.net/publication/277249737_Thite_M_Kavanagh_MJ_Johnson_R_D_2012_Evolution_of_human_resource_management_human_resource_information_systems_The_role_of_information_technology_In_Kavanagh_MJ_Thite_M_Johnson_R_D_Eds_Human_Resource_ (accessed on 30 May 2023).
136. Li, Y.; Zhu, Z.; Zhou, Y.; Xia, Y.; Shen, W.; Fishman, E.K.; Yuille, A.L. 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*; Springer Nature: Cham, Switzerland, 2019; pp. 69–91. [CrossRef]
137. Park, H.; Bayat, A.; Sabokrou, M.; Kirschke, J.S.; Menze, B.H. Robustification of Segmentation Models Against Adversarial Perturbations in Medical Imaging. In *Predictive Intelligence in Medicine: Third International Workshop, PRIME 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, 8 October 2020, Proceedings 3*; Springer Nature: Cham, Switzerland, 2020; Volume 12329, pp. 46–57. [CrossRef]
138. Li, Y.; Zhang, H.; Bermudez, C.; Chen, Y.; Landman, B.A.; Vorobeychik, Y. Anatomical Context Protects Deep Learning from Adversarial Perturbations in Medical Imaging. *Neurocomputing* **2020**, *379*, 370–378. [CrossRef]
139. Wu, D.; Liu, S.; Ban, J. Classification of Diabetic Retinopathy Using Adversarial Training. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *806*, 012050. [CrossRef]
140. Anand, D.; Tank, D.; Tibrewal, H.; Sethi, A. Self-Supervision vs. Transfer Learning: Robust Biomedical Image Analysis against Adversarial Attacks. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging, Iowa City, IA, USA, 3–7 April 2020; pp. 1159–1163. [CrossRef]
141. Ma, L.; Liang, L. Increasing-Margin Adversarial (IMA) training to Improve Adversarial Robustness of Neural Networks. *Comput. Methods Programs Biomed.* **2023**, *240*, 107687. [CrossRef]
142. Cheng, K.; Calivá, F.; Shah, R.; Han, M.; Majumdar, S.; Pedoia, V. Addressing The False Negative Problem of Deep Learning MRI Reconstruction Models by Adversarial Attacks and Robust Training 2020, PMLR, 21 September 2020; pp. 121–135. Available online: https://proceedings.mlr.press/v121/cheng20a.html (accessed on 21 September 2023).
143. Raj, A.; Bresler, Y.; Li, B. Improving Robustness of Deep-Learning-Based Image Reconstruction 2020; pp. 7932–7942. Available online: https://proceedings.mlr.press/v119/raj20a.html (accessed on 25 August 2023).
144. Huq, A.; Pervin, T. Analysis of Adversarial Attacks on Skin Cancer Recognition. In Proceedings of the 2020 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 5–6 August 2020. [CrossRef]
145. Liu, Q.; Jiang, H.; Liu, T.; Liu, Z.; Li, S.; Wen, W.; Shi, Y. Defending Deep Learning-Based Biomedical Image Segmentation from Adversarial Attacks: A Low-Cost Frequency Refinement Approach. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020, Proceedings, Part IV 23*; Springer Nature: Cham, Switzerland, 2020; Volume 12264, pp. 342–351. [CrossRef]
146. Watson, M.; Al Moubayed, N. Attack-agnostic adversarial detection on medical data using explainable machine learning. In Proceedings of the 2020 25th International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2021; pp. 8180–8187. [CrossRef]
147. Pervin, M.T.; Tao, L.; Huq, A.; He, Z.; Huo, L. Adversarial Attack Driven Data Augmentation for Accurate And Robust Medical Image Segmentation. 2021. Available online: http://arxiv.org/abs/2105.12106 (accessed on 25 August 2023).
148. Uwimana1, A.; Senanayake, R. Out of Distribution Detection and Adversarial Attacks on Deep Neural Networks for Robust Medical Image Analysis. 2021. Available online: http://arxiv.org/abs/2107.04882 (accessed on 25 August 2023).
149. Daza, L.; Pérez, J.C.; Arbeláez, P. Towards Robust General Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part III 24*; Springer Nature: Cham, Switzerland, 2021; Volume 12903, pp. 3–13. [CrossRef]
150. Gupta, D.; Pal, B. Vulnerability Analysis and Robust Training with Additive Noise for FGSM Attack on Transfer Learning-Based Brain Tumor Detection from MRI. *Lect. Notes Data Eng. Commun. Technol.* **2022**, *95*, 103–114. [CrossRef]
151. Yang, Y.; Shih, F.Y.; Roshan, U. Defense Against Adversarial Attacks Based on Stochastic Descent Sign Activation Networks on Medical Images. *Int. J. Pattern Recognit. Artif. Intell.* **2022**, *36*, 2254005. [CrossRef]
152. Alatalo, J.; Sipola, T.; Kokkonen, T. Detecting One-Pixel Attacks Using Variational Autoencoders. In Proceedings of the World Conference on Information Systems and Technologies, Budva, Montenegro, 12–14 April 2022; Volume 468, pp. 611–623. [CrossRef]
153. Rodriguez, D.; Nayak, T.; Chen, Y.; Krishnan, R.; Huang, Y. On the role of deep learning model complexity in adversarial robustness for medical images. *BMC Med. Inform. Decis. Mak.* **2022**, *22*, 1–15. [CrossRef] [PubMed]
154. Ma, L.; Liang, L. Adaptive Adversarial Training to Improve Adversarial Robustness of DNNs for Medical Image Segmentation and Detection. 2022. Available online: https://arxiv.org/abs/2206.01736v2 (accessed on 25 August 2023).
155. Xie, Y.; Fetit, A.E. How Effective is Adversarial Training of CNNs in Medical Image Analysis? In *Annual Conference on Medical Image Understanding and Analysis*; Springer: Cham, Switzerland, 2022; Volume 13413, pp. 443–457. [CrossRef]

156. Wang, Y.; Li, Y.; Shen, Z. Fight Fire with Fire: Reversing Skin Adversarial Examples by Multiscale Diffusive and Denoising Aggregation Mechanism. *arXiv* **2022**, arXiv:2208.10373. [CrossRef]
157. Ghaffari Laleh, N.; Truhn, D.; Veldhuizen, G.P.; Han, T.; van Treeck, M.; Buelow, R.D.; Langer, R.; Dislich, B.; Boor, P.; Schulz, V.; et al. Adversarial attacks and adversarial robustness in computational pathology. *Nat. Commun.* **2022**, *13*, 1–10. [CrossRef]
158. Maliamanis, T.V.; Apostolidis, K.D.; Papakostas, G.A. How Resilient Are Deep Learning Models in Medical Image Analysis? The Case of the Moment-Based Adversarial Attack (Mb-AdA). *Biomedicines* **2022**, *10*, 2545. [CrossRef]
159. Sun, S.; Xian, M.; Vakanski, A.; Ghanem, H. MIRST-DM: Multi-instance RST with Drop-Max Layer for Robust Classification of Breast Cancer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2022; Volume 13434, pp. 401–410. [CrossRef]
160. Pandey, P.; Vardhan, A.; Chasmai, M.; Sur, T.; Lall, B. Adversarially Robust Prototypical Few-Shot Segmentation with Neural-ODEs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2022; Volume 13438, pp. 77–87. [CrossRef]
161. Roh, J. Impact of Adversarial Training on the Robustness of Deep Neural Networks. In Proceedings of the 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 23–25 September 2020; pp. 560–566. [CrossRef]
162. Le, L.D.; Fu, H.; Xu, X.; Liu, Y.; Xu, Y.; Du, J.; Zhou, J.T.; Goh, R. An Efficient Defending Mechanism Against Image Attacking on Medical Image Segmentation Models. In *MICCAI Workshop on Resource-Efficient Medical Image Analysis*; Springer: Cham, Switzerland, 2022; Volume 13543, pp. 65–74. [CrossRef]
163. Chen, C.; Qin, C.; Ouyang, C.; Li, Z.; Wang, S.; Qiu, H.; Chen, L.; Tarroni, G.; Bai, W.; Rueckert, D. Enhancing MR image segmentation with realistic adversarial data augmentation. *Med. Image Anal.* **2022**, *82*, 102597. [CrossRef]
164. Shi, X.; Peng, Y.; Chen, Q.; Keenan, T.; Thavikulwat, A.T.; Lee, S.; Tang, Y.; Chew, E.Y.; Summers, R.M.; Lu, Z. Robust convolutional neural networks against adversarial attacks on medical images. *Pattern Recognit.* **2022**, *132*, 108923. [CrossRef]
165. Sitawarin, C. DARTS: Deceiving Autonomous Cars with Toxic Signs. CoRR 2018, abs/1802.06430. Available online: http://arxiv.org/abs/1802.06430 (accessed on 28 August 2023).
166. Su, J.; Vargas, D.V.; Sakurai, K. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 828–841. [CrossRef]
167. Croce, F.; Hein, M. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks. In Proceedings of the International Conference on Machine Learning, virtual, 12–18 July 2020.
168. Decencière, E.; Zhang, X.; Cazuguel, G.; Lay;, B.; Cochener, B.; Trone, C.; Gain, P.; Ordóñez-Varela, J.R.; Massin, P.; Erginay, A.; et al. Feedback On a Publicly Distributed Image Database: The Messidor Database. *Image Anal. Stereol.* **2014**, *33*, 231–234. [CrossRef]
169. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). 2019. Available online: https://arxiv.org/abs/1902.03368v2 (accessed on 28 August 2023).
170. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3462–3471. [CrossRef]
171. Chowdhury, M.E.H.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M.A.; Mahbub, Z.B.; Islam, K.R.; Khan, M.S.; Iqbal, A.; Al-Emadi, N.; et al. Can AI help in screening Viral and COVID-19 pneumonia? *IEEE Access* **2020**, *8*, 132665–132676. [CrossRef]
172. Gonzalez, T.F. *Handbook of Approximation Algorithms and Metaheuristics*; CRC Press: Boca Raton, FL, USA, 2007. [CrossRef]
173. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015. Available online: https://arxiv.org/abs/1409.1556v6 (accessed on 29 August 2023).
174. Bortsova, G.; González-Gonzalo, C.; Wetstein, S.C.; Dubost, F.; Katramados, I.; Hogeweg, L.; Liefers, B.; van Ginneken, B.; Pluim, J.P.W.; Veta, M.; et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Med. Image Anal.* **2021**, *73*, 102141. [CrossRef] [PubMed]
175. Han, C.; Rundo, L.; Murao, K.; Nemoto, T.; Nakayama, H. Bridging the Gap Between AI and Healthcare Sides: Towards Developing Clinically Relevant AI-Powered Diagnosis Systems. In Proceedings of the 16th IFIP WG 12.5 International Conference, Neos Marmaras, Greece, 5–7 June 2020; pp. 320–333. [CrossRef]
176. Saeed, W.; Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowl. Based Syst.* **2023**, *263*, 110273. [CrossRef]
177. Busnatu, Ștefan; Niculescu, A.G.; Bolocan, A.; Petrescu, G.E.D.; Păduraru, D.N.; Năstasă, I.; Lupușoru, M.; Geantă, M.; Andronic, O.; Grumezescu, A.M.; et al. Clinical Applications of Artificial Intelligence—An Updated Overview. *J. Clin. Med.* **2022**, *11*, 2265. [CrossRef]
178. Paul, M.; Maglaras, L.; Ferrag, M.A.; Almomani, I. Digitization of healthcare sector: A study on privacy and security concerns. *ICT Express* **2023**, *9*, 571–588. [CrossRef]
179. Wang, Y.; Sun, T.; Li, S.; Yuan, X.; Ni, W.; Hossain, E.; Poor, H.V. Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey. 2023. Available online: https://arxiv.org/abs/2303.06302v1 (accessed on 27 September 2023).

180. Ding, C.; Sun, S.; Zhao, J. Multi-Modal Adversarial Example Detection with Transformer. In Proceedings of the 2022 International Joint Conference on Neural Networks, Padua, Italy, 18–23 July 2022. [CrossRef]
181. Cao, H.; Zou, W.; Wang, Y.; Song, T.; Liu, M. Emerging Threats in Deep Learning-Based Autonomous Driving: A Comprehensive Survey. 2022. Available online: https://arxiv.org/abs/2210.11237v1 (accessed on 27 September 2023).