広島大学

Research on Adversarial Attacks and Enhancing Defense Capabilities in Medical Image Deep Learning Systems

医療画像ディープラーニングシステムにおける敵対的攻撃と防御能力の強化に関する研究

[LI YANG]

Graduate School of Advanced Science and Engineering
Informatics and Data Science

Hiroshima University
September, 2024

ABSTRACT

With the widespread application of deep learning technology in the medical field, it has demonstrated excellent performance in areas such as medical image analysis, drug development, and clinical decision support. Deep learning systems used for medical image analysis can learn effective feature representations from large-scale datasets, providing accurate and rapid diagnoses. Adversarial attacks, by adding slight perturbations to the original images, can cause models to make severe classification errors. However, medical images often contain significant amounts of noise, making them potential targets for adversarial attacks. Such attacks are particularly critical in the fields of life and health, potentially leading to incorrect diagnoses and treatment decisions, thereby threatening patient safety and health. Therefore, it has become a key research focus to test the security of deep learning systems based on medical images and to develop effective defense mechanisms to enhance their security and reliability.

This research investigates adversarial attacks and defenses in deep learning systems based on medical images. We tested the security vulnerabilities in deep learning systems for medical images using both white box and black box adversarial attacks. Subsequently, we proposed more effective defense strategies to defend against these adversarial attacks.

Specifically, we used white box adversarial attack algorithms to attack different deep learning systems based on medical images, testing the security of the systems. We found that deep learning systems based on medical images are vulnerable to attacks, leading to misclassification of medical images. Subsequently, we proposed better defense methods against white box adversarial attacks to defense against these attacks. Furthermore, to better simulate real-world scenarios, we proposed a black box adversarial attack algorithm and tested it on different medical deep learning systems. This algorithm demonstrated superior attack capabilities. Finally, to defend against black box adversarial attacks, we proposed a defense mechanism with improved performance, thereby enhancing the security of deep learning systems based on medical images.

KEY WORDS： Adversarial attack, defense, deep learning, security, medical image

Table of Contents

1 Introduction

1.1 Deep Learning in medical image analysis

As early as the 1950s, British mathematician Alan Turing's research on computer thinking laid the main theoretical foundation for artificial intelligence. He proposed the famous "Turing Test", which is a question-and-answer method to determine whether a computer has the level of human intelligence [1]. In 1956, the first symposium on artificial intelligence in history was held at Dartmouth College in the United States, which is considered to be the symbol of the birth of artificial intelligence. At the meeting, John McCarthy for the first time put forward the concept of "artificial intelligence", the goal of artificial intelligence is to realize a machine that can use knowledge to solve problems like humans [2]. In recent years, with the rapid development of information technology, Artificial Intelligence (AI) has been a great success [3].

Machine Learning (ML) is a sub-field of Artificial Intelligence (AI) [4]. Its core goal is that computer systems can be able to perform tasks by learning experiences from data. With the popularization of the Internet and the rapid development of digital technology to produce large-scale data, this data explosion for machine learning provides a huge data resource, while thanks to the continuous improvement of hardware performance, the machine learning algorithm's training speed has been significantly improved, thus making its rapid development [5].

Deep Learning (DL) is a subset of Machine Learning [6]. The roots of Deep Learning can be traced back to artificial neural network models in the 1940s, but in the past decades, Deep Learning has not been widely used due to the difficulty of training deep neural networks, computational power, and computer hardware limitations. However, with the improvement of hardware performance and the availability of large-scale data, deep learning has finally gained rapid development and application.

Deep neural networks have shown powerful capabilities for image recognition and classification [7]. However, Szegedy found that deep neural networks have fatal weaknesses in image classification tasks when adding perturbations to the input image that are difficult to detect with the eye, which can cause the models to generate classification errors [8]. An adversarial sample is defined as follows. An adversarial sample is generated by applying subtle perturbations (that are difficult to detect by the naked eye but are acceptable to the deep learning model) to the original data, leading to the input data being misjudged by the deep learning model. The input data are denoted by *x*, the deep learning model is denoted by *g*, the classification result is denoted by *g(x)*, and the perturbation is denoted by ϵ. Suppose there is a slight perturbation ϵ:

$$|\,\epsilon\,| < \delta \ and\ g\,(x + \epsilon) \ \neq \ g(x)$$

Then, $x + \epsilon$ can be called an adversarial sample.

Deep learning technology has gradually entered the field of medical image analysis and has brought great changes and innovations to the fields of medical detection, diagnosis, and assisted treatment. In the field of medical imaging, it has greatly improved doctors' work efficiency and diagnostic accuracy by quickly and accurately analyzing medical images such as X-ray images, Computed Tomography images, and MIR images. However, with the wide application of AI technology in various fields, its security issues are gradually exposed. Research has shown that there are huge security loopholes in AI systems, for example, in the classification task of image recognition, people deliberately add slight perturbations in the input image, although such perturbations are difficult to distinguish and detect with the naked eye, it will likely lead to classification errors in the model. As compared with other deep learning systems, the security of medical image-based deep learning systems is crucial, because medical images are related to personal health data, and the correct recognition and analysis of medical images through deep learning systems is an important tool to assist doctors in diagnosis, and any wrong diagnosis about a condition may bring irreparable harm. Additionally, dishonest individuals could attack the medical deep learning system to tamper with the outcomes of medical image diagnosis, which would subsequently lead to insurance fraud. Therefore, in this paper, we study the security of medical image deep learning system.

1.2 Contributions

1. We tested the security of medical image deep learning systems by employing white box adversarial attack. It was discovered that adversarial attack algorithms are capable

of compromising medical deep learning models, leading to incorrect classification of images. Moreover, a single adversarial attack exhibits high transferability across different models, posing a significant threat to the security of medical deep learning systems. We reveal the vulnerability of the medical deep learning systems against white-box attacks.

2. To address the security vulnerabilities in the deep learning system for medical images, we built a defense deep learning system for medical images with better defense performance. In order to defend against white-box adversarial attacks, we propose a more generalized defense method. The method not only defends against a single form of attack, but also effectively resists multiple different attack methods. The effectiveness of this method has been verified through rigorous testing, effectively improving the security and reliability of the medical image deep learning system.

3. We propose a decision-based black-box attack method that enables efficient attacks with a limited number of queries. This technique reduces the number of queries required to perform an effective attack. Through experimental validation on a variety of deep learning models of medical images, we confirm the effectiveness and applicability of this black-box attack method. The experimental results show that the method can achieve fast and efficient attacks on a variety of models, validating its potential and importance in practical applications. Our proposed black-box attack method demonstrates good generalization ability to different deep learning models of medical images. This finding is important for understanding the vulnerability of existing medical image deep learning systems in terms of security protection and also guides designing more secure deep learning systems for medical images.

4. We reveal the vulnerability of medical image deep learning systems in the face of black-box attacks, and based on these findings, we propose an effective defense strategy based on medical image deep learning systems. This strategies not only improve the model's resistance to black-box attacks but also guide designing safer and more reliable deep learning systems for medical images.

1.3 The structure of this thesis

The chapter structure of this paper is organized as follows:

Chapter 1: Introduction

Introduces the background and importance of the research on deep learning systems based on medical images, and describes the motivation, purpose, and contribution of the research.

Chapter 2: Background
It mainly introduces the basic concepts of deep learning and medical images, including the development history of deep learning, model construction techniques, and classification methods for medical images.

Chapter 3: White Box Adversarial Attacks on Medical Image Deep Learning Systems
The security of the deep learning system based on medical images is tested using the white-box adversarial attack algorithm, revealing potential security vulnerabilities of the system.

Chapter 4: Defense against White Box Adversarial Attacks
Proposes a defense method for medical image-based deep learning systems against white-box attacks.

Chapter 5: Black Box Adversarial Attacks on Medical Image Deep Learning Systems
Propose black-box attack algorithms that are more suitable for the real world and use the algorithms to test the security of a deep learning system based on medical images.

Chapter 6: Defense against Black Box Adversarial Attack
Explore the defense strategies against black-box attacks and propose a defense method against black-box attacks for deep learning systems based on medical images.

Chapter 7: Conclusion and Future work
Summarizes the full paper and research results and discusses future research directions, including potential improvement points and research directions, etc.

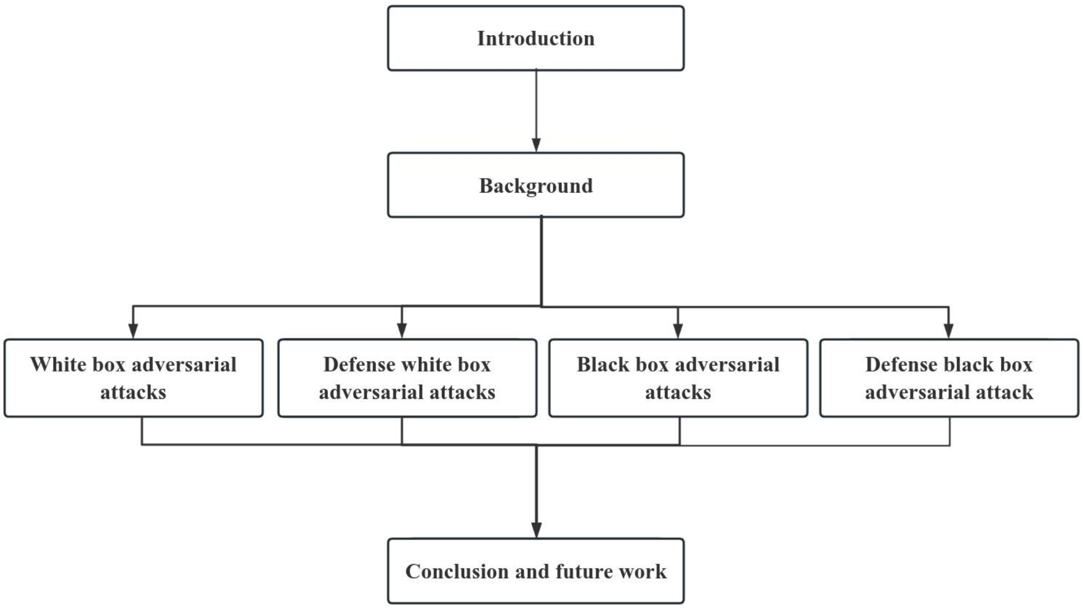

Introduction
Background
White box adversarial attacks
Defense white box adversarial attacks
Black box adversarial attacks
Defense black box adversarial attack
Conclusion and future work

Figure1-1 The structure of thesis

2 Background

This chapter introduces the basics of deep neural networks, adversarial samples, and medical images.

2.1 Artificial Neural Network, ANN

Artificial Neural Network (ANN), humans discovered that neurons in the brain collaborate to complete the processing and transmission of information, by imitating the structure and function of biological neural networks and designing mathematical models for information processing [9]. It consists of a large number of artificial neurons, which are connected through connection weights to form a network. Artificial neural networks are commonly used for deep learning tasks.

2.1.1 Multilayer Perceptron, MLP

American psychologist Frank Rosenblatt proposed a neural network with a single layer of computational units called the Perceptron in 1958 [10]. The Perceptron transmits and processes information by simulating human vision to receive information from the environment and utilizing connections between neurons for information transfer (Figure 2-1). Single-layer perceptron is the most basic binary classification neural network, but cannot effectively classify nonlinear data.

$$\mathrm{m}=\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{w_i x_i}+\mathrm{b}$$

$$y = sign(m) = \begin{cases} +1, m > 0 \\ -1, m \le 0 \end{cases}$$

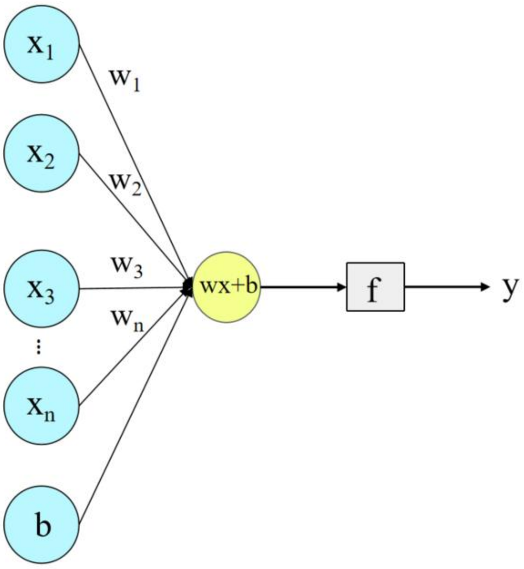


Figure 2-1 The Structure of the perceptron

Multilayer Perceptron is a structure based on feed-forward neural networks, consisting of an input layer, a hidden layer, and an output layer, where the hidden layer can have multiple layers [11]. The neurons in each layer are connected to the neurons in the neighboring layers, and the learning and prediction of complex problems are realized by constantly adjusting the weights between neurons, which has a powerful processing and expression ability and is widely used in a variety of tasks such as classification, regression, and recognition (Figure 2-2).

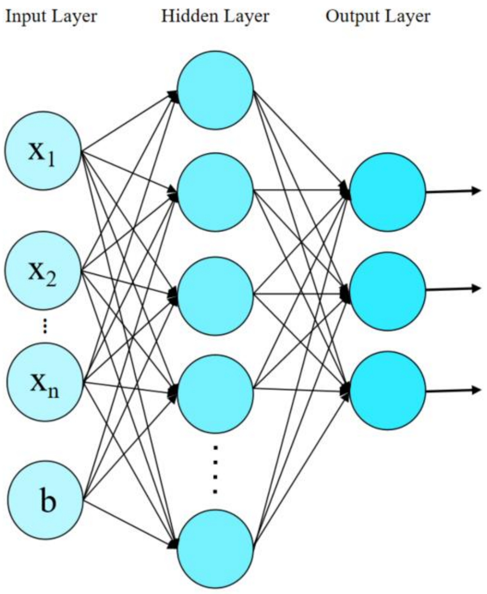


Figure 2-2 The Structure of multilayer perceptron

2.1.2 Convolutional Neural Network, CNN

Convolutional Neural Networks (CNN) is a class of feed-forward neural networks containing convolutional computation with deep structure, which is one of the representative algorithms for deep learning [12]. It is mainly used to process and analyze data with a grid structure, such as images and videos. CNN is involved in the development of computer vision, pattern recognition, and artificial intelligence.

The basic architecture of CNN mainly contains structures such as the input layer, convolutional layer, pooling layer, fully connected layer, and output layer [13]. Based on the basic network architecture of CNN, the network structure can be freely set up with multiple convolutional layers and pooling layers connected, which can fully learn various features of the input information. The input data is processed by the input layer and fed into the CCN model, firstly through the convolutional layer for feature extraction then through the pooling layer for feature selection, and then in the fully connected layer for feature integration, and finally the corresponding labels are predicted by the output layer, and the results are output.

（1） Convolutional layer

The convolutional Layer is one of the core components of CNN, which extracts features from the input data through convolutional operations [14]. The convolution operation is the core of the convolutional layer and is performed on the input data through a convolution kernel. This convolution kernel moves over the input data through a sliding window and performs dot multiplication and summation operations at each position. This helps in detecting different features like edges, texture, etc. in the input data. CNNs learn the weights of the convolutional kernels during training, thus enabling the network to automatically learn the most efficient feature extraction for a given task. The size, number, and weights of the convolutional kernels are hyperparameters that are determined based on the network structure and task requirements.

（2） Pooling Layer

Pooling Layer is often used in convolutional neural networks along with convolutional layers, mainly to bracket reduce the data dimensions, reduce the amount of computation, and extract features [15]. The main purpose of the pooling layer is to retain the most important information and reduce computation by downsampling the input data. Common pooling operations are Max Pooling and Average Pooling. In CNNs, a pooling layer is usually added after the convolutional layer to gradually reduce the size of the feature map. By stacking multiple convolutional and pooling layers, the network can extract more abstract features layer by layer. This helps CNNs achieve excellent performance in tasks such as image recognition and classification.

（3）Fully Connected Layer

Fully Connected Layer, also called Dense Connected Layer, in Fully Connected Layer, each neuron is connected to all the neurons in the previous layer, and each connection has a weight, which will lead to a relative increase in the number of parameters in Fully Connected Layer, especially when the input data increases, the number of parameters in Fully Connected Layer will increase rapidly [16]. The fully connected layer maps the features of the previous layer to the final output space so that the network can integrate global information about the entire input data, not just local features. The network's ability to perceive the overall structure and context is improved, thus enabling the network to perform specific tasks such as image classification, target detection, etc.

（4） Activation Function

An activation function is a nonlinear transformation in a neural network that introduces nonlinearity to the network, allowing it to learn complex mapping relationships. In convolutional neural networks, activation functions are typically applied to the outputs of the convolutional and fully connected layers [16]. In the following, we will introduce several common activation functions.

The Sigmoid activation function maps the input to a continuous output of [0, 1], thus making the raw output of the neural network map a range representing probability [17]. Therefore, it is well suited for binary classification tasks. The specific mathematical expression is as follows:

$$f(x)=\frac{1}{1+e^{-x}}$$

In a convolutional neural network, we compute the gradient by using the chain rule several times, which can potentially lead to the gradient becoming progressively very small, leading to the problem of vanishing gradients (Figure 2-3).

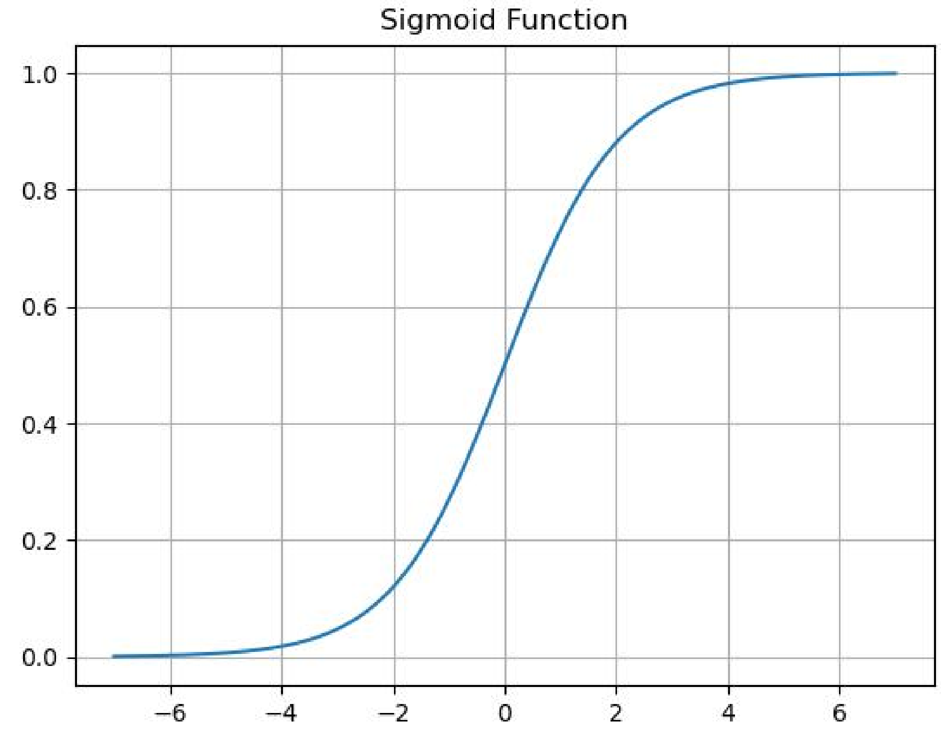


Figure 2-3 The sigmoid function

The ReLU activation function helps with this problem by setting the negative number to 0 and the derivative of ReLU is set to 1 in the positive part so that the gradient does not vanish in the positive interval [18] (Figure 2-4).

The Leaky ReLU activation function is a variant of ReLU. In Leaky ReLU, negative inputs are not suppressed completely but are multiplied by a small positive slope, usually a very small constant [19]. This helps to solve the problem of neuron "death" that can occur in ReLUs with negative inputs, i.e., the neurons may stop learning during training because they always output 0 on negative inputs (Figure 2-5).

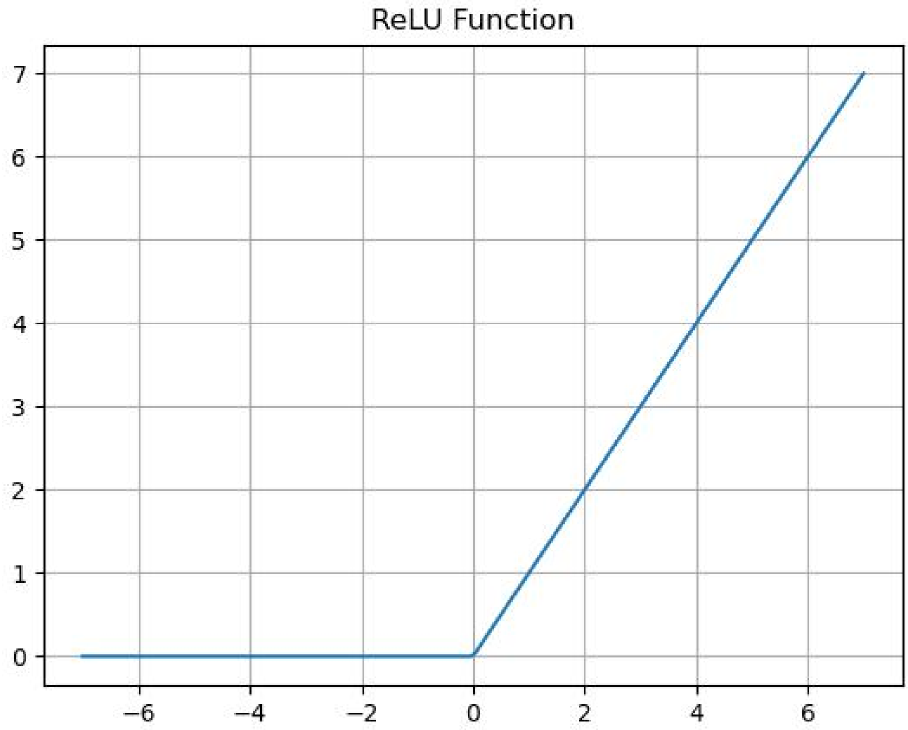


Figure 2-4 The ReLU function

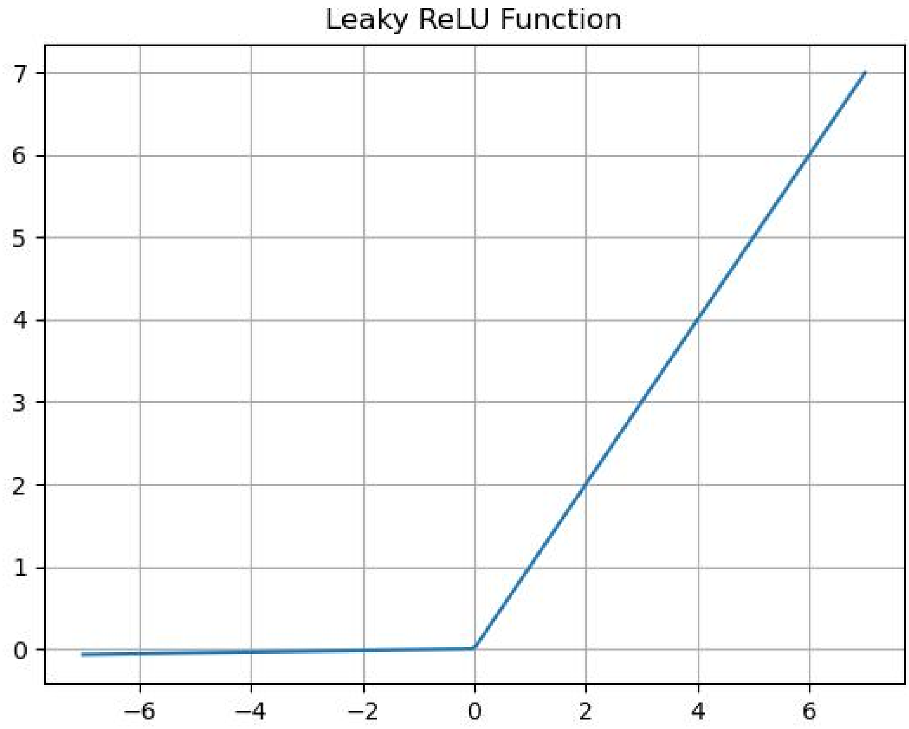


Figure 2-5 The Leaky ReLU function

The Tanh activation function is an S-shaped function, similar to the Sigmoid activation function, but the output of the Tanh function is in the range of (-1, 1), which is close to zero means compared to the Sigmoid function [20] (Figure 2-6). This helps to alleviate the problem of mean shift in the data, which can lead to the problem of vanishing gradient or gradient explosion when the mean value of the input deviates from zero.

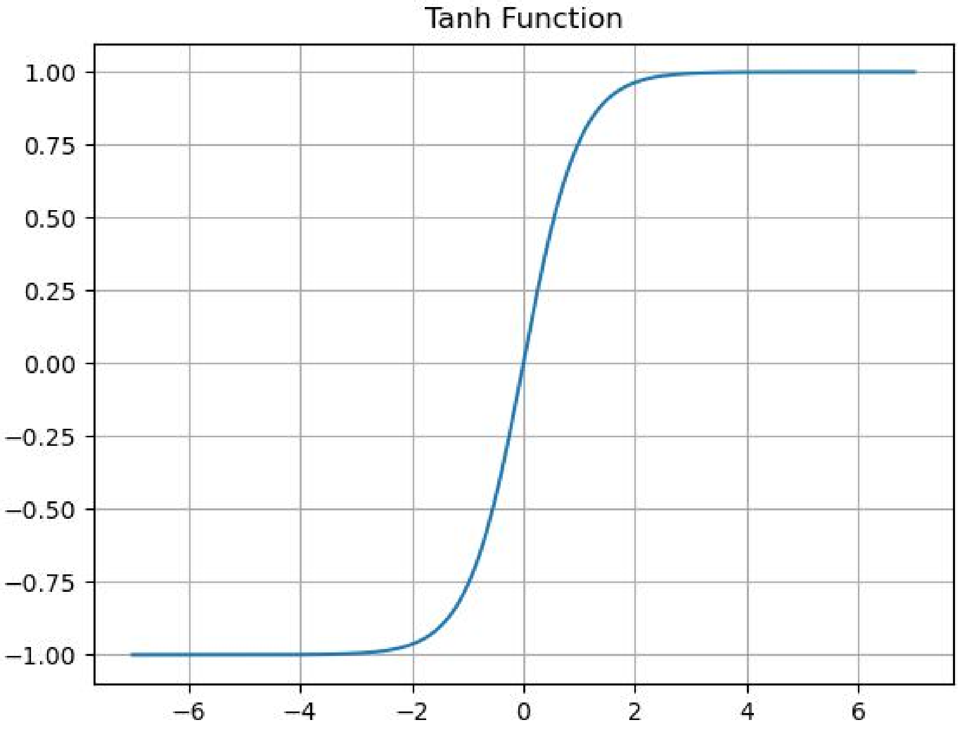


Figure 2-6 The Tanh function

Softmax is commonly used in the output layer of neural networks, especially for multi-categorization problems [21]. The Softmax function can convert the raw network output into a probability distribution such that the output of each category is represented as the probability of that category and the output probability of each category is between 0 and 1, while the sum of the probabilities of all the categories is 1. The activation function of Softmax makes the neural network able to output probability estimates for multiple categories, thus enabling the classification of input samples. Mathematical expression for the Softmax function:

Given an original output vector $S = (S_1, S_2, S_3, \ldots, S_m)$ with M categories, the Softmax function transforms them into a probability distribution $P = (P_1, P_2, P_3, \ldots,$

P_m), where *e* is the base of the natural logarithm and the denominator is the exponential sum of the outputs of all categories:

$$p_i = \frac{e^{s_j}}{\sum_{j=1}^{M} e^{s_j}}, i = 1, 2, 3, ..., m$$

2.2 Adversarial sample

An adversarial sample is generated by applying subtle perturbations (that are difficult to detect by the naked eye but are acceptable to the deep learning model) to the original data, leading to the input data being misjudged by the deep learning model [22]. The input data are denoted by x, the deep learning model is denoted by g, the classification result is denoted by g(x), and the perturbation is denoted by ϵ. Suppose there is a slight perturbation ϵ:

$$|\, \epsilon \,| < \delta \textit{ and } g\ (x + \epsilon)\ != g(x)$$

Then, $x + \epsilon$ can be called an adversarial sample.

2.3 Medical images

The development of medical imaging can be traced back to the end of the 19th century and the beginning of the 20th century, when X-ray technology was discovered and widely used in medical diagnosis [23]. With the advancement of science and technology and the development of computer technology, medical imaging has made tremendous progress. Medical images are based on physical principles that produce interactions with organisms to extract information about the morphology, structure, and certain physiological functions of tissues or organs in the organism, and to provide imaging information for biological tissue research and clinical diagnosis. Clinical medical imaging predominantly includes modalities such as X-ray, Computed Tomography (CT), and Magnetic Resonance imaging (MRI) scans, along with Ultrasound (US), retinal photography, and dermoscopy images, each offering unique insights into the internal workings of the human body [24]. Medical imaging plays a crucial role in modern medicine, not only helping doctors diagnose diseases more accurately but also monitoring treatment effects and guiding treatment processes such as surgery [25]. With the development of technology, new imaging techniques and methods are constantly emerging, improving the resolution, speed, and security of medical imaging and providing patients with a better diagnosis and treatment experience [26].

X-ray imaging (X-ray): This is one of the most commonly used medical imaging techniques and is primarily used to screen for broken bones, chest conditions (such as pneumonia), and certain cancers [27].

Computed Tomography (CT): CT scans provide a more detailed image of the body than X-rays by combining multiple X-ray images to produce a cross-sectional image of the inside of the body, and are commonly used to detect various types of cancer, brain damage, internal bleeding, and more [28].

Magnetic Resonance Imaging (MRI, Magnetic Resonance Imaging): MRI uses a strong magnetic field and radio waves to produce detailed images of the inside of the body, and is particularly useful for soft tissues such as the brain, spinal cord, muscles, joints, and internal organs [29].

Positron Emission Tomography (PET): PET scanning is a nuclear medicine imaging technique used to study metabolic activity and blood flow by detecting the distribution of radiopharmaceuticals in the body, and is often used for studies of cancer and brain function [22].

Ultrasound: Ultrasound uses high-frequency sound waves to produce images of internal body structures and is often used to examine pregnant women, internal organs, blood vessels, and soft tissues [30].

Digital Subtraction Angiography (DSA): DSA is an imaging technique used to show the inside of blood vessels in detail. It enhances the image of the blood vessels by injecting a contrasting agent and is often used to check for blood vessel abnormalities such as stenosis, blockages, or aneurysms [31].

Deep learning has been a tremendous success in recent years and is increasingly being used in a variety of domains[32], including autonomous driving, speech recognition, and drug discovery [33]. With the development of technologies such as artificial intelligence and machine learning, medical images have entered a new era. Traditional medical image analysis relies on the expertise of radiologists and other professionals to manually identify abnormal areas in images. This process is not only time-consuming but also susceptible to factors such as personal experience and fatigue, which can lead to misdiagnosis or missed diagnoses. As the volume of medical data increases and manual analysis becomes increasingly impractical, automated solutions are needed to improve the accuracy and efficiency of analysis. Deep learning has powerful data processing capabilities and therefore plays an increasingly important role in medical image analysis. These technologies can help doctors automatically analyze and identify lesions in images, improving the accuracy and efficiency of diagnosis.

The main applications of deep learning to medical impact are currently as follows：

2.3.1 Medical image classification

We input medical images into a deep learning model to get diagnostic results, such as whether the results are normal or abnormal. It is worth noting that because the data of medical images are often much smaller than the number required by computer vision models, it is difficult to achieve good results with the data of medical images alone. Therefore, the transfer learning approach is often used to solve the problem of insufficient medical image data by training the generated network with other images [34].

2.3.2 Medical image segmentation

Medical image segmentation is the accurate segmentation of organs or other parts of the body, which facilitates the quantitative analysis of clinical metrics such as the volume and shape of the target objects on the image [35]. Medical image segmentation is the process of dividing a digital medical image into multiple parts to isolate specific regions or anatomical structures. This involves delineating the boundaries of organs, tumors, or other relevant anatomical features from surrounding tissues, using medical image types such as MRI, CT, and ultrasound. Deep learning enhances the reliability of medical diagnosis by clearly delineating different tissue types, thereby reducing the likelihood of human error when diagnosing complex cases.

In addition, it simplifies the workflow in medical diagnosis by automating the measurement and analysis process.

2.3.3 Medical image detection

The detection of medical images includes the localization of organ boundaries and the detection of lesion locations. Lesion location detection is a critical step in clinical medical diagnosis and one of the most time-consuming tasks for physicians. Compared with the medical image classification problem, the lesion is relatively small, and lesion detection needs to classify each pixel, and the classified part is relatively small, so the number of non-target classes exceeds the target classes, resulting in a classification imbalance of training data. Therefore, how to solve the data imbalance is important for the lesion detection task [36] .

3 White Box Adversarial Attacks on Medical Image Deep Learning Systems

This chapter tests the security of medical image deep learning systems using white-box adversarial attacks.

3.1 Introduction

Towards the end of 2019, the coronavirus disease 2019 (COVID-19), caused by the SARS-CoV-2 virus emerged; infections were mainly transmitted through respiratory droplets spread rapidly, and eventually, were recognized as a global pandemic by the WHO [37,38]. The most common clinical symptoms of COVID-19 infection include a cough, fever, headache, and so on [39,40]. Particularly, in high-risk populations such as the elderly or those with numerous disorders where COVID-19 may induce lung damage, infection with COVID-19 is more likely to result in viral pneumonia. Severely infected patients may develop acute respiratory distress syndrome, severe lung infection, fibrosis, and even death [41]. The rapid global spread of COVID-19 has caused serious damage to human health, the world economy, and public health security [42–44]. At the beginning of the COVID-19 epidemic, its clinical diagnosis was based on a patient's epidemiology, clinical presentation, a chest X-ray, chest CT, and RT-PCR [45,46]. As compared with other diagnostic methods, chest CT, which is the main tool for screening and diagnosing COVID-19, can detect pulmonary lesions and can also classify patients into early, intermediate, or severe cases based on CT manifestations in the chest [47–49]. CT images of the lungs of patients with COVID-19 show patchy or ground glass shadows [50]. As the disease progresses, the severity of the lung lesions may become more significant, and pulmonary fibrosis may develop, with a white coloration of both lungs detected by CT lung examination [50,51]. Therefore, it is crucial to display information about the lungs of COVID-19 cases through CT. Doctors can correctly evaluate patients' CT images, thus, diagnosing patients' conditions for early detection and treatment.

Breast cancer is one of the most common malignant tumors in women, its incidence is the highest among female malignant tumors, and it is developing earlier and becoming more prevalent [52–54]. Clinical studies have shown that early detection and precise treatment of breast cancer can effectively reduce the risk of death in patients, thus increasing the success rate of breast cancer treatment [55,56]. Therefore, the accurate identification and diagnosis of pathological images in breast cancer clinics is crucial for patients. It can help doctors make accurate judgments and assessments of a patient's condition to provide precise treatment for the patient's condition. Radiomics technology provides great assistance in the adjuvant treatment and prediction of breast cancer [57–59]. Medical radiologists mainly use radionics to observe the characteristics of breast pathological tissue information for the quantitative analysis of breast cancer cells, lymphocytes, and glands, to effectively diagnose breast pathology images and assess disease [60].

Pneumonia is an acute exudative inflammatory condition affecting the pulmonary tissues, representing a frequently encountered pathology within the spectrum of respiratory system diseases [61]. Populations such as children and the elderly exhibit heightened vulnerability to infectious pneumonia, primarily due to the typically diminished robustness of their immune systems [62]. Consequently, the expeditious diagnosis and treatment of conditions such as pneumonia are imperative in these demographic groups [63]. The primary modalities for diagnosing pulmonary disorders include X-ray, Magnetic Resonance Imaging (MRI), and Computed Tomography (CT). Among these, X-rays and CT scans are often favored in clinical settings due to their relative cost-effectiveness and satisfactory image quality [64]. However, the task of detecting pneumonia from chest radiographs presents substantial challenges, necessitating a high level of professional knowledge and extensive clinical experience from medical practitioners. The daily requirement for physicians to examine a large volume of chest X-ray images can lead to visual fatigue, increasing the risks of diagnostic errors and oversights. In response to these challenges, the development of Computer-Aided Design (CAD) systems has emerged as a pivotal technological advancement [65]. These systems aim to augment the diagnostic process, offering crucial support in analyzing imaging data, thereby reducing the likelihood of misdiagnosis and enhancing the overall efficiency and accuracy of pneumonia detection [66]. This integration of AI technology into medical imaging represents a significant stride forward in the battle against one of the most common diseases in clinical medicine [67].

Deep learning has been a tremendous success in recent years and is increasingly being used in a variety of domains, including autonomous driving [68], speech recognition [69], and medical image analysis [70]. In the field of medical imaging, deep learning has been increasingly applied with significant impact, primarily in three key areas: medical image classification, lesion detection in medical images, and medical image segmentation [71]. The revolutionary impact of deep learning on medical image diagnosis is undeniable. However, the security of deep learning systems in this context has garnered significant attention from the research community. Szegedy et al. initially exposed a fundamental vulnerability in deep neural networks used for image classification tasks. Their findings revealed that introducing subtle perturbations into input samples, imperceptible to humans, could mislead the model into generating highly confident but erroneous outputs. This phenomenon has prompted a heightened focus among researchers on the security aspects of deep learning. The stakes are notably higher in the realm of medical imaging as compared to natural image processing. The security of deep learning systems that handle medical images is of paramount importance due to the sensitive nature of personal health data involved. Accurate identification and analysis of medical images by deep learning systems are crucial in assisting physicians with diagnoses. Misdiagnoses resulting from system errors could lead to serious, irreversible consequences. Moreover, there is a risk of malicious entities attempting to compromise medical deep learning systems, potentially manipulating diagnostic results for purposes such as insurance fraud.

Given these considerations, this paper investigates the security of deep learning systems based on three common medical images. Unlike natural images, medical images are a more prevalent medium in medical imaging, making them a critical focus of our security testing. We aim to verify the possibility of threat transferability in such systems, underscoring the need for robust security measures in medical deep learning applications to prevent potential misdiagnoses and safeguard against malicious manipulations.

3.2 Preliminary

In this section, we introduce the concept of the adversarial sample, the classification of attack methods, and methods of generating adversarial samples.

3.2.1 Adversarial sample

An adversarial sample is a uniquely crafted version of the original data that is manipulated through the application of subtle perturbations. These perturbations, while typically imperceptible to the human eye, are designed to be significant enough for a deep learning model to misinterpret the data. The original input data is represented as x, the deep learning model as g, and the output classification by the model as g(x). The perturbation applied to the data is denoted by ϵ. The criteria for an adversarial sample are based on the perturbation ϵ being sufficiently small yet effective. The magnitude of ϵ is constrained such that $|\epsilon| < \delta$, ensuring that the perturbation remains subtle. The perturbed input $x+\epsilon$ leads to a different classification result compared to the original input, i.e., *g(x+*ϵ*) != g(x)*.

3.2.2 The classification methods of adversarial sample

There are various classifications of attacks based on their attack environments; therefore, attacks can be classified as black-box, white-box, and gray-box attacks [72].

Black-box attacks mean that the attacker does not know the internal structure of the attacking model, the training parameters, or the defense methods, and can only interact with the model through the output.

White-box attacks are unlike black-box models, as the attacker knows everything about the model, including the network structure and parameters. Most of the current attack algorithms are white-box attacks.

Gray-box attacks are found between black-box and white-box attacks, and only a part of the model is known (e.g., realizing the output probability of the model or understanding the model structure but not the parameters).

Concerning the purpose of the attack, attacks can be divided into targeted and untargeted attacks [73].

An untargeted attack is associated with image classification, namely in the sense that the attacker only needs to make the target model misclassify the sample but does not specify which classification is wrong.

A targeted attack means that the attacker specifies a class so that the target model not only misclassifies the samples but also misclassifies them into the specified type. In terms of difficulty, targeted attacks are more challenging to implement than untargeted attacks.

3.2.3 Generating Adversarial Samples

There are several adversarial attack methods proposed in the literature, but we only discuss the ones that are most relevant in this section.

(1) Gradient-Based Generation of Adversarial Samples

The gradient is obtained from the input data in the training phase, then the input data are updated stepwise according to the loss function, and finally, the adversarial sample is obtained.

Fast Gradient Sign Method, FGSM Goodfellow et al. demonstrated a simple and fast way to generate adversarial samples to spoof deep learning models by designing the FGSM adversarial attack algorithm and explaining why deep neural networks are so vulnerable to adversarial samples [74]. The FGSM algorithm utilizes the model's gradient information by adding a small perturbation to the input data along the direction of the gradient, causing the model to produce false outputs. The fast gradient sign method (FGSM) is a gradient-based method for generating adversarial samples that maximize the loss function in the opposite direction of the decreasing gradient during the data propagation and updating of a neural network. The expression of FGSM is shown below.

$$x_{adv} = x + \epsilon \cdot sign\ (\nabla x\ J\ (\theta, x, y))$$

where x is the input sample, y is the label corresponding to sample x, x_{adv} is the adversarial sample, θ is the weight parameter of the model, the manually set perturbation parameter of the model is ϵ, and the loss function of the model is *J()*. The FGSM algorithm is shown below (Table 1).

Algorithm 1 FGSM

```
Input: original image, orig_im; original_target, orig_tar;
Output: adversarial image, adv_im; adversarial target, adv_tar;
adv_im = orig_im
iteration = 1
while iteration < max_iteration and adv_tar = orig_tar
adv_im = orig_im + iteration * step_size * sign(gradient(orig_im))
adv_im = clip(adv_im, min, max)
iteration + = 1
end
return adv_im
```

The simplicity and efficiency of the FGSM method have attracted much attention, and it has become one of the classic methods in the study of adversarial attacks. The proposal of FGSM triggered attention to the robustness of deep learning models and drove the rapid development of the field of adversarial machine learning, which became one of the foundations for evaluating model robustness, designing defense mechanisms, and gaining insights into model behavior.

Basic Iterative Method, BIM, proposed by Kurakin et al. in 2016, is an improvement of the FGSM algorithm that makes the generated adversarial samples more robust by introducing the idea of iteration [75]. FGSM performs only one gradient update, which may not be sufficient against more complex defense mechanisms, and BIM builds adversarial samples gradually by iterating over several iterations, introducing small perturbations each time. This iterative process increases the attacker's influence on the target model, making it more difficult for the model to resist attacks. However, compared to FGSM, BIM requires multiple iterations, resulting in higher computational complexity and higher computational cost.

Projected Gradient Descent PGD is an iterative attack method. It generates adversarial samples by computing the gradient in each iteration and updating the data in the opposite direction of the gradient at each step [76]. With multiple iterations, PGD can search in a larger perturbation space and find more challenging adversarial samples more easily. Due to its iterative nature, PGD is typically better at generating more challenging adversarial samples than FGSM. Multiple iterations allow PGD to search a wider perturbation space, thus generating adversarial samples that are more challenging for the model and improving the robustness of the model.

Momentum Iterative Fast Gradient Sign Method, MI-FGSM is based on FGSM, and the attack can be improved by adjusting the iteration step size and number [77]. For example, the accuracy and efficiency of the attack can be improved by using an adaptive step size or a dynamic adjustment strategy based on the response of the target model.

$$\mathrm{g}_{\mathrm{t+1}} = \mu \cdot g_t + \frac{\nabla_x J\left(x_t^{adv}, y\right)}{\| \nabla_x J\left(x_t^{adv}, y\right) \|_1}$$

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign(g_{t+1})$$

(2) Optimization-Based Generation of Adversarial Samples

In the training phase of the model, the value of the loss function is continuously reduced by calculating the loss function between the predicted and true values of the sample data, adjusting various parameters of the model in the backward transfer process, and iteratively calculating the parameters of each layer of the model to generate adversarial samples.

C&W Carlinr et al. [70] proposed a set of adversarial C&W attacks based on optimization, considering both a high attack rejection rate and low adversarial disturbance.

$$r_n = \frac{1}{2}(\tanh(w_n) + 1 - x^n$$

$$\min_{w_X} \| r_n \| + cf(\frac{1}{2}(\tanh(w_n) + 1))$$

$$\text{where f(x')=max(max}\{\mathrm{Z(x')}i_X : i \neq t\} - Z(\mathrm{x}')t, -k)$$

(3) Adversarial Network-Based Generation of Adversarial Samples

In 2014, Goodfellow proposed exciting adversarial attack networks (GANs) [78], and then various studies on GANs have also emerged. GANs consist of two parts: a generator and a discriminator. A generator *(G)* is used to generate realistic samples from random noise, and a discriminator *(D)* is trained to discriminate the real data from the generated data, both are trained at the same time until a balance is reached, in which the data generated by the generator is indistinguishable from the real data, and the discriminator cannot distinguish the generated data from the real data correctly. Similarly, GAN-based networks can generate adversarial samples more efficiently. AdvGAN is a method for generating adversarial samples based on GANs models; given the input x, the perturbation *G(x)* is generated by the generator network [79]. On the one hand, *G(x) + x* is sent to the discriminator network for training, and on the other hand, *G(x) + x* is sent to the attacked network. The objective loss function is continuously optimized, and *G(x)* is the perturbation when the model reaches optimality. The target loss can be decomposed into three parts, expressed as:

$$L = L_{adv}^{\mathrm{f}} + \alpha L_{adv} + \beta L_{hinge}$$

where L_{adv}^{f} is the misleading misclassification loss, L_{adv} is the loss function of the GAN, and L_{hinge} is used to restrict the perturbations to a certain range.

We summarized the main adversarial attack algorithms in Table 3-1.

Table 3-1 The classification of adversarial attack algorithms

Method	Attack	Attack Box	Attack Target
Optimization-based method	JMSA [80]	White	Target
	L-BFGS [81]	White	Target
	C&W	White	Target
Gradient-based method	FGSM	White	Target
	BIM	White	No target
	PGD	White	No target
	MI-FGSM	White	No target
Adversarial network-based method	AdvGAN	White–black	No target
	AdvGAN++ [82]	White–black	No target
	AdvFaces [83]	White–black	No target

3.3 Methodology

The experimental methodology detailed in this paper encompasses three integral components: building three deep learning systems based on medical images, and the testing of the transferability of adversarial samples derived from medical images. To elucidate the security vulnerabilities, our approach begins with the construction of three medical deep learning systems that are capable of accurately recognizing and classifying normal and abnormal medical images respectively. Following this, the systems were subjected to adversarial attacks employing algorithms Fast Gradient Sign Method (FGSM) and Momentum Iterative Fast Gradient Sign Method (MI-FGSM). The final phase involves testing the transferability of these adversarial samples generated from chest X-ray images, thereby providing a comprehensive analysis of the deep learning system's resilience and susceptibility to such attack.

3.3.1 Building the medical image deep learning systems

Building the medical deep learning system involved training and testing a deep neural network (Figure 3-1). The training and testing stages meant that datasets had to be selected. The deep neural network needed to be carefully selected. In this section, we address these issues.

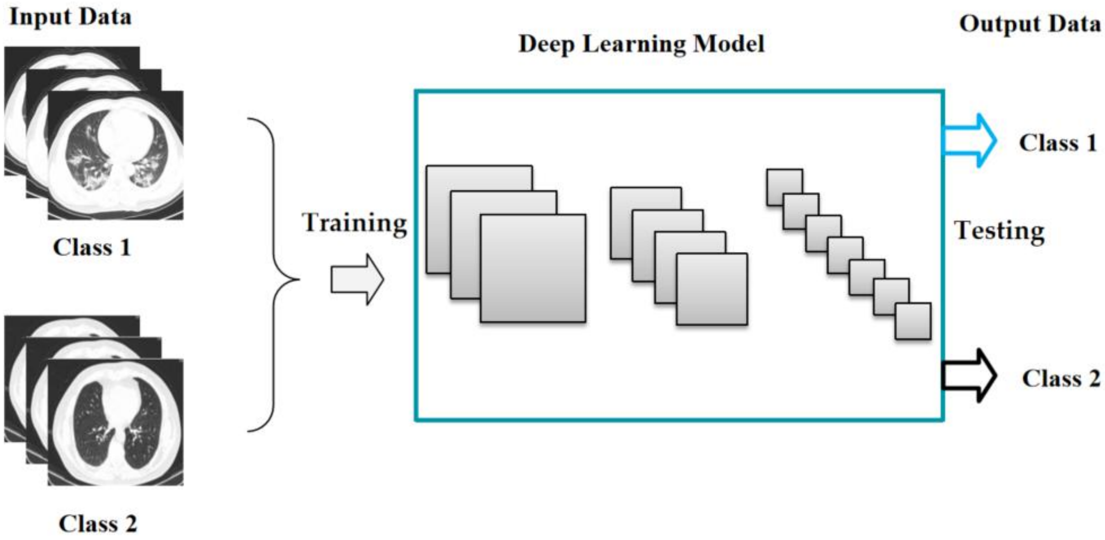


Figure 3-1 The pipeline of the medical images deep learning systems

(1) Datasets

The CT image data in this paper were obtained from publicly available datasets extracted from the medRxiv and bioRxiv preprints of COVID-19 by Xingyi Yang at the University of California, San Diego [84]. These datasets are anonymous and can be applied to the study of COVID-19. The datasets contained 349 CT images of COVID-19 infection cases (COVID-19 CT images) and 397 CT images of cases without

COVID-19 infection (non-COVID-19 CT images) (Figure 3-2).

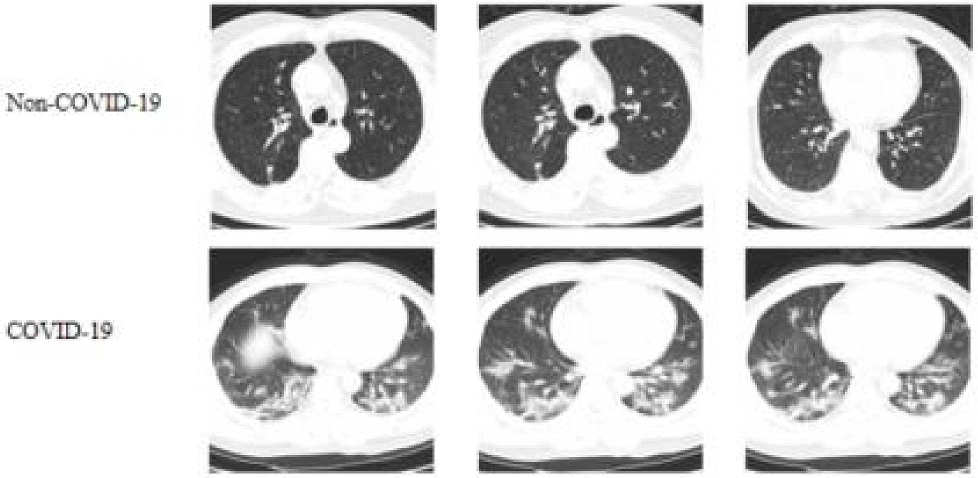


Figure 3-2 The Chest CT image

The breast cancer pathology image data in this paper were obtained from the breast cancer pathological database (BreakHis) [85]. This dataset is anonymous and publicly available for non-commercial studies on breast cancer images. It contains 644 benign breast tumor pathology images and 903 malignant breast tumor pathology (breast cancer) images. We divided these three different medical image datasets into three datasets: training set, validation set, and testing set (Figure 3-3).

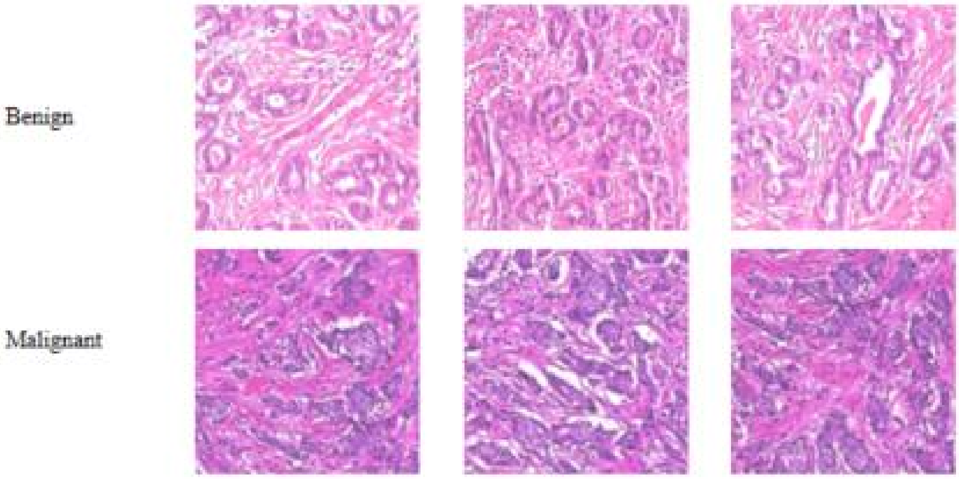


Figure 3-3 The breast tumor image

The chest X-ray image data utilized in this study were sourced from publicly accessible datasets. These datasets are anonymized, ensuring their suitability for research in the field of pneumonia [86]. The collection encompasses 4220 chest X-ray images indicative of pneumonia and 1580 chest X-ray images classified as normal (Figure 3-4).

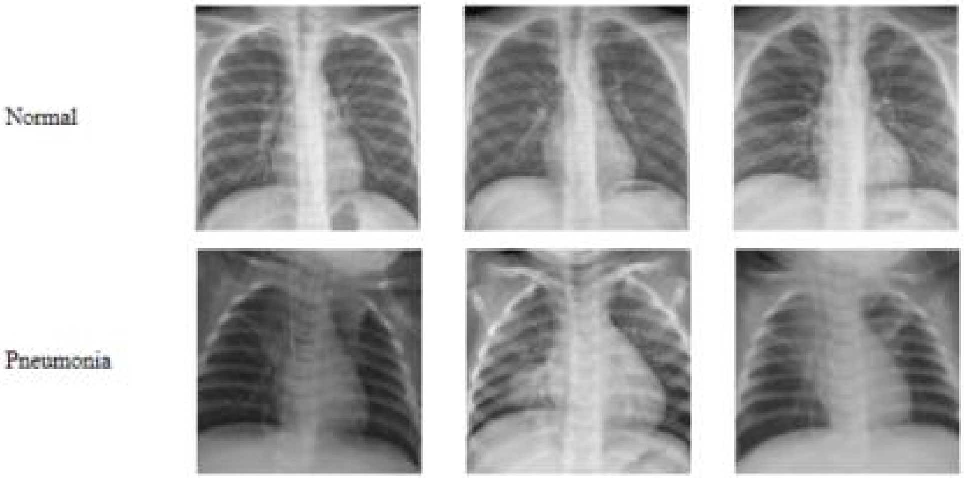


Figure 3-4 The chest X-ray image

The entire dataset has been systematically segmented into three distinct sets: the training set, the validation set, and the testing set (Table 3-2)

Table 3-2 The classification of datasets

Datasets	Medical images	Training set	Validation set	Testing set	Total
CT images	Non-COVID-19	317	40	40	397
	COVID-19	279	35	35	349
Breast cancer images	Benign	515	64	65	644
	Malignant	722	90	91	903
X-ray images	Normal	1264	158	158	1580
	Pneumonia	3376	422	422	4220

(2) Deep learning model

As compared with machine learning, the advantage of deep learning is that the network capacity is large enough to accommodate richer feature information, and the deep learning effect always improves as the number of data increases and deepens. Deep learning is a complex machine learning algorithm, and with continuous research, many classical deep learning models have emerged, which have greatly improved the performance of deep learning. In our research, we opted for two classical convolutional neural network models: ResNet [87] and DensNet [88]. The performance of deep learning models is significantly influenced by the dataset size, as larger datasets typically yield more effective training outcomes. Transfer learning is an effective method of applying knowledge gained in one domain to another, and is especially useful when the amount of data is limited.

We chose the classical ResNet model, the winning model of ImageNet 2015, which offers several advantages such as a very low error rate; it also presents little complexity and only requires small computational effort. One of the factors for better performance of deep learning is the dataset; a large dataset can make the model achieve better training results. Given the relatively small number of medical images in the public dataset, training a deep learning model from scratch would likely result in suboptimal performance. Consequently, we leveraged migration learning to enhance the training process for our medical images deep learning models by employing a pre-trained ResNet-50 model, wherein we froze the model parameters and replaced the pooling and fully connected layers. The optimization was carried out using the adaptive moment estimation (Adam) algorithm, with fine-tuning implemented via stochastic gradient descent at a learning rate of 1×10^{-3} [89]. The fully connected layer was also modified to differentiate between the two classes. Similarly, for the DenseNet-based system, we utilized a pre-trained DenseNet-121 model. This choice was aimed at achieving superior results, and the model was trained using the Adam optimizer as well, with a batch size of 32 and an initial learning rate of 1×10^{-3}. Both models underwent data augmentation to enhance the dataset. These methodological choices were instrumental in optimizing the performance of our medical image deep learning systems.

(3) Metrics

In this study, the performance of the breast cancer deep learning system was evaluated using the metric of accuracy. The accuracy metric was used to measure the overall correctness of the model's classifications [90]. True positives (TPs) indicated the number of COVID-19 images that were correctly classified as COVID-19 CT images. False positives (FPs) indicated the number of non-COVID-19 images that were incorrectly classified as COVID-19 images. True negatives (TNs) denoted the number of non-COVID-19 images that were correctly classified as non-COVID-19 CT images. False negatives (FNs) indicated the number of COVID-19 images that were incorrectly classified as non-COVID-19 CT images.

Similarly, True positives (TPs) represented the number of breast cancer images that were correctly identified as breast cancer images. False positives (FPs) indicated the number of benign tumor images that were incorrectly classified as breast cancer images. True negatives (TNs) represented the number of benign tumor images that were accurately identified as benign tumor images. Finally, false negatives (FNs) indicated the number of breast cancer images that were mistakenly classified as benign tumor images.

For the chest X-ray medical deep learning system, True positives (TPs) were defined as the number of Pneumonia images accurately identified as Pneumonia CT images. False positives (FP) were the instances where normal images were erroneously classified as Pneumonia images. True negatives (TN) represented the count of normal images correctly recognized as normal CT images. False negatives (FN) were the occurrences where Pneumonia images were mistakenly classified as normal CT images.

$$Accuracy = \frac{TPs + TNs}{TPs + TNs + FPs + FNs}$$

3.3.2 Adversarial attack on medical deep learning system

To rigorously evaluate the reliability of the deep learning system based on medical images, this study introduces slight perturbations to the test set images. These alterations, while being imperceptible to the human eye, possess the potential to be misclassified by the model. Essentially, pre-trained models are attacked and influenced by adversarial attack algorithms, a strategic approach to evaluating model classification accuracy under intentionally manipulated conditions (Figure 3-5).

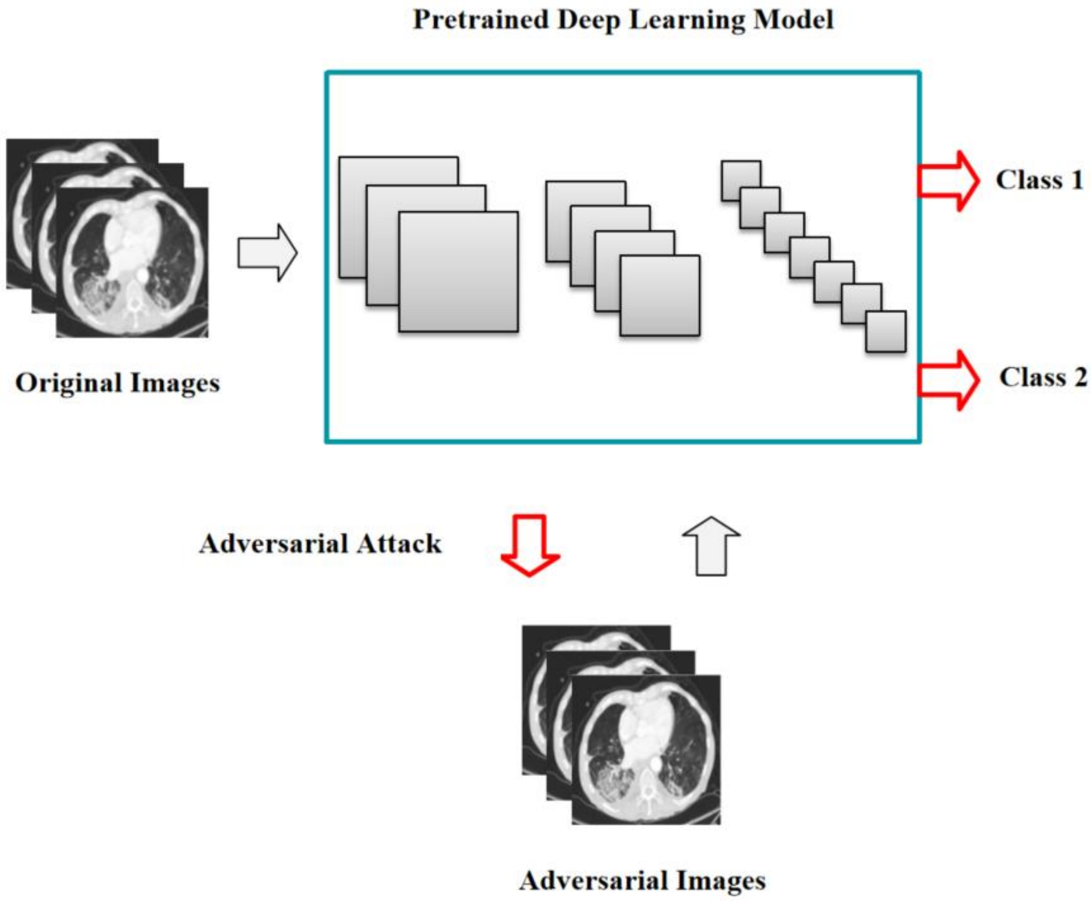


Figure 3-5 The pipeline of adversarial attack against medical deep learning system

3.3.3 Testing the transferability of adversarial samples in medical image deep learning systems

To evaluate the transferability and potential impact of adversarial samples generated from a chest X-ray image-based deep learning system, our methodology involves a two-step process using adversarial attack algorithms. Initially, we trained a ResNet-based deep learning system, specifically designed for medical image analysis. We attacked the model with an adversarial attack algorithm to generate adversarial samples, which were then used to test a standalone deep learning system based on the DenseNet architecture. The aim here is to assess the accuracy of the DenseNet-based model in the presence of these adversarial samples (Figure 3-6).

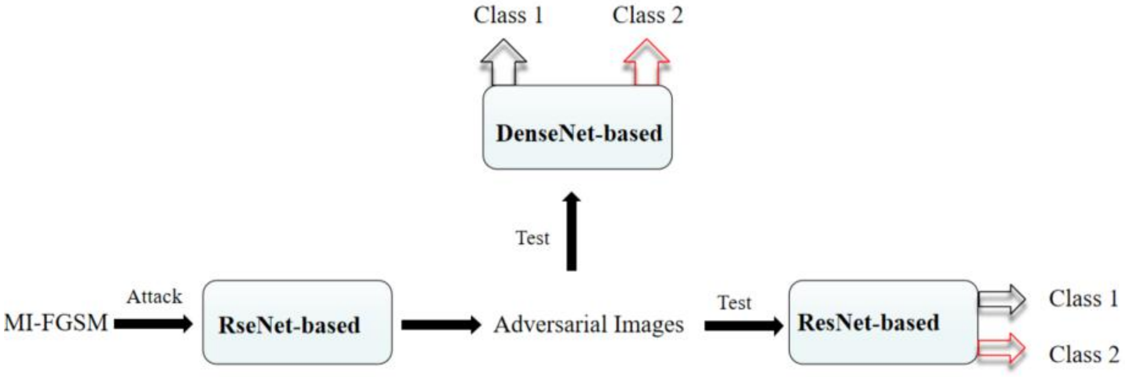


Figure 3-6 Adversarial attack on DeseNet-based medical image deep learning system

In a similar vein, the DenseNet-based medical image analysis deep learning system is again subjected to an adversarial attack, generating a new set of adversarial samples. These samples are subsequently employed to attack the same ResNet deep learning system, providing a measure of the system's resilience and accuracy when confronted with its adversarial samples (Figure 3-7).

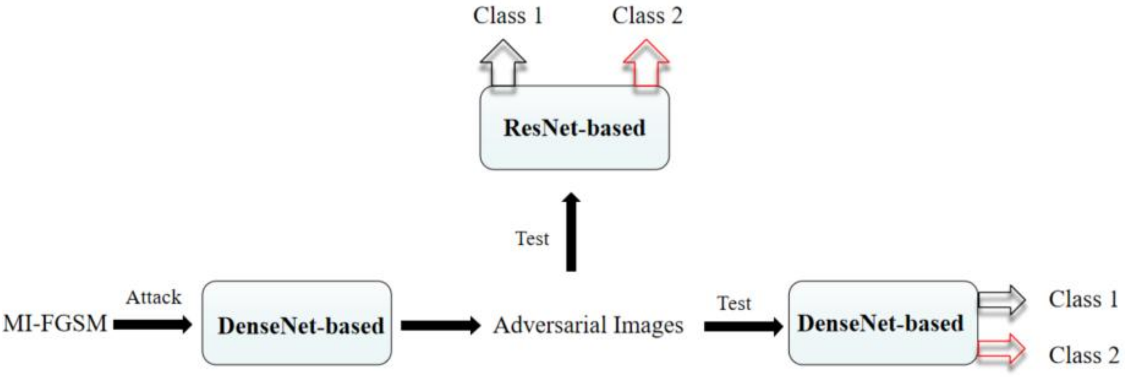


Figure 3-7 Adversarial attack on ResNet-based medical image deep learning system

This approach allows for a comprehensive assessment of the adversarial samples' transferability across different deep learning architectures. Furthermore, it elucidates the potential hazards these adversarial samples pose to medical image-based deep learning systems, underscoring the importance of robust model design to mitigate such risks.

3.4 Results and Discussions

After training, we tested the deep learning model based on ResNet and DenseNet for COVID-19 CT images and obtained an accuracy of 76.27% and 84%, respectively. This indicates that the model can accurately identify COVID-19 CT images and non-COVID-19 CT images and possesses good recognition accuracy. Likewise, the deep learning system based on ResNet and DenseNet models developed for analyzing breast pathology images has achieved accuracy rates of 95.51% and 98.72%, which amply demonstrates the ability of the model to accurately differentiate between benign and malignant breast tumors with a high degree of accuracy. Similarly, we evaluated the deep learning system based on chest X-ray images using the same test set. The empirical outcomes demonstrated a remarkable accuracy rate of 95.86% for the ResNet-based model and 96.03% accuracy for the DenseNet-based model. This demonstrates that our constructed deep learning system based on chest X-ray images is highly accurate in distinguishing between pneumonia and normal images (Table 3-3).

Table 3-3 The classification accuracy (%) of the medical image deep learning systems

Datasets	Accuracy	
	ResNet-based	DenseNet-based
CT images	76.27%	84.00%
Breast cancer images	95.51%	98.72%
X-ray images	95.86%	96.03%

To reveal the threat of adversarial attacks on medical deep learning systems and better simulate the security risk of real-world deep systems based on breast cancer images, we used three different types of medical images as the research object and attacked the pre-trained model with the FGSM and MI-FGSM adversarial attack algorithm and then tested the defense capabilities of the medical deep learning model against the adversarial attacks, and the results are shown in Table 3-4. When the deep learning model based on COVID-19 medical images is attacked by adversarial attack algorithms FGSM, the accuracy of the ResNet-based model decreases from 76.27% to 5.33%, and the accuracy of the DenseNet-based model decreases from 84.00% to 40%. Furthermore, when the same model is attacked with MI-FGSM, the accuracy of the ResNet-based model decreases from 76.27% to 0% and the accuracy of the DenseNet-based model decreases from 84.00% to 0%. This indicates that the model is successfully attacked by the adversarial attack algorithm and the attack can severely damage the performance of the model.

Similarly, in deep learning models for breast cancer image analysis, adversarial attack algorithms have a significant impact on the accuracy of the models, the accuracy of the ResNet model subjected to the FGSM attack plunged from 95.51% to 7.69%, while the accuracy of the DenseNet model under the same type of attack dropped from 98.72% to 33.97%. We tested the models with the MI-FGSM algorithm and found that the accuracy based on the ResNet-based model plunged from 95.51% to 0%, while the accuracy of the DenseNet-based model under similar attacks dropped from 98.72% to 0%. This shows that two different attack algorithms can attack the model, making it misclassify medical images.

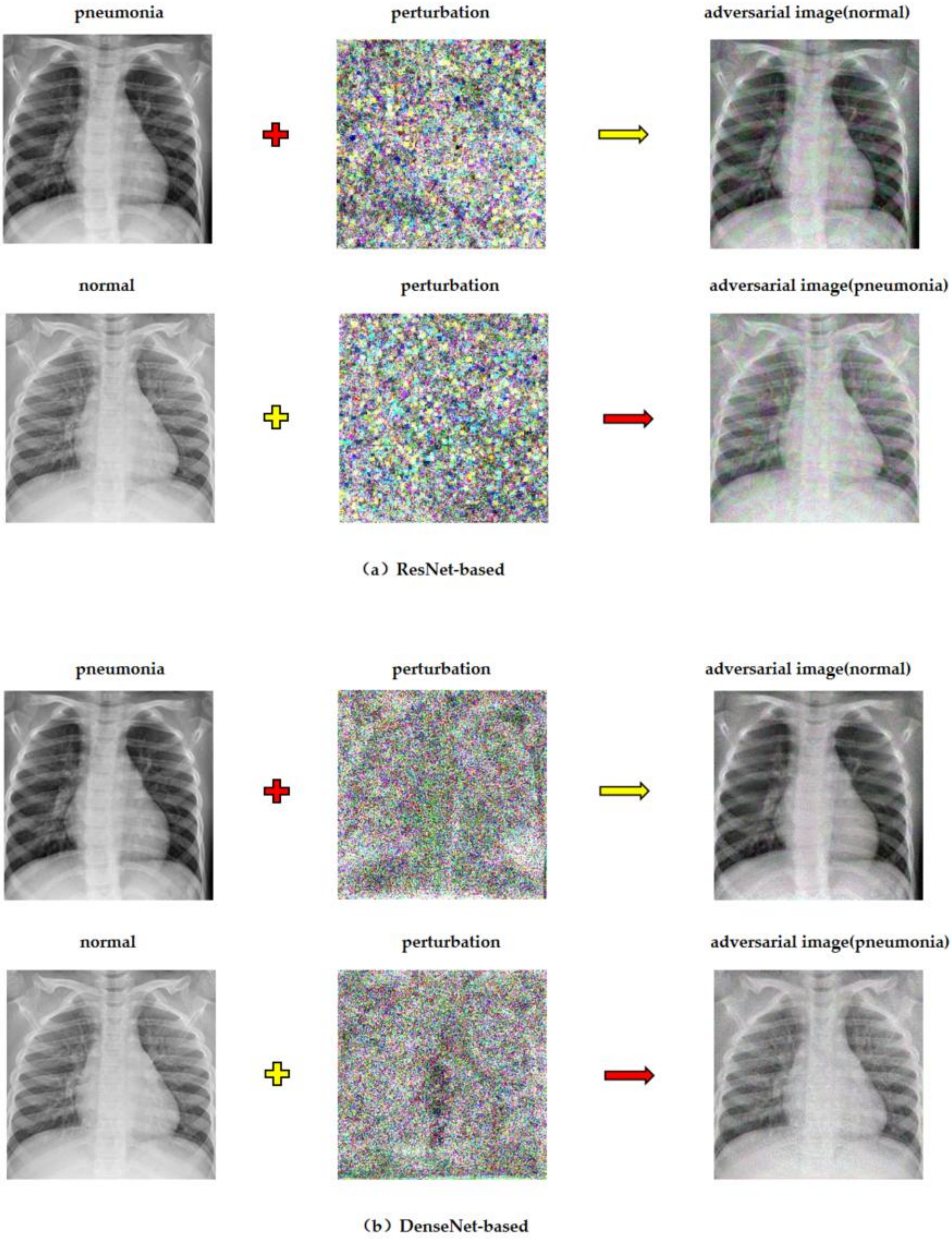
In order to study the impact of different color spaces grayscale and RGB, on deep learning models against sample attacks. We perform the attack by converting the grayscale image of chest X-ray to RGB format. For the deep learning system for medical images based on chest X-ray images, when ResNet-based and DenseNet-based models, are subjected to well-designed adversarial attacks FGSM, their performance drops dramatically. Specifically, the recognition accuracy of the ResNet-based model drops from the original 95.86% to 72.76% after being attacked, while the accuracy of the DenseNet-based model drops directly from 96.03% to 83.28%. When subjected to the MI-FGSM attack, the performance of the ResNet-based model experiences a substantial decline in accuracy, decreasing from 95.86% to 39.83%. Similarly, the DenseNet-based model demonstrates a pronounced vulnerability, with its accuracy plummeting from 96.03% to 0%. In addition, we found that when converting grayscale images to RGB format for attacking, you find that the attack efficiency of the adversarial attack algorithms undergoes a significant decrease relative to the other models. This finding can lead to a discussion on the adaptability of attack algorithms to different types of images and the impact on model robustness. This contributes to a deeper understanding of the mechanisms of the adversarial sample attack and has important implications for improving the robustness and security of the model.

Table 3-4 The accuracy of medical images deep learning systems subjected to adversarial attack

Datasets	Accuracy			
	ResNet-based		DenseNet-based	
	FGSM	MI-FGSM	FGSM	MI-FGSM
CT images	5.33%	0%	40%	0%
Breast cancer images	7.69%	0%	33.97%	0%
X-ray images	72.76%	39.83%	83.28%	0%

These results not only reveal the vulnerability of deep learning systems for medical images in the face of malicious attacks, but also highlight the urgency of enhancing the resilience of the models before clinical application. Future research needs to be dedicated to developing techniques that can effectively defend against such attacks to ensure the security and accuracy of medical image analysis.

To better illustrate the hidden characteristics of adversarial samples, we compare the adversarial samples generated based on the ResNet model and DenseNet model with the original images. Figure 3-8 shows the comparison between the adversary samples generated based on the MI-FGSM based attack and the original samples, where a is the adversary sample generated based on the ResNet model, and b is the adversary sample generated by the DenseNet model. Figure 3-8 shows that when we examine them with the naked eye, there is difficulty in identifying the obvious discrepancies between the adversarial images and the original images. This finding is critical because it reinforces the substantial threat that these adversarial samples pose to deep learning systems based on chest X-ray images. It is difficult for us to visually distinguish the adversarial image from the original image, which highlights the stealthy nature of adversarial samples. This covertness poses a serious challenge to the diagnostic accuracy and reliability of medical deep learning systems, emphasizing the need for techniques to detect and defend against adversarial samples in medical image deep learning systems.


pneumonia
perturbation
adversarial image(normal)
normal
perturbation
adversarial image(pneumonia)
pneumonia
perturbation
adversarial image(normal)
normal
perturbation
adversarial image(pneumonia)

（a） ResNet-based

（b） DenseNet-based

Figure 3-8 Comparison between the adversarial image and the original image

To test the transferability of the deep learning system, developed for the medical images, to security breaches instigated by adversarial samples, thereby causing erroneous classifications. We tested and investigated the medic deep learning system for chest X-ray images as an example. Initially, MI-FGSM was utilized to introduce minor perturbations to the original images processed by the ResNet model, resulting in the generation of adversarial samples. We used the same test set to test the accuracy of the two models separately independently. It was found that the accuracy of the ResNet-based model decreased from 95.86% to 39.83%, while the accuracy of the DenseNet-based model decreased from 96.03% to 73.62% (Table 3-5). These demonstrate the vulnerability of deep learning systems based on chest X-ray images in the face of adversarial attacks, ultimately leading to significant misclassification rates.

Table 3-5 The accuracy of the model for generating adversarial images based on the ResNet

Chest X-ray images	Accuracy	
deep learn model	ResNet-based	DenseNet-based
Original	95.86%	96.03%
MI-FGSM-attack	39.83%	73.62%

In a parallel approach, we applied the MI-FGSM to introduce slight perturbations to the original images within the DenseNet model, resulting in the creation of adversarial samples. These adversarial samples were then utilized to independently assess the accuracies of both the DenseNet and ResNet models. The results indicated a substantial decline in accuracy for both models, the Resnet model showed a decrease in accuracy from 95.86% to 72.76%. The accuracy of the DenseNet model similarly plummeted from 96.03% to 0% (Table 3-6). This outcome underscores the pronounced susceptibility of these models to adversarial attacks under the given experimental conditions.

Table 3-6 The accuracy of the model for generating adversarial images based on the DenseNet

Chest X-ray images	Accuracy	
deep learn model	ResNet-based	DenseNet-based
Original	95.86%	96.03%
MI-FGSM-attack	72.76%	0%

This outcome demonstrates the high transferability and potential threat posed by the adversarial samples within the context of the deep learning system designed for chest X-ray image analysis. This result suggests that adversarial samples are highly transferable and potentially threatening in deep learning systems designed for chest X-ray image analysis. Thus, it highlights the need to improve the security and robustness of the model when developing a medical deep learning system based on chest X-ray images.

3.5 Summary

We tested and studied deep learning systems based on three different kinds of medical images. In contrast to previous studies on deep learning from medical images, our study focuses on the security of deep learning systems based on medical images and demonstrates that deep learning systems based on medical images have security vulnerabilities and are susceptible to attacks through adversarial attacks.

We developed and trained three kinds of deep learning systems based on a transfer learning approach using medical images to identify and classify abnormal and normal medical images. These deep learning systems are based on the ResNet-50 and DenseNet-121 architectures and achieve an average higher accuracy. To comprehensively assess the security risks inherent in real-world deep learning systems used for medical image analysis, we employed the Fast Gradient Sign Method (FGSM) and Momentum Iterative Fast Gradient Sign Method (MI-FGSM) as a case study. Our findings revealed a significant decline in image recognition accuracy post-attack, underscoring the vulnerability of these models to adversarial interference. Furthermore, the study demonstrated that adversarial images crafted for one model could effectively compromise another completely separate deep learning system designed for medical image analysis, thereby evidencing a concerning level of transferability in these attacks. This phenomenon suggests the need to improve the security and defenses of systems when designing medical images deep learning systems.

4 Defense against White Box Adversarial Attacks

This chapter proposes a generic defense against white-box attacks on medical image deep learning systems.

4.1 Introduction

Deep learning has greatly improved the accuracy of medical image recognition and classification, and it has also provided effective diagnostic aids in the early detection and graded treatment of diseases. However, the security of medical systems based on medically assisted diagnosis is more important than in other deep learning systems. Adversarial attacks are the biggest potential security vulnerability in medical imaging deep learning systems. This can lead to the misdiagnosis of a patient's condition and thus miss the time for treatment. Therefore, the security and reliability of medical deep learning systems is a topic of concern. From the beginning of the design of medical deep learning systems, security and reliability should be prioritized. The application of deep learning to medicine is to better assist physicians in improving the efficiency and accuracy of medical diagnosis, and if the security of deep systems is compromised, it is extremely harmful to both physicians and patients. Improving the security of healthcare systems is a very important topic [91].

Hence, it becomes crucial to explore effective strategies for defending against adversarial examples. Presently, the bulk of research in this area concentrates on altering the architecture of the model. However, such methods often prove impractical for large-scale commercial network models in use today. Therefore, in this paper, we study the security and defense of deep learning systems based on medical images. Based on the trained model in the previous section, for white-box attack algorithms, we construct a generic defense model against such attacks.

4.2 Preliminary

Faced with the threat of adversarial samples, researchers have proposed many methods of adversarial sample defense to protect deep learning models. We classify the defenses as follows：

4.2.1 Gradient masking

Most white-box attacks get the adversarial sample by computing the gradient of the model, so if the gradient of the model cannot be computed, the attack will be ineffective. Gradient masking will change the model to some extent thus making the gradient useless and thus resisting the adversarial sample well.

4.2.2 Detection of adversarial samples

Pre-processing operations are performed on the input data before the model is trained, such as adding detectors and removing adversarial samples in advance if they are detected to avoid affecting the training of the model.

4.2.3 Adversarial training

In the training phase of the model add countermeasure samples or input data for various data augmentation operations, and then train all the data together so that a hardened model can be obtained that can produce a better defense against countermeasure attacks.

4.3 Methodology

The method proposed in this paper consists of three parts. Firstly, normal and abnormal medical images are used as input images, and the two types of images are trained based on the transfer learning method with Desnet-121 as the skeleton to obtain a deep learning model that can accurately identify the two types of images. The trained model is attacked with adversarial attacks to generate adversarial images, which makes the model misclassify the adversarial images, thus also leading to a decrease in the accuracy of the model for image recognition. To overcome this problem, we propose a defensive approach against adversarial attacks, building up a more secure, reliable, and robust defense deep learning system.

Adversarial training is very easy to implement as we can quickly generate adversarial samples by an adversarial attack algorithm and then retrain the adversarial samples. However, it is difficult to exhaust all adversarial samples in adversarial training, and the model is always passive in terms of defense. So is there a way to better simulate the confrontation samples and thus improve the security and defense of the model.

Consider the set of original samples to be $O = \{O_1, O_2, ..., O_n\}$, and the set of adversarial samples to be $A= \{a_1, a_2, ..., a_n\}$, where each O_i represents an original sample and a_i represents the corresponding adversarial sample. The mixed sample set M can then be constructed by alternately arranging the original and adversarial samples as follows:

$$M = \{O_1, a_1, O_2, a_2, ..., O_n, a_n\}$$

To introduce an element of randomness, Gaussian noise characterized by a mean of μ and a variance of σ^2 is added to each sample in the mixed set M; hence, the noise-augmented sample m_i', is given by:

$$m_i' = m_i + N(\mu, \sigma^2)$$

Here, $N(\mu, \sigma^2)$ denotes the Gaussian noise with mean μ and variance σ^2, which is applied to the ith mixed sample.

Consequently, the noise-augmented mixed sample set M' is represented as:

$$M' = (m_1', m_2', \quad ..., m_{2n}')$$

In this way, we obtain a set of mixed sample sets processed with Gaussian noise.

Algorithm ANT

```
Input:
$O$: Set of original samples
$A$: Set of adversarial samples corresponding to $O$
$\mu$: Mean of the Gaussian noise
$\sigma^2$: Variance of the Gaussian noise
Output: $M$: Mixed samples
Initialize $M$ as an empty list
for $i = 1$ to $n$
    Append $O_i$ to $M$
    Append $a_i$ to $M$
Initialize $M'$ as an empty list
for each $m$ in $M'$ :
        noise ← random Gaussian noise with mean $\mu$ and standard deviation $\sigma$
        $m' \leftarrow m + noise$
        Append $m'$ to $M'$
Return $M'$
```

4.3.1 Deep learning system construction based on transfer learning

We constructed three deep learning systems for medical images using the transfer learning approach. The model construction consists of two parts: model training and performance testing (Figure 4-1). In this section, we will discuss these questions in detail.

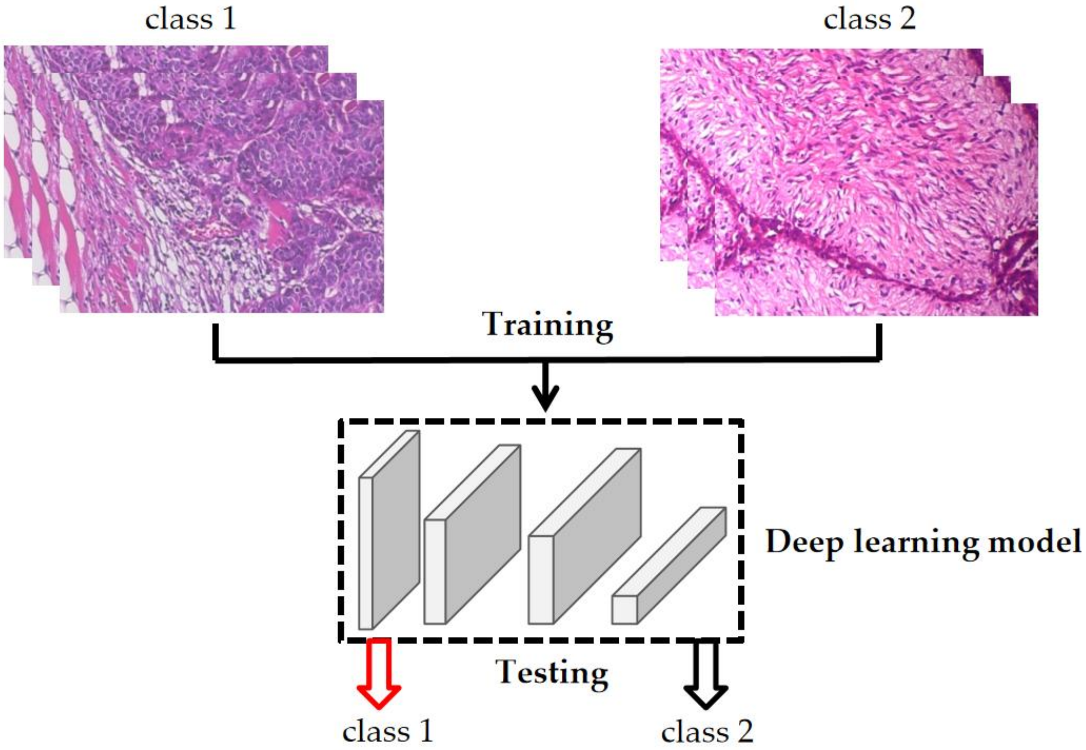


Figure 4-1 The pipeline of building deep learning systems for medical imaging

（一）Datasets

In this study, the medical image datasets used were obtained from public websites, with detailed information provided in Section 3.3 of Chapter 3. The dataset is anonymous and publicly available for non-commercial medical research. We divided the different datasets into three parts: a training set, validation set, and testing set for the medical images deep learning model construction and adversarial attack and defense experiments. The training set was used to train the deep learning model, the validation set was used to tune the hyperparameters of the model, and the test set was used to test and evaluate the performance of the trained deep learning model. Details of the dataset are shown in Table 3-2.

（二） Transfer learning from the DenseNet121 model

Transfer learning is a common approach in deep learning, whereby trained models are used to accomplish new tasks by exploiting the similarity between models and targets. By using transfer learning, we can take an existing trained model, migrate it to our task, and then fine-tune the model for our task-specific requirements to save training costs and time and quickly achieve the task requirements. Due to the small amount of data from the medical images, we adopted a transfer learning approach using the DenseNet121 model pre-trained on medical images to achieve better results. The parameters of the model were frozen, the pooling layer and fully connected layer were replaced, and the dropout layer rate was set to 0.5. The optimizer used adaptive moment estimation (Adam), performed fine tuning using stochastic gradient descent with a learning rate of 1×10^{-3}, and fully changed the connected layer to two classifications. Preprocessing and data augmentation operations were performed on all medical image datasets

4.3.2 Adversarial attack on medical image deep learning systems

One of the security risks of medical image deep learning systems is that the original normal images are modified into abnormal images by maliciously tampering with medical images, thus making the models misclassify them. This leads to the misdiagnosis of the patient’s condition, thus making the patient miss the best time for treatment (Figure 4-2). To test the security of the breast cancer deep learning system, we conducted an adversarial attack on the trained model using breast cancer images as the research object and added subtle interference to the test set, which caused the deep learning model to misclassify the images. We use the FGSM and MI-FGSM algorithms to attack the trained model and generate adversarial samples that are difficult to distinguish with the eye.

4.3.3 Defense against adversarial attacks in medical image deep learning systems

We use an adversarial attack to attack a deep learning model by generating an adversarial image by slightly altering the pixels of the original image to trick the model and cause it to misclassify the image. Adversarial training is a common defense, but usually, only one attack can be methodically performed against an attack. If noise is added to the original image before training the model, the new image is trained so that the model can obtain more feature information from the noisy image. We use the noisy images as input data to train and build a defense deep learning system with the same model and parameters as the original and test the performance of the defense model using the original test set without added noise. Gaussian noise is noise whose probability density function obeys a Gaussian distribution. In the construction of the adversarial defense model, we choose to mix the original images and the adversarial samples and then add Gaussian noise, and finally get a new input image dataset (Figure 4-2).

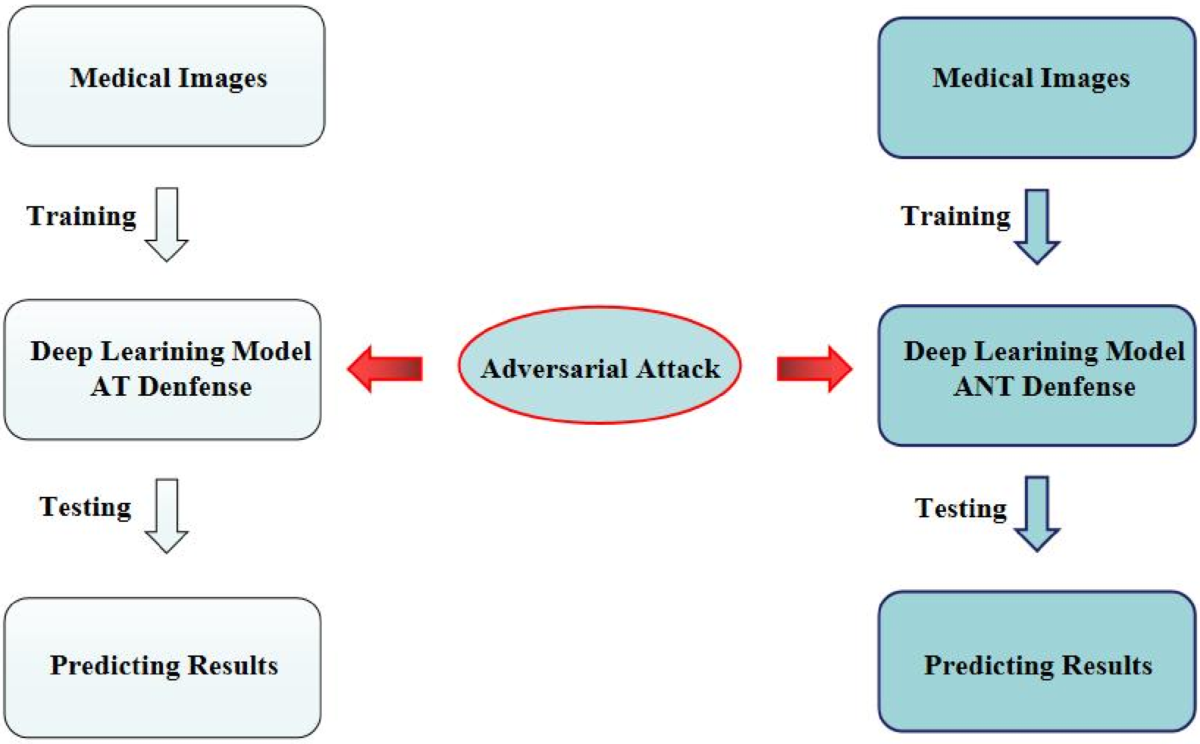


Figure 4-2 Defense against adversarial attacks based on deep learning systems for medical images

4.3.4 Simulate real-world Scenarios to defend against adversarial attacks with breast cancer image deep learning system

We simulate real-world adversarial attack and defenses using the example of breast cancer image deep learning system. One of the security risks of medical image deep learning systems is that the original breast cancer images are modified into benign tumor images by maliciously tampering with medical images, thus making the models misclassify them. This leads to the misdiagnosis of the patient’s condition, thus making the patient miss the best time for treatment. In order to test the security of the breast cancer deep learning system, we conducted an adversarial attack on the trained model using breast cancer images as the research object and added subtle interference to the test set, which caused the deep learning model to misclassify the images. We use the FGSM algorithm to attack the trained model and generate adversarial samples that are difficult to distinguish with the eye.

Unlike the above where the adversarial samples are added to the training of the model, we process the original samples with noise only in our simulated real-world experiments, and then compare the defense performance of the original model and the noisy defense model against FGSM attack. We used an adversarial attack to attack the deep learning model by slightly altering the original image's pixel to generate an adversarial image, thus fooling the model and making it misclassify the image. If we added noise to the original image before training the model, the new image was trained so that the model could obtain more feature information from the noisy image (Figure 4-3).

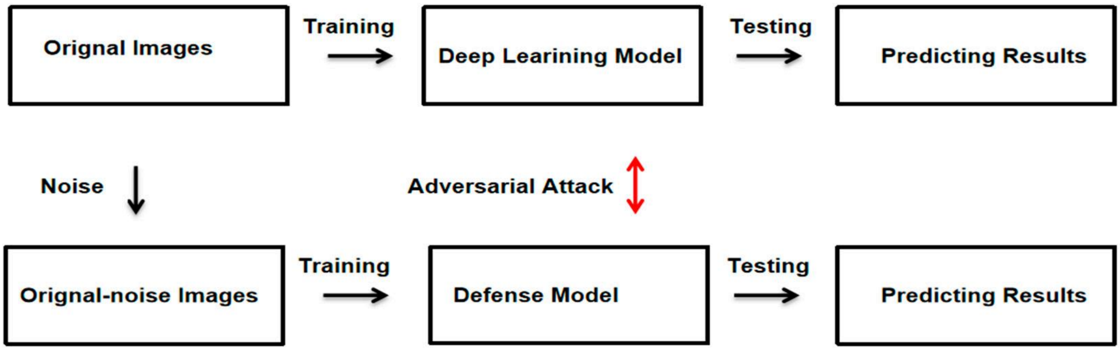


Figure 4-3 Adversarial attack and defense in breast cancer deep learning system

We used the noisy images as input data to train and build a defense deep learning system with the same model and parameters as the original model and test the performance of the defense model with the original test set without added noise. In the construction of the adversarial defense model, we chose to add noise to the original image, and the comparison between the noise image and the original image was as follows. We normalized the original image so that the pixel values are distributed between 0 and 1. We then created a matrix with a noisy image and added noise to the original image to get a new image with noise (Figure 4-4).

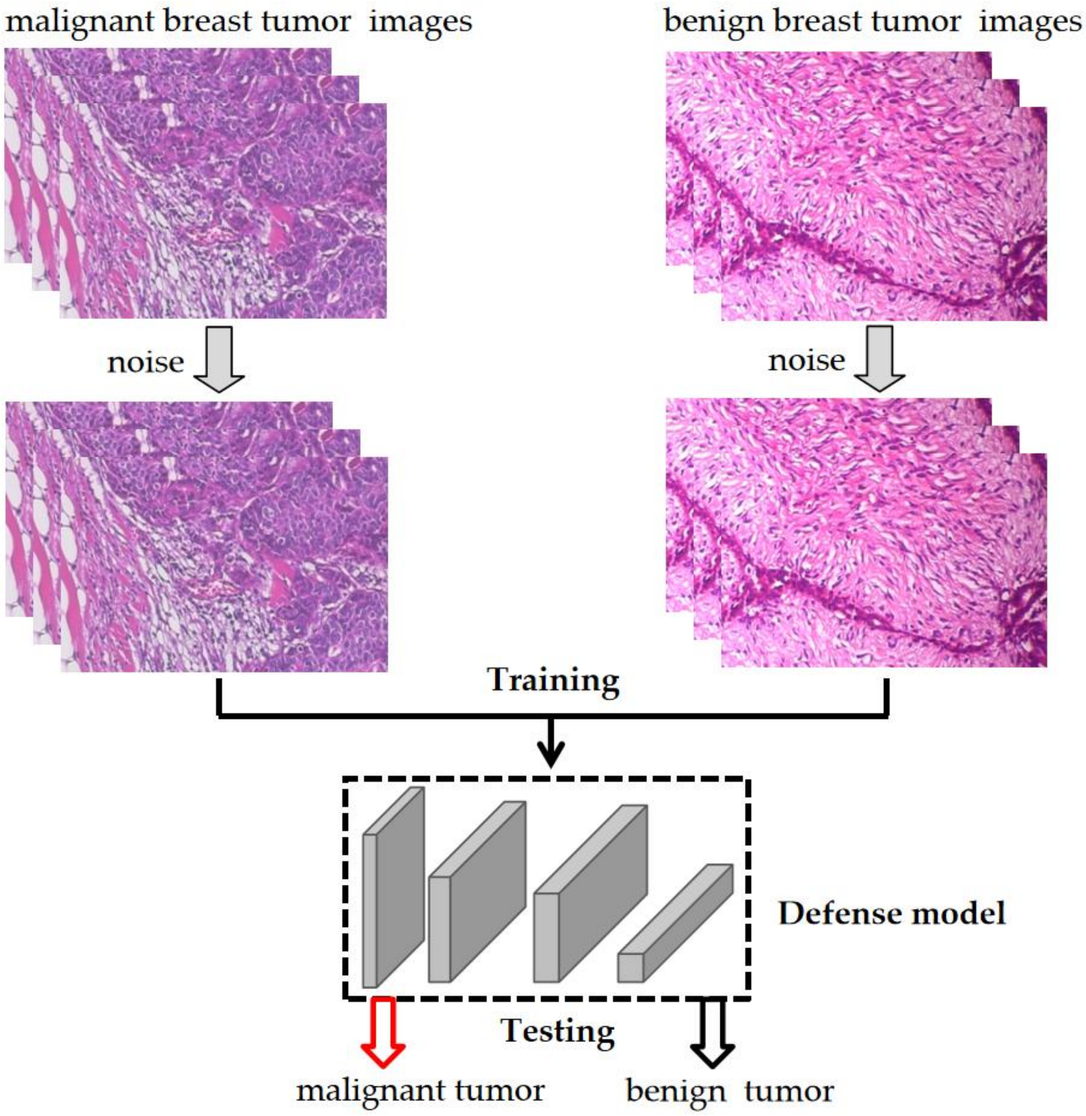


Figure 4-4 Construction of a defense deep learning model for breast cancer

4.4 Results and Discussions

To test the effectiveness of our proposed white-box attack defense algorithm, we compare a series of experiments including a careful comparative analysis of the no-defense, adversarial defense (AT) algorithm and adversarial noise defense (ANT) algorithm. Further, to comprehensively evaluate the performance of the algorithms, we selected three different medical image datasets for testing. The core of the study is to evaluate the robustness and effectiveness of our proposed defense strategy in diverse attack environments.

The performance test results of the deep learning model based on CT images are shown in Table 4-1. On the same test set, the accuracy of the DenseNet-based deep learning system is 84.00% under no-attack conditions, which indicates that the models are both able to accurately recognize COVID-19 and normal medical images with good recognition and classification capabilities. However, when the models are attacked by the adversarial attack algorithms of FGSM and MI-FGSM, the accuracy of the models both decreases rapidly, which indicates that the system is attacked by the adversarial attack algorithms resulting in misclassification of images. When we use the AT defense algorithm to defend against attacks, the accuracy of the model obtained based on MI-FGSM confrontation training is 74.67%in the face of MI-FGSM attacks, however, the accuracy of the model in the face of FGSM attacks is this 49.33%, which indicates that the AT defense can defend against MI-FGSM attack algorithms very well, but cannot defend against the FGSM algorithm. However, when we use the ANT defense algorithm to defend against the attack, the accuracy of the RENT model based on the FGSM confrontation training is 76.00% in the face of the FGSM attack, however, the accuracy of the model in the face of the MI-FGSM attack is 73.33%, which shows that the ANT defense algorithm can defend against the FGSM attack algorithm well, and it can defend against the MI-FGSM algorithm well, with better defense generality and robustness.

Table 4-1 The accuracy of the model based on X-ray image defense against adversarial attack

CT images	Accuracy	
	FGSM	MI-FGSM
No defense	40%	0%
AT defense	74.67%	49.33%
ANT defense	76.00%	73.33%

The performance of the test results of the deep learning models based on breast cancer images is described in Table 4-2. Under no-attack experimental conditions, the deep learning systems achieved high accuracy rates of 98.72% on the test set, confirming the efficiency and reliability of these models in identifying the benign and malignant nature of breast tumors. However, under the condition of no defense again, when the system was subjected to adversarial attacks such as FGSM and MI-FGSM, the performance of both models suffered severely, and the accuracy rate dropped drastically. This phenomenon suggests that although the models perform well under normal conditions, they are still vulnerable in the adversarial attack environment, leading to significant degradation of recognition and classification accuracy. To defend against adversarial attack, we tested the AT defense and ANT defense strategy. Both models have an accuracy of 98.72% for the same test set in the no-attack condition, which shows that both defense methods can recognize breast pathology images well. The AT defense method was able to maintain 96.79% accuracy against the FGSM attack and 44.23% accuracy against the MI-FGSM attack. This result indicates that the AT defense strategy performs better against the FGSM attack algorithm, but is slightly weaker against the MI-FGSM attack. Comparatively, after using the ANT defense algorithm, the model shows a better defense effect against MI-FGSM and FGSM attacks with an accuracy of 94.87% and 59.62%, respectively, compared to the AT defense algorithm. This indicates that the ANT defense strategy can effectively enhance the robustness of the model against different adversarial attack algorithms.

Table 4-2 The accuracy of model based on breast cancer image defense against adversarial attack

Breast cancer images	Accuracy	
	FGSM	MI-FGSM
No defense	33.97%	0%
AT defense	96.79%	44.23%
ANT defense	94.87%	59.62%

The test results of the deep learning models for medical images based on chest X-rays are shown in Table 4-3. Under no-attack conditions, the models based on DenseNet architecture achieve 96.03% accuracy on the test set, which validates their excellent performance in recognizing classified pneumonia and normal chest images. However, these models show significant performance degradation and a substantial reduction in accuracy when faced with the adversarial attacks of FGSM and MI-FGSM without any special defense measures. This suggests that although these models perform well under normal conditions, they exhibit significant vulnerability when subjected to adversarial attacks, leading to a severe degradation in classification accuracy. To deal with these types of adversarial attacks, we comparatively tested the AT defense algorithm and the ANT defense strategy. With AT defense, the model was able to maintain 91.21% accuracy against FGSM attacks and 75.69% against MI-FGSM attacks. These data show that the AT defense method has better resistance to MI-FGSM attacks, but the defense against FGSM attacks is slightly insufficient. Then when the ANT defense method is used, the accuracy of the model in defending against FGSM and MI-FGSM attacks is at 92.24% and 85.34%, respectively. This indicates that the ANT defense method can effectively improve the model's resistance to different types of adversarial attacks, showing broader defense generality and better robustness. While this may not appear as a substantial advancement, it indisputably demonstrates the ANT defensive method's superior security and robustness in the face of adversarial attack challenges.

Table 4-3 The accuracy of model based on chest X-ray images defense against adversarial attack

X-ray images	Accuracy	
	FGSM	MI-FGSM
No defense	83.28%	0%
AT defense	91.21%	75.69%
ANT defense	92.24%	85.34%

To better demonstrate the performance of the two defense methods on different medical datasets, we plot the graphical data as bar charts. The bar chart presented in Figure 4-5 demonstrates the performance of two different defense methods across three distinct types of medical images: CT images, breast cancer images, and X-ray images. The methods evaluated are AT_FGSM (AT defense against FGSM), ANT_FGSM (ANT defense against FGSM), AT_MI-FGSM (defense against MI-FGSM), and ANT_MI-FGSM (ANT defense against MI-FGSM). Overall, the performance of AT defense and ANT defense methods in defending against the FGSM algorithm is quite similar, but when confronted with the more powerful MI-FGSM, ANT_MI-FGSM outperforms the other methods in different image datasets, especially for X-ray images, which demonstrates its robustness and effectiveness in defending against counter-attacks in medical imaging applications.

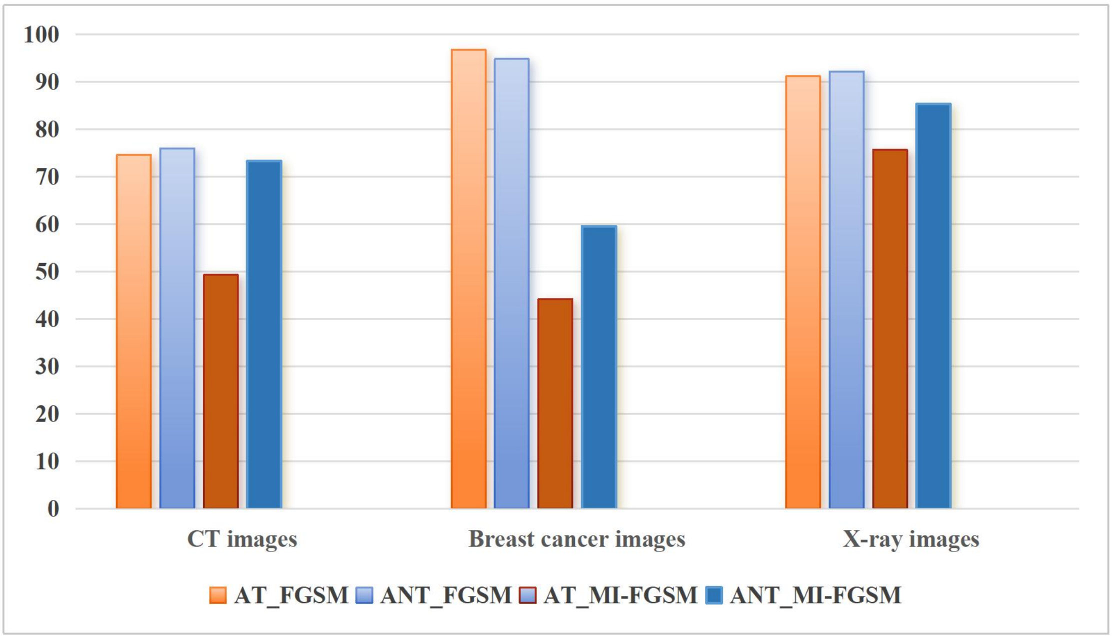


Figure 4-5 The performance comparison of different defense methods across various medical image datasets

The experimental results of using deep learning system for breast cancer images to simulate real-world scenarios against adversarial attacks are as follows. The performance test results of breast cancer image deep learning model are shown in Table 4-4. The accuracy of the original deep learning model was 98.72% and the accuracy of the defense model was 98.08% in the same test set, which indicates that when noise is added to the input image does not affect the recognition and classification ability of the defense model for the image, both the defense model and the original model can accurately identify the medical images of benign and malignant breast tumors with good recognition and classification ability.

Table 4-4 The accuracy of two deep learning models on the same test set

Metrics	Original model	Defense model
Accuracy(%)	98.72	98.08

Figure 4-6 shows the change in the accuracy of the models as the number of training increases. Because the dataset has few images, we train the deep learning model by transfer learning method, the accuracy of the model increases rapidly and the performance reaches saturation quickly.

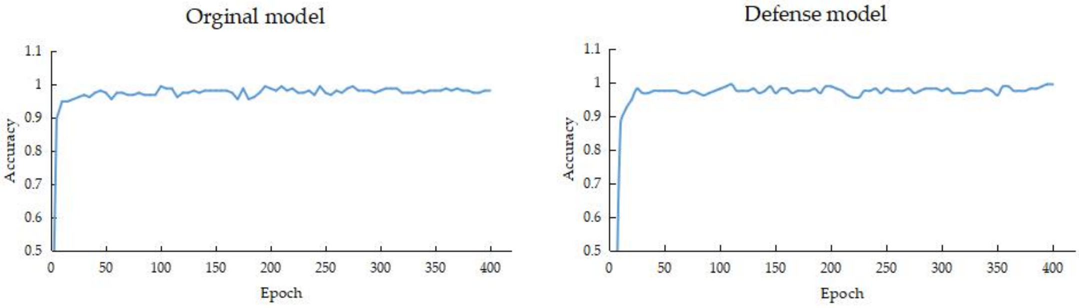


Figure 4-6 Accuracy of deep learning models for breast cancer

To reveal the threat of adversarial attacks on breast cancer deep learning systems and better simulate the security risk of real-world deep systems based on breast cancer images, we used breast cancer images as the research object and attacked the pretrained model with FGSM adversarial attack algorithm, and then tested the defense capability of the original model and the defense model against the adversarial attacks, and the results are shown in Table 4-5. When the original model is attacked by the adversarial attack algorithm, the recognition accuracy of the model decreases from 98.90% to 10.99%, which indicates that the model is successfully attacked by the adversarial attack algorithm and the attack can severely damage the performance of the model. Similarly, when the defense model is attacked by the adversarial attack algorithm, the recognition accuracy of the model decreases from 96.70% to 27.47%, which indicates that the defense model is also successfully attacked by the adversarial attack algorithm, but the recognition accuracy of the defense model increases by 16.48% compared to the original model when facing the same adversarial attack, thus indicating that the defensive model has certain defensive capability against the adversarial attack algorithm compared to the original model. The defense model with better security, reliability, and robustness performance.

Table 4-5 The accuracy of two breast cancer deep learning models subjected to adversarial attack

Attack	Accuracy（%）	
	Original model	Defense model
No-attack	98.90	96.70
FGSM attack	10.99	27.47

To better illustrate the adversarial attack, we compared the original image with the adversarial image, as shown in Figure 4-7. We took a breast cancer image as an example and used the FGSM algorithm to perform the adversarial attack on the deep learning model. The adversarial image was generated by a slight perturbation on the original image, and it was difficult for us to distinguish the difference between the two images with our eyes, but the deep learning model misclassified them, which further illustrates that the adversarial attack is a great threat to the deep learning system.

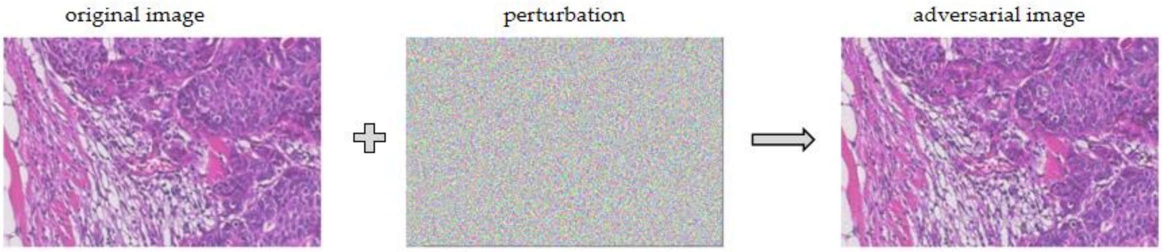


Figure 4-7 Characteristic results of adversarial image and original image

4.5 Summary

In modern medical images diagnosis, deep learning techniques are widely used for automatic analysis and classification of medical images. These systems can learn effective feature representations from large-scale datasets, providing accurate and fast diagnosis. However, medical images often contain a high amount of noise, which makes them a potential target for adversarial attacks. By adding small perturbations to the original image, adversarial attacks can make the model produce serious errors in classification. In particular, white-box attacks allow the attacker to have full knowledge of the structure and parameters of the target model, thus generating highly threatening adversarial samples.

Based on the research in Chapter 3, we tested that white-box adversarial attack algorithms can attack the medical image system, which makes the model misclassify the images. To address the hazard of white-box adversarial attack algorithms on the security of medical image-based deep learning systems in Chapter 3. In order to address the threat of adversarial attacks on the security of medical image-based deep learning systems, we propose a method that can defend against adversarial attacks, thereby effectively reducing the success rate of adversarial attacks and improving the security and reliability of deep learning systems. Medical images contain more noise, resulting in a malicious attacker adding noise value to the image of the interference is also not easy to be detected by the human eye, through the adversarial training defense method can improve the model's ability to defend against adversarial samples, however, the adversarial training is usually only able to resist one attack method. Compared to ordinary adversarial training, we introduce clean samples and Gaussian noise during the adversarial training process, which not only improves the model's recognition and classification accuracy of clean samples but also makes the model have generalized defense capability against different white-box attack algorithms.

In order to better simulate the real-world attack and defense, we only take the breast cancer image as the attack target, and use FGSM algorithm to attack only the breast cancer image to make it recognized as a conscientious tumor image. Secondly, in order to compare the defense method without confrontation training, we constructed a medical image deep learning defense model with noise only, and found that its defense ability is relatively limited, while the defense model with the addition of adversarial samples and noise has better defense performance.

5 Black Box Adversarial Attacks on Medical Image Deep Learning Systems

This chapter proposes black-box attack methods with more efficient attacks, capable of generating smaller perturbations with the same query budget.

5.1 Introduction

White-box attacks have full information about the target model, such as model structure, parameters, etc., and can precisely design the attack method to maximize the effect, which means that fewer attempts are needed to quickly and efficiently counter the sample. However, in real-world application scenarios, it is difficult for an attacker to obtain the internal information of the target model, which makes white-box attacks less common in practice than black-box attacks.

Black-box attacks do not require the attacker to know the internal structure and parameters of the target model and only need to infer and construct the attack through the inputs and outputs of the model. This type of attack is more in line with real-world attack scenarios. Black-box attacks are particularly problematic in deep learning systems for medical images. For example, an attacker can manipulate a diagnostic system by subtly altering medical images to produce erroneous outputs. This manipulation may be difficult for a human observer to detect, but it can cause the AI system to misdiagnose a condition, which may lead to a delay in the patient's condition. This vulnerability is exacerbated by the reliance on deep learning models, which are inherently complex and operate like black boxes. The opaque nature of these models makes it difficult to detect when images have been tampered with or when the model has been compromised in the decision-making process. In addition, the deployment of machine learning models in healthcare environments often requires adherence to strict regulatory standards for accuracy and reliability. By studying and testing black-box attacks, we can better understand and strengthen the security vulnerabilities of deep learning models for medical image analysis. This can help build more secure systems that prevent malicious modification of medical images or misleading diagnostic results.

5.2 Preliminary

The majority of deep learning models in the real world are black-boxed, where the internal structure and parameters of the model are invisible to the user, thus making it impossible for an attacker to understand the details of the target model. The black-box nature makes it necessary for the attacker to attack the model by other means. Black-box attacks can be further categorized into query-based attack, decision-based attack, transfer-based attack, etc., depending on how the attacker interacts with the target model and the level of knowledge.

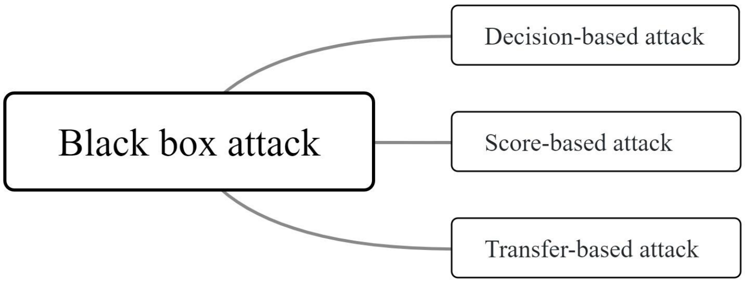


Figure 5-1 The classification of black box attack

5.2.1 Decision-based attack

Brendel et al [92] first introduced the concept of decision-based attack, and designed and proposed the Boundary Attack algorithm. Boundary Attack is an algorithm for generating adversarial samples, and its main goal is to adjust the input samples by iteration in black-box attack scenarios, i.e., when the attacker does not have direct access to the gradient information of the target model, and then keep moving along the decision boundary between the adversarial and non-adversarial regions, which gradually converges to the decision boundary of the model while maintaining the adversarial nature of the model to make the model misclassified.

Cheng et al. [93] proposed the OPT attack, which redefines the decision-based attack as an optimization problem of finding a direction with the shortest distance from the original point to the decision boundary. They have framed the generation of adversarial examples in black-box settings with hard labels as a real-valued optimization problem. In this framework, the variable θ represents the search direction, and the function $g(\theta)$ denotes the distance from the original sample x_0 to the decision boundary along the direction θ. Essentially, the optimization problem aims to identify an optimal direction θ, in which the distance from the initial sample to the decision boundary is minimized. However, in the process of finding the distance, the bisection search consumes a large number of queries, so the OPT algorithm still lacks query efficiency.

Chen, J. et al [94] HSJA investigates and optimizes decision-based attacks and presents a series of novel algorithms HopSkipJumpAttack (HSJA) for generating targeted and untargeted adversarial examples. The algorithms are inherently iterative, with each iteration involving three steps: estimation of the gradient direction, step search via a geometric progression, and bounded search via dichotomy, and a theoretical analysis of the pairwise optimization framework and gradient direction estimation is presented.

5.2.2 Score-based attack

Score-based black-box adversarial attacks are a specific class of attacks that target systems that can only provide model output scores, e.g., classification probabilities, without providing information about the internal gradient. These attacks are crucial for evaluating the robustness of machine learning models because they can be performed without knowing the internal working mechanism of the model. In this attack method, the attacker does not need to access the internal structure of the model or the gradient information and generates adversarial samples only by accessing the output of the model, i.e., the confidence scores of the classification. The core idea of score-based black-box attacks is to iteratively adjust the input data by the scores of the model outputs until an adversarial sample is found that can make the model misclassify.

5.2.3 Transfer-based attack

Transfer-base attack is carried out by attacking the migration ability, by attacking the source model so that it can get all the information of the model, calculating the perturbation of a certain picture, and then attacking the target model directly with the perturbed antagonistic picture, so the focus of this class of methods is how to enhance the migration of the perturbation.

5.3 Methodology

Black box attack is where the attacker lacks complete internal information and access to the target model and can only observe the behavior of the model through inputs and outputs. The attacker cannot obtain detailed information about the structure, parameters, algorithms, etc. of the target model. In black-box attacks, the attacker usually infers the behavior of the model through trial and analysis and generates specific inputs to deceive or destroy the target model. Boundary-based black box attack does not rely on internal access to the target model, such as gradients or model parameters. This makes it suitable for attacking black-box models, even if the attacker only has access to the inputs and outputs of the model. It is more efficient in finding successful adversarial samples than other gradient-free attack methods because it optimizes the attack direction by estimating the direction of the decision boundary, reducing the number of queries required. It can do this by approximating the decision boundary and jumping around it. It can generate small but effective perturbations that have a large impact on the model but a small impact on human perception by approximating the decision boundary and jumping around it.

Although the previous HSJA adversarial box attack has shown effectiveness in a variety of scenarios, it has certain limitations, particularly the problem of gradient conflict. Gradient clashes occur when the estimated gradient that guides the adversarial perturbation is not accurately aligned with the true gradient direction of the model. This misalignment can lead to suboptimal perturbations that reduce the effectiveness of the attack and increase the number of queries required to generate successful adversarial examples. Due to gradient conflicts, similar approaches typically require a large number of queries to the model to generate valid adversarial examples. This inefficiency is a key drawback, especially in cases where the query budget is limited and the number of queries allowed is restricted. In such cases, achieving a successful attack within a limited number of queries becomes challenging. PCGrad is an optimization method for multi-task learning, which aims to solve the problem of gradient conflicts between different tasks [95]. In multi-task learning, the objective functions of different tasks may lead to conflicts in the direction of model parameter updates, thus affecting the training effect. It canreduces the gradient conflicts between different tasks by projecting the gradients, thus improving the training effect of the model. To address the limitations of this problem, we propose a new method Gradient Based Black Box Attack (GBBA) based on the HSJA algorithm.

The main goal of GBBA is to mitigate the gradient conflict problem by employing a more accurate gradient estimation technique, thereby improving the efficiency and effectiveness of black box attacks.

Algorithm GBBA

```
Input：
Estimation Gradient $g$
Original sample $x$,
Adversarial sample $x_{adv}$
Output:
  Final Gradient $g_{adv}$
Procedure:
    Require:
     $g$: The gradient estimated from the model.
     $x$: The original sample from the dataset.
     $x_{adv}$: The adversarial sample generated to test the model's robustness.
  Ensure:
     The output is the adjusted gradient $g_{adv}$.
  If  $cos(g,\ x_{adv} - x$）$<0$:
      $g_{adv} = g - \frac{g \cdot (x_{adv} - x)}{||x_{adv} - x||^2}(x_{adv} - x)$
  Else:
     $g_{adv} = g$
  Output $g_{adv}$
```

5.4 Experiment

In this section, we will provide a detailed introduction to the experiment.

5.4.1 Development of medical images deep learning systems

Training and testing deep neural models are key steps when building a deep learning system for medical images (Figure 5-2). This process requires the selection of appropriate datasets and careful selection of suitable deep neural network models. These key aspects are explored in detail in this section.

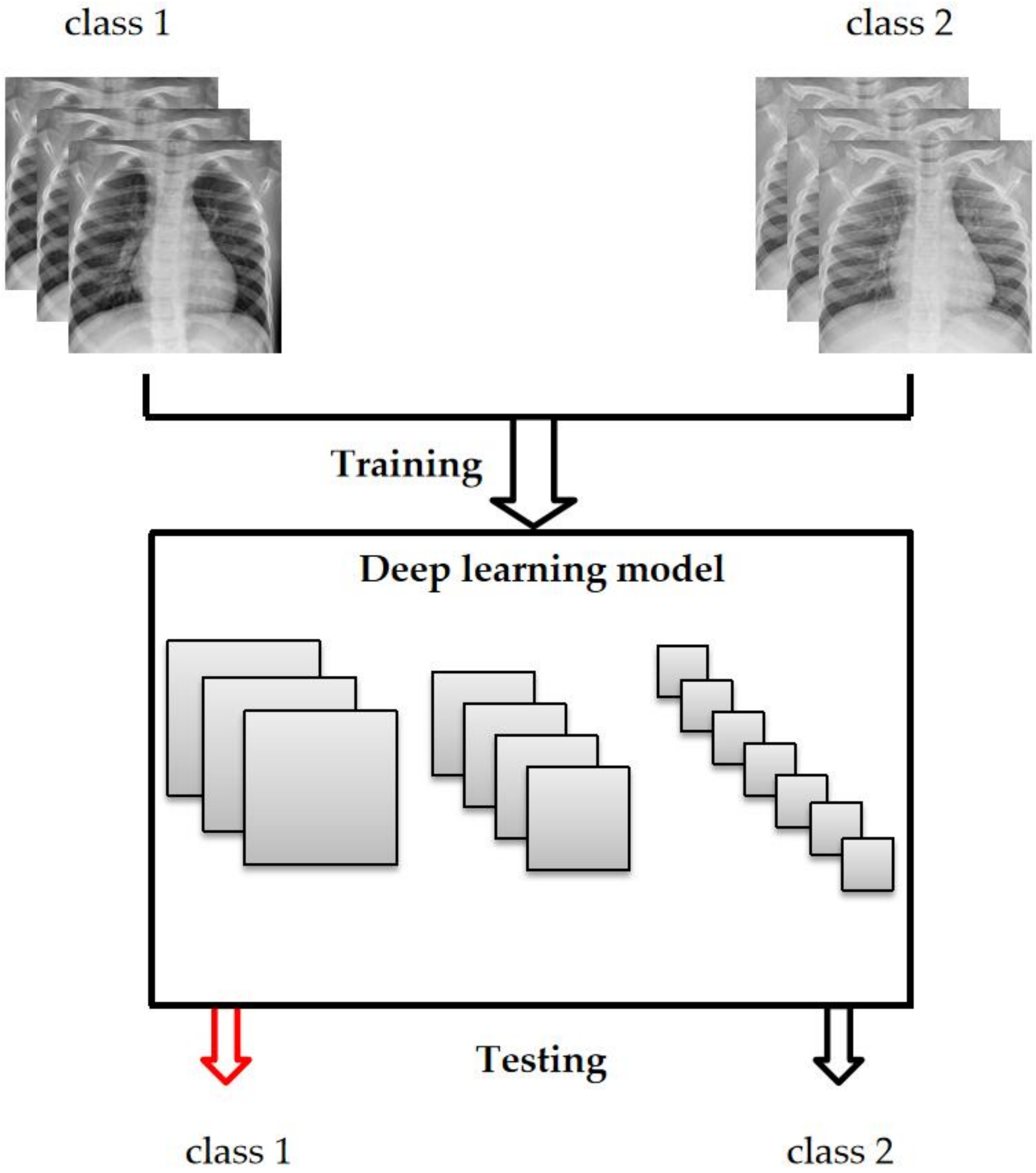


Figure 5-2 The pipeline of the medical images deep learning systems based on black box attack

（一）Dataset

In this research, We selected three types of public datasets, which are the chest CT image dataset, chest radiograph dataset, and breast cancer pathology image dataset. The medical image datasets utilized were sourced from publicly accessible websites. Detailed descriptions and sourcing information for these datasets are thoroughly documented in Section 3.3 of Chapter 3. The datasets are completely anonymized and are made available for non-commercial medical research purposes. We organized the entire dataset into two distinct subsets: a training set, and a testing set, which are specifically designed for processing medical images. Details of these datasets are presented in Table 5-1.

Table 5-1 The classification of datasets

Datasets	Medical images	Training set	Testing set	Total
CT images	Non-COVID-19	238	159	397
	COVID-19	209	140	349
Breast cancer	Benign	515	129	644
	Malignant	722	181	903
X-ray images	Normal	1264	316	1580
	Pneumonia	3376	844	4220

（二）Transfer Learning from the DenseNet121 Model

Transfer learning has become an important technique in deep learning, and is particularly useful in situations where labeled data is scarce or training from scratch is computationally expensive. In medical imaging, where data is often limited and highly specialized, migration learning offers a practical solution. The DenseNet network was designed to connect each layer directly to its preceding layers to achieve the reuse of features and to effectively solve the gradient disappearance problem while designing each layer of the network to be particularly narrow, requiring only a very small number of feature maps to be learned, thus substantially reducing the number of parameters. In this study, we use DenseNet121 models as the subjects of our black-box attack experiments, and we employ some fine-tuning strategies to adapt these models to our specific tasks in medical imaging. First, data augmentation techniques were applied to enhance the diversity of the limited dataset, including rotation, translation, and scaling modifications. Subsequently, we resized all input images to a uniform size of 224×224 pixels to match the input size requirements of the pre-trained network. The model was trained for 400 rounds using the Adam optimizer. We set the batch size to 32 and initiated training at a learning rate of 0.001 and dynamically adjusted it based on the performance metrics observed during training.

5.4.2 Black-box adversarial attack on the medical images deep learning systems

A black-box attack against a medical deep learning system is a type of cyber attack in which the attacker does not know the internal structure of the system. The goal is usually to manipulate the output or functionality of the system by providing carefully crafted inputs that are designed to exploit vulnerabilities in the model.

5.4.3 Metrics

The Euclidean distance corresponds to the actual straight-line distance between two points, which is very intuitive in a geometric sense [96]. This intuition makes it easy to understand and explain, especially in physical space and real-world applications. His calculations are simple and easy to implement. For two points, simply calculate the sum of the squares of the differences based on their coordinates and then take the square root to get the distance between them. Euclidean distance can be used in any dimension of space, from the two-dimensional plane to high-dimensional data spaces. This makes it widely used in several fields, such as machine learning, data mining, computational geometry, image analysis, etc.

The mathematical formula for Euclidean distance can be expressed as follows: Given the coordinates of two points $P_1(x_1, y_1, z_1, ...)$ and $P_2(x_2, y_2, z_2, ...)$ in n-dimensional Euclidean space, the Euclidean distance d between these two points can be computed using the following equation:

$$d\left(P_1,\ P_2\right) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 + \ldots}$$

The *lp* distance refers to the median *lp* distance between the original image and the perturbed image. In black-box attacks, the initialization phase makes the initial point lie within the adversarial region and maintains the adversarial nature throughout the attack. During the attack, the *lp* distance allows a visual comparison of the performance and effectiveness of each algorithm. In this chapter, the l_2 distance is used to measure the performance and performance of the algorithm, which is measured by the l_2 mean distance and median distance for the same image set respectively. Under the same number of queries, the smaller the *lp* distance is, the more efficient the algorithm is in attacking.

5.5 Results and Discussion

This section shows the experimental results.

To verify the performance of our proposed black-box attack algorithm, we compare it with black-box attack algorithms such as OPT and HSJA. The experiments apply all the attacks to the same dataset and target model, and also to compare the performance of the algorithms, we test the black-box attack algorithms using three different medical image datasets.

Tables 5-2 show the average l_2 distance and median l_2 distance of the OPT, HSJA, and GBBA attack algorithms for attacking sums with different query budgets for the CT medical image dataset. From the following tables, it is obvious that our proposed GBBA algorithm generates the smallest l_2 distance perturbation under a limited number of queries compared to the OPT, and HSJA block box algorithms. Taking the chest CT images dataset as an example, our proposed HJSA method produces the smallest perturbation among the four algorithms for a query count of 1000. This proves that the GBBA algorithm is effective in reducing the size of the added perturbations at the beginning of the attack. As the number of queries increases, both the mean and median l_2 distances are minimized, which indicates that the GBBA algorithm attacks are significantly better than the other three algorithms.

Table 5-2 The query counts and l_2 distances for the CT image deep learning system

CT images	**Attack**	**1K**	**5k**	**10k**	**15k**	**20k**
Average l_2 distance	OPT	34.00	28.43	23.46	21.98	18.55
	HSJA	32.56	20.68	15.62	13.75	12.16
	GBBA	29.11	13.84	7.19	4.98	4.56
Median l_2 distance	OPT	31.26	23.91	19.15	16.10	13.43
	HSJA	26.65	12.69	9.12	7.54	6.33
	GBBA	20.11	7.04	4.19	2.74	2.33

In order to better compare the effects of the three black-box attacks on different medical image datasets, we plot the experimental results into graphs, where Figure 5-3 shows the deep learning system based on COVID-19 medical images. The solid line in the figure represents the mean l_2 distance and the dashed line represents the median l_2 distance. It is obvious from the figure that with the increase of the number of queries, the l_2 distance becomes smaller and smaller, which indicates that increasing the number of queries can significantly reduce the gap between the adversarial image and the original image. Specifically, it can be observed in the graph that with the increase in the number of queries, the various methods' l_2 distance all show a decreasing trend. This indicates that by increasing the number of queries, the black-box attack can generate adversarial samples that are similar to the original image more effectively, thus reducing the visual difference between the adversarial samples and the original image. This is of great significance for improving the stealthiness and attack effect of the black box attack. Among the three different black-box attack methods, the l_2 distance of GBBA is smaller than the other two algorithms. Both the mean l_2 distance and the median l_2 distance, GBBA exhibits lower values. This clearly shows that GBBA outperforms the other algorithms and can generate higher-quality adversarial samples in fewer queries.

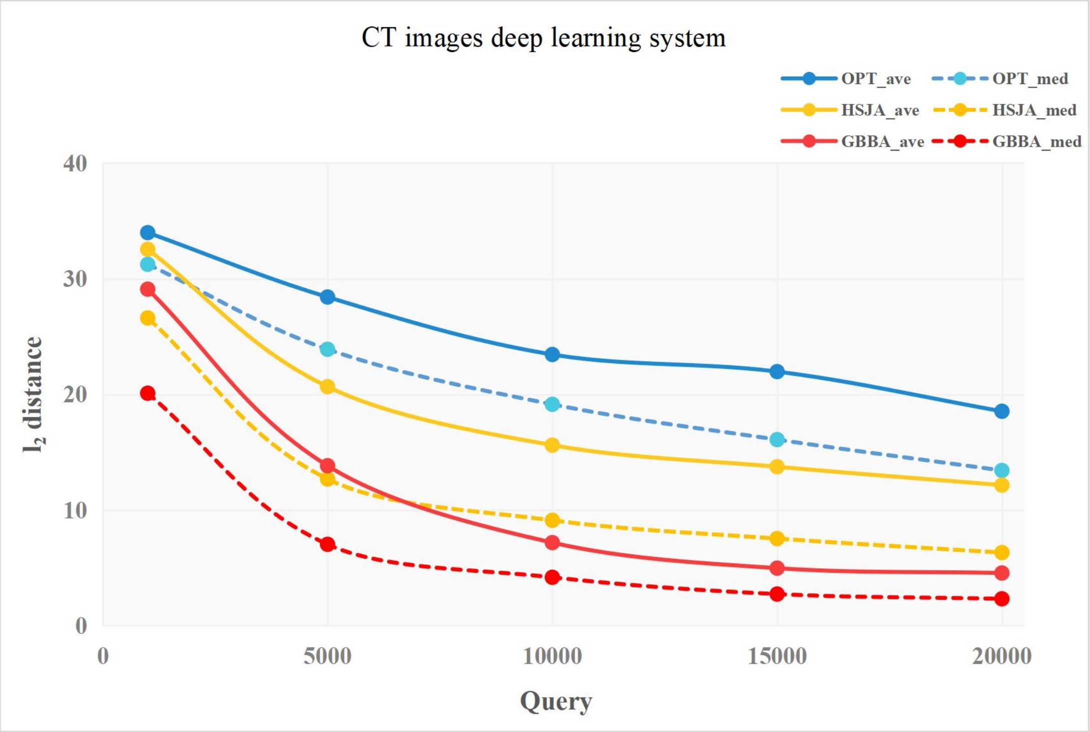


Figure 5-3 The plots of query counts and l_2 distances for CT image deep learning system

To better demonstrate the difference between the original image and the adversarial image, we compared the original image and the adversarial image based on the COVID-19 deep learning system under different query counts, Figure 5-4 shows the comparison graphs of the original image and the adversarial image of the GBBA algorithms under different query counts, and it can be seen that, as the query counts increase, the noise of the adversarial image is getting smaller and smaller and the difference with the original image is getting smaller and smaller and the noise of the adversarial image produced by the GBBA algorithm, which reflects the superior performance of the GBBA algorithm.

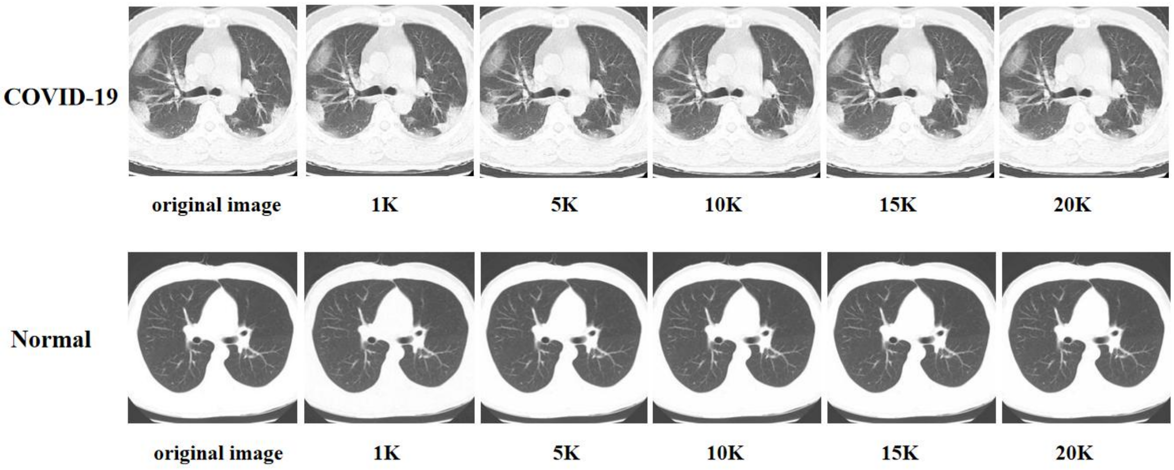


Figure 5-4 Comparison of adversarial samples and original samples of CT images under GBBA attack

Using the dataset of breast cancer images as an example, our proposed GBBA method introduces the least amount of perturbation compared to the other four algorithms when evaluated at 1000 queries. This demonstrates the efficiency of the GBBA algorithm in minimizing the initial perturbation during an attack. Furthermore, as the query count rises, both the average and median l_2 distances continue to decrease, signifying that the GBBA algorithm performs substantially better in reducing distortions than the other three algorithms.

Table 5-3 The query counts and l_2 distances for breast cancer image deep learning system

Breast cancer images	**Attack**	**1K**	**5k**	**10k**	**15k**	**20k**
Average l_2 distance	OPT	25.28	19.13	15.58	13.13	11.62
	HSJA	19.49	10.02	7.44	6.43	5.58
	GBBA	15.78	6.91	4.55	3.53	2.92
Median l_2 distance	OPT	23.64	16.66	12.84	10.66	9.28
	HSJA	15.71	8.07	5.88	5.03	4.45
	GBBA	12.13	5.18	3.57	2.77	2.28

To better compare the effectiveness of the three black-box attacks on different medical image datasets, we plotted the experimental results as graphs, as shown in Figures 5-5. It is obvious from the figure that the distance of l_2 is getting smaller and smaller with the increase in the number of queries, which indicates that increasing the number of queries can significantly reduce the gap between the adversarial image and the original image. In addition to this, the l_2 of GBBA is smaller than the other three algorithms on three different datasets, which all indicates that GBBA performs better than the other algorithms.

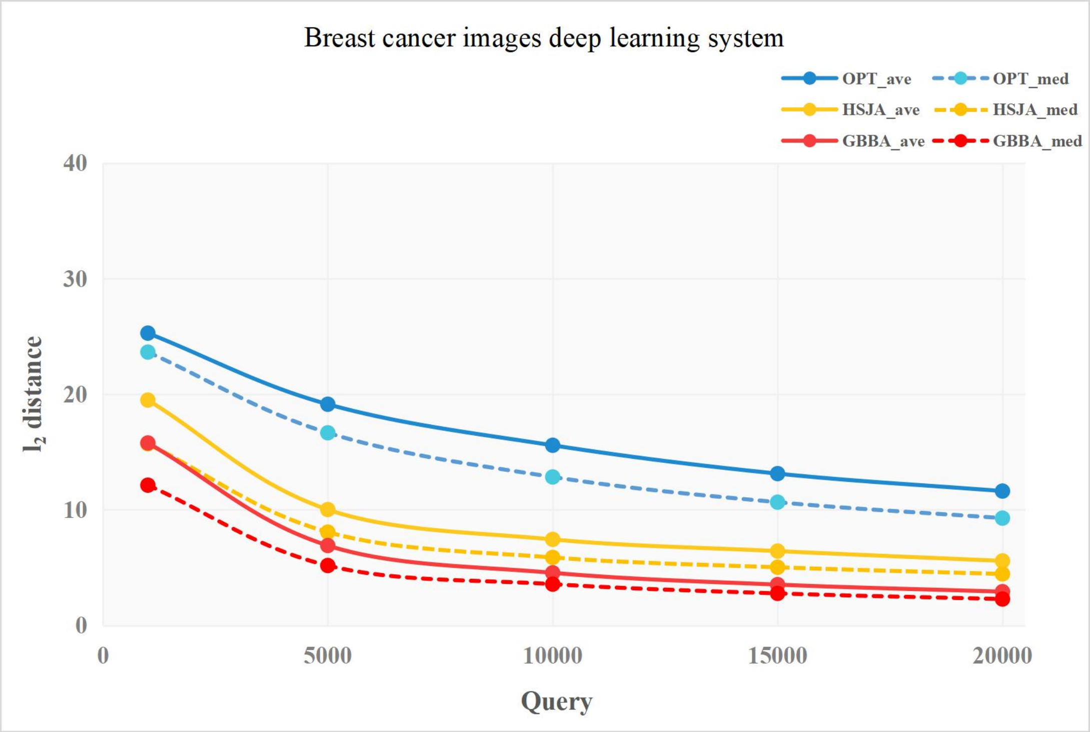


Figure 5-5 The plots of query counts and l_2 distances for breast image deep learning system

In order to better demonstrate the difference between the original image and the adversarial image, we compare the original image and the adversarial image based on the breast cancer deep learning system under different numbers of queries. Figures 5-6 show the comparison graphs of original and adversarial images of GBBA algorithms under different numbers of queries. When the initial number of queries is low, there is obvious noise and differences between the generated adversarial image and the original image. However, as the number of queries increases, these noises and differences gradually decrease. This indicates that by increasing the number of queries, the black-box attack can more effectively generate adversarial samples that are more similar to the original image, thus reducing the differences between the adversarial image and the original image. This trend also better reflects the threat of adversarial samples to deep learning systems for medical images.

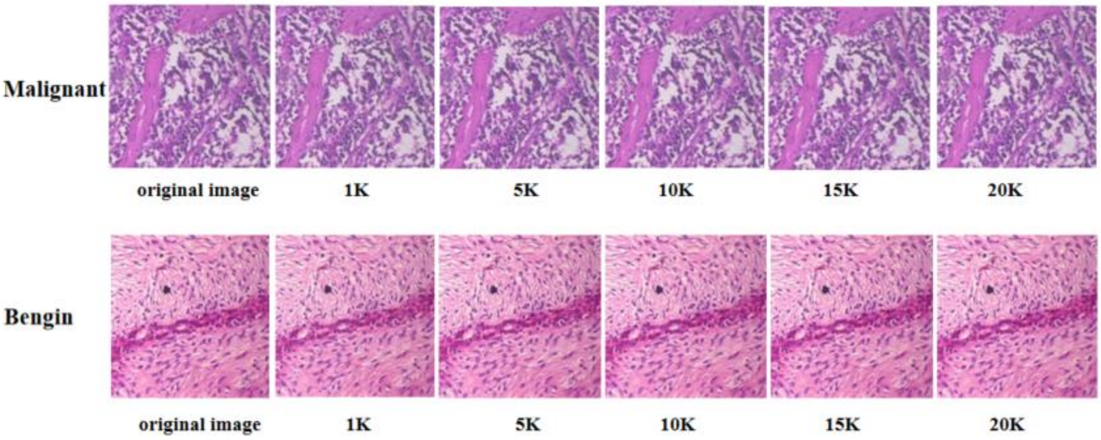


Figure 5-6 Comparison of adversarial samples and original samples of breast cancer images under GBBA attack

Similarly for the chest X-ray dataset, our proposed GBBA method introduces the least amount of perturbations compared to the other four algorithms when evaluated over 1000 queries (Table 5-4). This demonstrates the performance of the HJSA algorithm in minimizing the initial perturbations during the attack. Furthermore, both the mean l_2 distance and median l_2 distance continue to decrease as the number of queries increases. This trend suggests that the GBBA algorithm not only performs best in the initial phase of the attack but also improves its performance with more queries. This enhancement is useful because it shows that the algorithm improves its attack strategy by optimizing the perturbations more efficiently as the number of queries increases. Such performance highlights the robustness and adaptability of the GBBA algorithm.

Table 5-4 The query counts and l_2 distances for chest X-ray image deep learning system

X-ray images	**Attack**	**1K**	**5k**	**10k**	**15k**	**20k**
Average l_2 distance	OPT	39.27	36.13	32.53	29.47	26.81
	HSJA	32.81	14.44	7.47	5.29	4.03
	GBBA	29.35	11.05	4.74	2.95	2.23
Median l_2 distance	OPT	38.84	35.84	32.59	29.91	27.37
	HSJA	32.59	13.96	6.87	4.95	3.89
	GBBA	29.55	10.39	4.49	2.91	2.22

To better compare the effectiveness of the three black-box attacks on chest X-ray medical image datasets, we plotted the experimental results as graphs, as shown in Figure 5-7. The solid lines represent the mean l_2 distances, and the dashed lines represent the median l_2 distances. It is obvious from the figure that the l_2 distance decreases progressively with the increase in the number of queries. This trend indicates that increasing the number of queries can significantly reduce the gap between the adversarial image and the original image. The graph shows that initially, with fewer queries, the l_2 distances for all three black-box attack methods are relatively high, suggesting a significant difference between the adversarial samples and the original images. However, as the number of queries increases, the l_2 distances decrease substantially. This indicates that more queries allow the attacks to refine the adversarial samples more effectively, making them more similar to the original images. Among the three black-box attack methods, GBBA consistently shows a smaller l_2 distance compared to the other two algorithms across chest X-ray image datasets. This is evident in both the mean and median l_2 distances, demonstrating that GBBA outperforms the other algorithms.

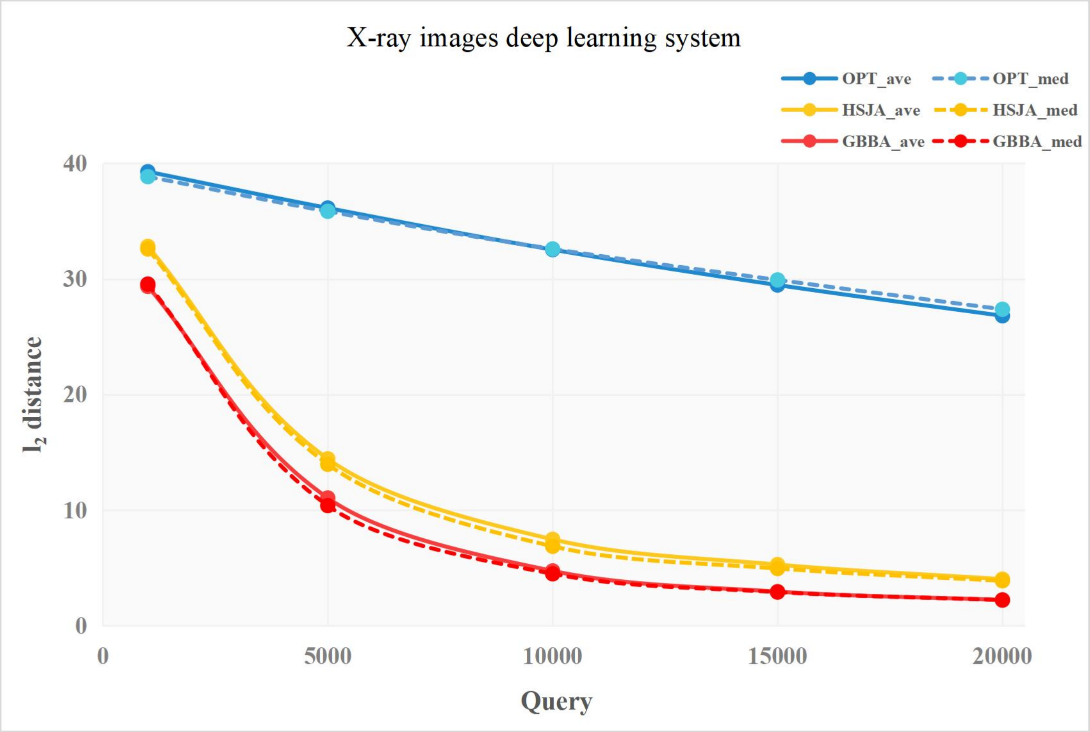


Figure 5-7 The plots of query counts and l_2 distances for chest X-ray image deep learning system

To better illustrate the differences between original images and adversarial images, we compared them under different query counts using a chest X-ray deep learning system. Figure 5-8 presents the comparison of original and adversarial images generated by GBBA algorithms at various query counts. The figure clearly shows that as the number of queries increases, the noise in the adversarial images decreases, making them more similar to the original images. Initially, with a low number of queries, there is significant noise and difference between the adversarial images and the original images. However, as the number of queries increases, this noise and difference gradually decrease. This indicates that increasing the number of queries allows black-box attacks to generate adversarial samples that are more similar to the original images, thereby reducing the gap between them. This indicates that the GBBA algorithm has a notable advantage in generating adversarial samples. Regardless of whether the query count is low or high, the adversarial images produced by the SA algorithm show smaller differences from the original images.

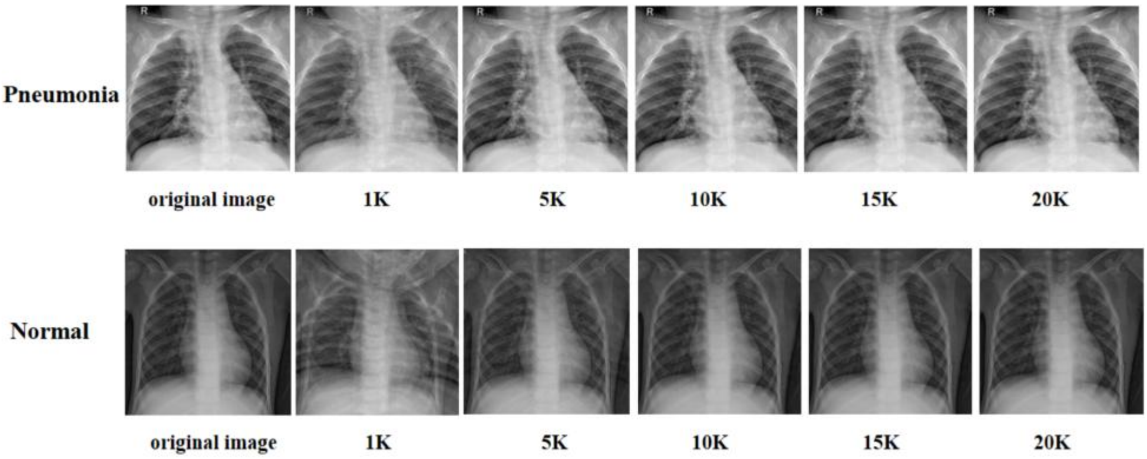


Figure 5-8 Comparison of adversarial samples and original samples of X-ray images under GBBA attack

The experimental results show that increasing the number of queries can effectively reduce the difference between the adversarial image and the original image. Among the three black-box attack methods studied, the GBBA algorithm consistently generates significantly less noisy adversarial images, demonstrating superior performance. This finding is crucial for practical applications, especially in the field of medical image analysis and security, where choosing the right attack methods and strategies can significantly improve the quality and effectiveness of adversarial samples.

5.6 Summary

In this chapter, we propose a black-box adversarial attack algorithm based on decision boundaries, which explores the structure of decision boundaries through bisection search and gradient decomposition to generate adversarial samples more efficiently in the black-box case. We demonstrate the effectiveness and advantages of this new approach through comparative experiments with algorithms such as OPT, and HSJA. Compared with other algorithms, this approach may exhibit higher attack efficiency and generate more misleading adversarial samples, which is crucial for the security and robustness of deep learning systems for medical images.

Our research aims to address the challenge of black-box attacks on deep learning systems for medical images, for which a feasible solution is proposed. By deeply analyzing the structure and characteristics of decision boundaries, our algorithm can generate adversarial samples more accurately, thus revealing the weaknesses of medical image deep learning models. Meanwhile, by comparing with existing attack algorithms, we verify the obvious performance advantages of the proposed method, which provides important support for future research on the security of medical image deep learning systems. The contribution of this study is not only to expand the research on attacks on medical image deep learning systems but also to provide new insights for understanding model robustness. By deeply exploring the structure of decision boundaries, we provide new ideas and methods for designing more secure and reliable deep learning systems for medical images.

6 Defense against Black Box Adversarial Attack

This chapter presents an efficient and easily deployable defense against black-box attacks on medical image deep learning systems.

6.1 Introduction

As deep learning technologies are widely used in various industries, especially in the medical and health fields, their security issues are gradually emerging, especially the threat of black-box attacks. Such attacks do not require in-depth knowledge of the internal structure of the target model and training data and only use the output of the model to generate adversarial samples that can mislead the model's decision, posing serious challenges to security-sensitive applications. The stealthiness and feasibility of this attack make medical images deep learning systems face serious security challenges. In the field of medical image processing, such attacks can lead to misdiagnosis or omission, thus posing a serious risk to patient safety. Although existing research has proposed various defense strategies, they still fall short in terms of effectiveness, complexity, and practicality. Therefore, it is crucial to ensure the security and reliability of medical deep learning systems against black-box attacks in real-world medical applications. In addition, exploring effective defense mechanisms against black-box attacks is crucial to ensure the security and reliability of deep learning models.

This study aims to lay a more solid foundation for the security applications of medical image deep learning by proposing a new defense framework that not only defends against existing black-box attack methods but also improves the model's ability to defend against unknown attacks. Through a series of experiments, we verify the effectiveness and superiority of the framework on multiple public datasets, further advancing the research in the field of deep learning security.

6.2 Preliminary

In a defense strategy against black-box attacks, especially when dealing with

Artificial Intelligence (AI) systems, we must take into account that an attacker usually does not have access to the internal system, but can attack through the system's external interfaces. Therefore, defense strategies should focus on strengthening these interfaces and the system's resistance to anomalous inputs and behaviors. The current black box attack defense methods are as follows:

6.2.1 Input validation and filtering

Validating and filtering all input data is the first line of defense against black box attacks. By ensuring that input data conforms to the expected format and scope, the likelihood of malicious input can be reduced. This includes cleaning the input to weed out non-compliant content that could be used in an attack.

6.2.2 Adversarial sample training

In the field of AI, the robustness of a model can be enhanced by adversarial training. This approach involves the deliberate addition of adversarial samples to the training data, which teaches the model to recognize and correctly process these inputs that have been carefully designed to mislead the model.

6.2.3 Anomaly detection systems

Implementing a robust anomaly detection system can monitor and alert on unusual query patterns or behaviors, which are often signs of black box attacks. By monitoring system activity in real-time, potential attack attempts can be quickly recognized and responded to.

6.2.4 Using Deep Learning Defenses

Deep learning can be used to augment traditional anomaly detection systems, for example by training models to identify attack patterns or unusual behavior. These systems can be trained at multiple levels to recognize more sophisticated attack strategies.

In many practical applications, query-based black-box attacks pose a serious threat to deep learning systems. Zeyu Qin et al [97] investigated a lightweight defense method called Random Noise Defense (RND), which adds appropriate Gaussian noise to each query. This method of RND can better defend against query-based black-box attacks.

6.3 Methodology

Considering the wide application of deep learning models in medical image analysis and the increasing security requirements, we propose a novel defense strategy aimed at enhancing the defense capability of deep learning systems for medical images against black-box attacks, which is referred to as double noise defense. With a limited model query budget, it is more realistic to assess the vulnerability of medical image deep learning systems under decision-based attacks. We can prevent attackers from carrying out large-scale attacks by setting a limit on the number of queries allowed in a specific time, which makes the query efficiency of black-box attacks low. We propose a black-box defence method called Double Noise Defence (DND), which aims to enhance system security by reducing potential vulnerabilities and countering malicious attacks. The algorithm for DND is as follows:

```
Algorithm DND
Require: Classifier $C$, Input Sample $x$, Constant $m, n$
Ensure: Output Prediction $y$
1: Initialize
    Generate noise samples $\alpha, \beta$ from Standard Gaussian distribution $N(0, 1)$
2: Prepare Modified Input
        $x' \leftarrow x + m * \alpha$
3: Classify
        $y \leftarrow C(x')$
4: Adjust Prediction
        $y' \leftarrow y + n * \beta$
5: Output $y'$
```

6.4 Experiment

In this section, we will provide a detailed introduction to the experiment.

6.4.1 Built the medical images deep learning systems

（一）Dataset

In the present study, we utilized three distinct types of public medical image datasets: chest CT images, chest X-ray images, and breast cancer pathology images. These datasets were obtained from websites that provide public access to medical image resources. Comprehensive details regarding the acquisition and characteristics of these datasets are meticulously detailed in Section 3.3, Chapter 3 of this document. Ensuring confidentiality and ethical compliance, all datasets have been fully anonymized and are exclusively available for non-commercial medical research applications. We segmented the entire collection into two subsets: a training set and a test set. A detailed overview of these datasets is systematically outlined in Table 5-1.

（二）Deep learning model

When using deep learning in medical image analysis, the stability of the model and the accuracy of its predictions are critical. Therefore we often need to test the model extensively to ensure that it performs as expected in the real world. Black-box attack experiments are one form of such testing, which detects the model's sensitivity and robustness to external disturbances. DenseNet-121, an effective convolutional neural network architecture, has been widely used for medical image analysis and recognition. This model architecture is capable of achieving high accuracy in a low number of parameters through enhanced feature transfer and utilization, and is particularly suitable for the capture of details in image recognition tasks, which is especially important for medical image analysis, as these images often contain subtle and critical biomarkers.

In this study, we continue to use the DenseNet-121 model that has been trained and validated in previous black-box attack experiments. This is because this model has demonstrated good processing capabilities and high robustness to medical images, and can maintain high recognition accuracy in the face of various attacks and interference. By continuing to use this already trained model, we can more effectively compare the performance changes under different experimental configurations, and thus deeply analyze the model's performance in real applications and potential room for improvement. In addition, maintaining experimental consistency also means that the variability of experimental results can be reduced, thus providing more reliable data to support further analysis and decision-making.

6.4.2 Defense of the black box attacks on the medical images deep learning systems

To verify the effectiveness of the defense algorithms, we compared the defense algorithms with the no-defense and RND defense algorithms respectively, and to ensure the validity of the experimental results, we tested them with three different deep learning systems for medical images.

6.4.3 Metrics

In this chapter, *lp* distance is used to comprehensively assess the effectiveness of the defense black box. *lp* distance serves as a key metric to quantify the median distance between the original image and its perturbed version. Specifically, we focus on the l_2 distance, a metric that allows for an intuitive comparison of the amount of perturbation introduced by different algorithms. The smaller the *lp* distance, subject to the same number of queries, indicates that the algorithm is more efficient in launching successful attacks.

6.5 Results and Discussion

To substantiate the efficacy of our proposed black-box attack defense algorithm, a comparative analysis was conducted against both a no-defense scenario and the RND black-box attack defense algorithm. In addition, to comprehensively evaluate the algorithms' performance, we conducted tests utilizing three distinct medical image datasets. Our approach aimed to gauge the robustness and effectiveness of the proposed defense mechanism under various attack scenarios. By comparing its performance against both an unprotected setting and an existing defense algorithm, we sought to ascertain its superiority in mitigating adversarial threats in medical image processing tasks**.**

Tables 6-1 present the average and median l_2 distances resulting from the application of No defense, RDN defense, and DND defense algorithms on chest CT medical image datasets, each with varying query budgets. The analysis reveals that our proposed DND defense algorithm consistently yields the maximal l_2 distance perturbations when constrained by a limited number of queries, outperforming the RND black box defense algorithms. Our DND defense method demonstrates the maximal perturbations among the four algorithms when queried 1000 times. This highlights the fact that the DND defense algorithm significantly outperforms the RDN defense method at the beginning of the attack. In addition, as the number of queries increases, the mean and median values of the l_2 distances decrease, but the distances are still greater than the RDN defense algorithm, which indicates that the DND defense algorithm outperforms the other methods. These findings underscore the effectiveness of our proposed defense mechanism in reducing perturbation sizes, particularly when operating under limited query budgets. The DND defense algorithm allows black-box attack algorithms to generate larger perturbations to the model than other defense algorithms, which highlights its potential to enhance the robustness of medical image processing systems against adversarial attacks.

Table 6-1 The query counts and l_2 distances for the CT image deep learning system

CT images	**Defense**	**1K**	**5k**	**10k**	**15k**	**20k**
Average l_2 distance	No defense	29.11	13.84	7.19	4.98	4.56
	RND	32.95	30.48	27.79	23.99	21.95
	DND	37.82	36.72	34.22	33.93	33.83
Median l_2 distance	No defense	20.11	7.04	4.19	2.74	2.33
	RND	32.61	25.19	23.80	22.44	18.83
	DND	34.84	32.86	32.03	31.15	29.80

To better compare the effectiveness of two defense algorithms against GBBA black-box attacks on different medical image datasets, we plotted the experimental results as line graphs. Figure 6-1 shows the comparison of defense algorithms in a deep learning system based on COVID-19 medical images. In the figure, solid lines represent the mean l_2 distances, while dashed lines represent the median l_2 distances. It is evident from the figure that as the number of queries increases, the L2 distance decreases, indicating that increasing the number of queries can significantly reduce the gap between the adversarial image and the original image. Compared to the scenario without any defense, both RND and DND defense algorithms significantly increase the l_2 distance at the same number of queries. Specifically, as the number of queries increases, the l_2 distance under the DND defense algorithm consistently remains higher than that of the RND algorithm. Whether considering the mean l_2 distance or the median l_2 distance, DND exhibits larger values. This indicates that the

DND defense algorithm is superior to the RND algorithm.

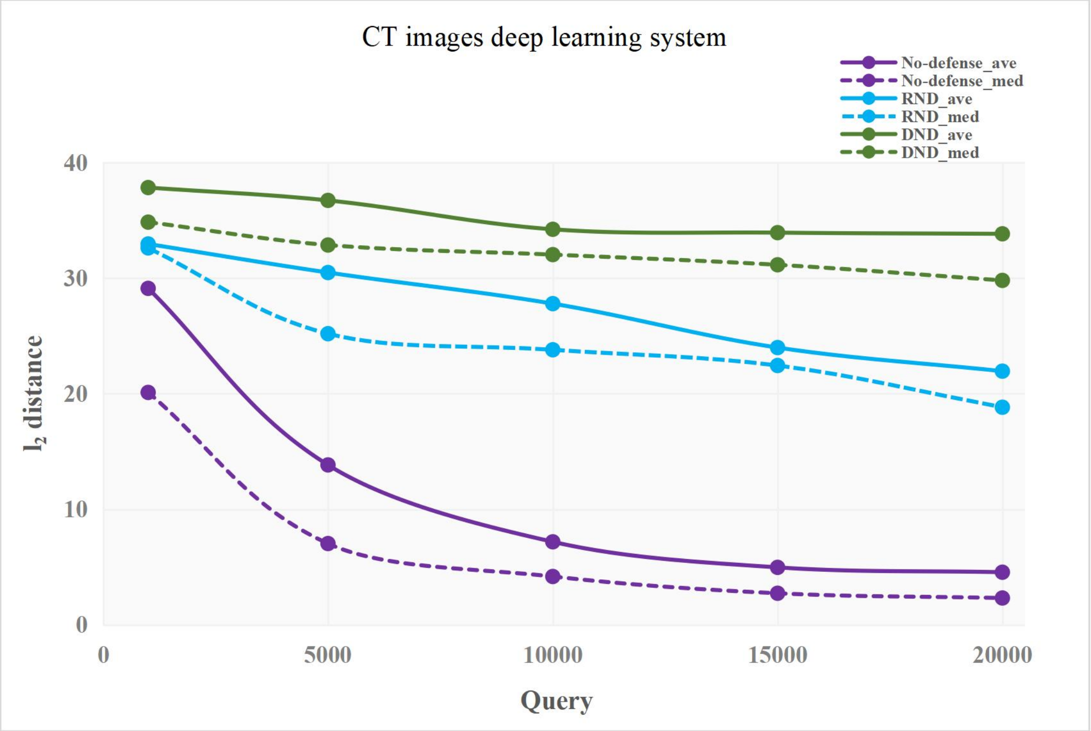


Figure 6-1 The plots of query counts and l_2 distances for CT images deep learning system

To better illustrate the differences between original and adversarial images under different defense methods, we first compared them under different numbers of queries using a deep learning system based on chest CT images. Figure 6-2 illustrates the comparison between the original and adversarial images generated by DND defense methods under different numbers of queries. It is clear from the figure that both no defense and defense with the increasing number of queries, the noise in the adversarial image decreases, making it more similar to the original image. Relative to the no-defense state, DND defenses have significant noise and differences between the resulting adversarial image and the original image with less number of queries. This indicates that the DND defense algorithm can better defend against black-box attacks with better defense performance.

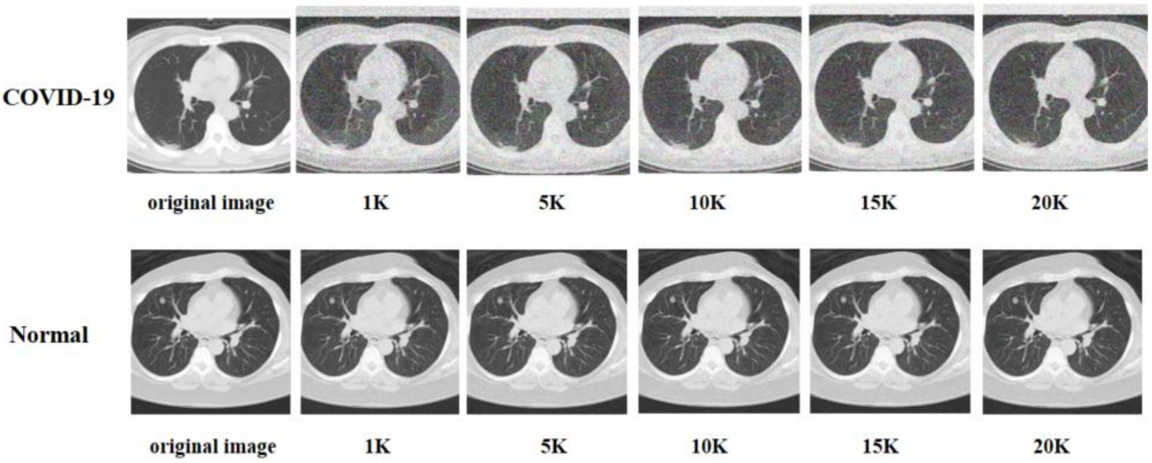


Figure 6-2 Comparison of adversarial images and original images of CT images under DND defense

As shown in Table 6-2, the performance of the no-defense and defense algorithms for the breast pathology image dataset is evaluated by calculating the mean and median of the l_2 distances. This study covers the baseline in the no-defense, RDN defense, and our proposed DND defense algorithm, each of which sets a different query budget. After comparative analysis, it is found that the DND defense algorithm can produce significantly larger l_2 distance perturbations than the other algorithms under the condition of a limited number of queries, indicating its superiority in defense effectiveness. In particular, when the number of queries reaches 1000, the DND defense method algorithm produces the largest perturbation among all compared algorithms, indicating that it can quickly outperform the RDN defense strategy at the early stage of the attack. Furthermore, despite the decrease in the mean and median values of the l_2 distance as the number of queries increases, the DND defense algorithm maintains a large perturbation distance, consistently outperforming the RDN defense algorithm.

Table 6-2 The query counts and l_2 distances for the breast cancer image deep learning system

Breast cancer images	**Defense**	**1K**	**5k**	**10k**	**15k**	**20k**
Average l_2 distance	No defense	15.78	6.91	4.55	3.53	2.92
	RND	28.98	24.43	21.52	19.81	18.70
	DND	31.94	30.35	29.78	29.64	29.01
Median l_2 distance	No defense	12.13	5.18	3.57	2.77	2.28
	RND	26.36	23.16	19.99	18.36	17.06
	DND	30.39	28.40	27.99	29.14	27.65

Figure 6-3 shows a comparison of defense algorithms in a deep learning system based on breast medical images. The solid line in the figure indicates the mean l_2 distance and the dashed line indicates the median l_2 distance. Again we can see from the figure that the l_2 distance tends to decrease as the number of queries increases. This indicates that increasing the number of queries can significantly reduce the difference between the adversarial image and the original image. Compared with the no-defense scenario, both RND and DND defense algorithms significantly improve the l_2 distance under the same number of queries. As the number of queries increases, the l_2 distance under the DND defense algorithm is always higher than the l_2 distance under the RND algorithm. Regardless of whether the mean l_2 distance or the median l_2 distance is considered, the value of DND is larger. This clearly shows that the DND defense algorithm is superior to the RND algorithm. These results suggest that the DND strategy is the most effective in reducing the l_2 distance, particularly under a high number of queries, thereby providing better protection for deep learning systems against adversarial attacks.

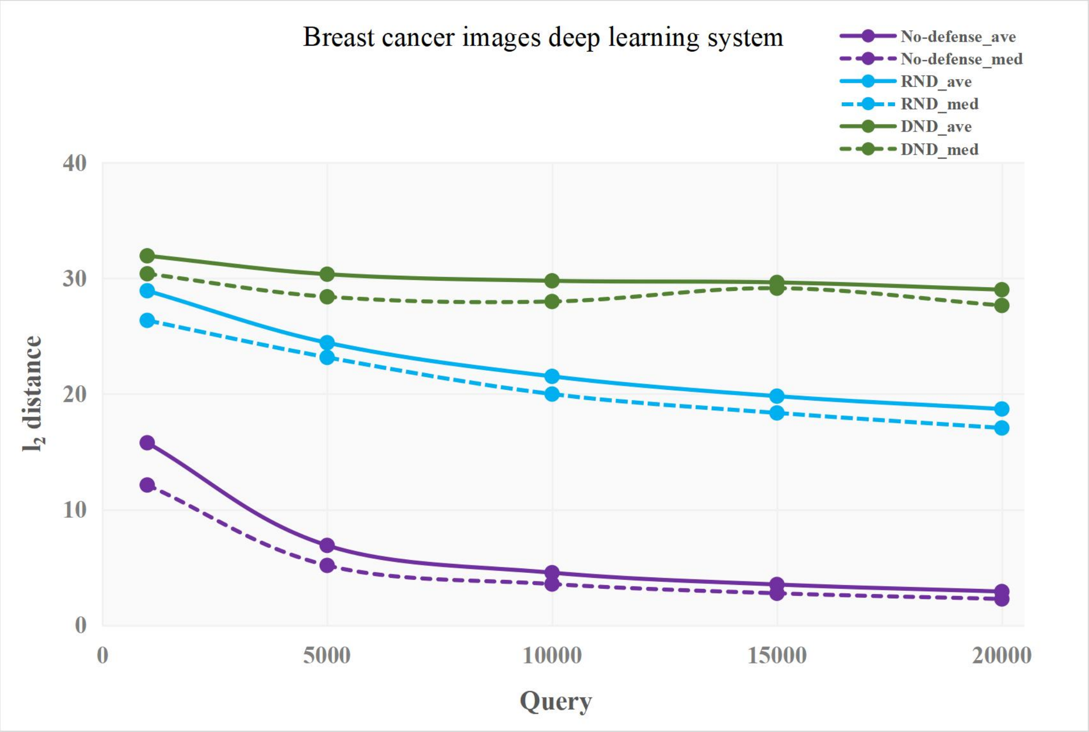


Figure 6-3 The plots of query counts and l_2 distances for breast cancer image deep learning system

To better illustrate the differences between original and adversarial images under the DND defense method, we analyzed them based on breast medical images deep learning system and varying the number of queries. Figure 6-4 shows a comparison between the original and adversarial images generated with DND defense methods across different numbers of queries. The figure clearly shows that as the number of queries increases, the noise reduction in the adversarial image is very little and there is a significant difference compared to the original image. Compared to the no-defense condition, DND defense methods show significant noise and differences between the adversarial and original images with fewer queries. This suggests that the DND defense algorithm provides better performance in defending against black-box attacks.

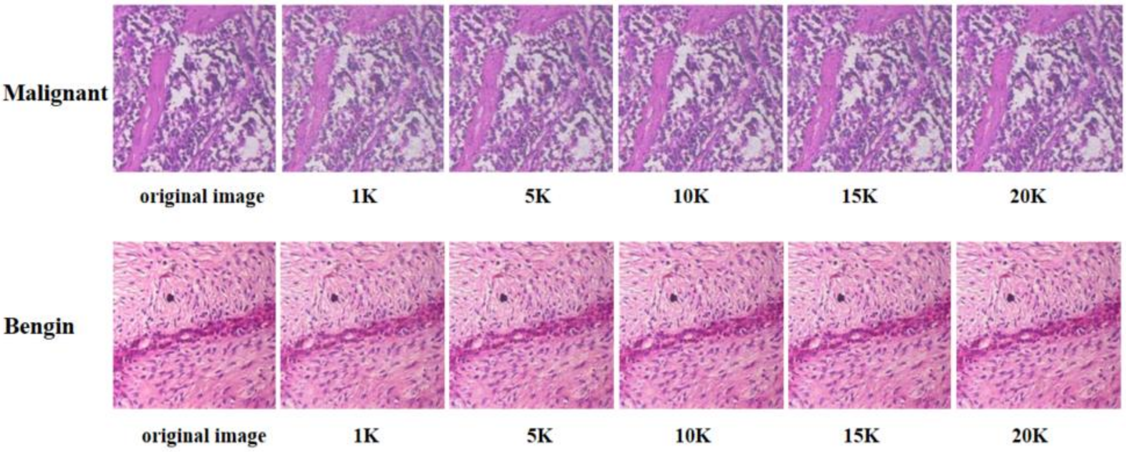


Figure 6-4 Comparison of adversarial images and original images of breast cancer images under DND defense

In addition, we evaluate the performance of several defense algorithms against chest X-ray medical imaging datasets, including the no-defense state, the RDN defense algorithm, and our newly proposed DND defense method. We analyze the efficacy of these algorithms by measuring the mean and median values of the l_2 norm distance. The results are summarized in Table 6 and broken down by different query budgets.

Our analysis shows that the DND defense approach produces the largest l_2 distance perturbation in scenarios where the number of queries is limited, significantly outperforming the RDN defense strategy. Specifically, the DND defense algorithm produces the maximum perturbation among all comparisons when the number of queries increases to 1000, proving its significant advantage in attack efficiency. Even when the number of queries increases, the mean and median l_2 distances induced by the DND defense decrease, but are still higher than the RDN defense, consistently demonstrating superior defense.

Table 6-3 The query counts and l_2 distances for X-ray image deep learning system

X-ray images	**Defense**	**1K**	**5k**	**10k**	**15k**	**20k**
Average l_2 distance	No defense	29.35	11.05	4.74	2.95	2.23
	RND	39.09	36.47	33.69	31.07	29.27
	DND	39.83	39.59	39.35	39.17	38.98
Median l_2 distance	No defense	29.55	10.39	4.49	2.91	2.22
	RND	38.73	36.60	33.78	30.99	29.30
	DND	39.51	39.27	39.29	38.81	38.73

Figure 6-5 illustrates the l_2 distance variation under different query numbers for three defense strategies: No-defense, RND defense, and DND defense. The analysis is conducted for both average (ave) and median (med) values. The x-axis represents the number of queries, while the y-axis represents the l_2 distance. From the figure, it is evident that the No-defense strategy maintains a low l_2 distance consistently, with negligible variation as the number of queries increases. In contrast, the RND strategy shows a gradual decrease in l_2 distance with an increasing number of queries, and the rate of decrease is relatively slow. The DND strategy demonstrates a fewer reduction in l_2 distance, maintaining high levels as the number of queries increases, indicating a more effective defense mechanism. Specifically, in the no-defense strategy, both the mean l_2 distance and the median l_2 distance are significantly reduced, while the median l_2 distance decreases to around 10 after 20,000 queries and stabilizes at a lower level as the number of queries increases. Under the RND strategy, the average l_2 distance decreases from nearly 40 to approximately 30. The DND strategy performs the best, with both the average and median l_2 distances remain around the 40 mark, stabilizing at high levels as the query number increases.

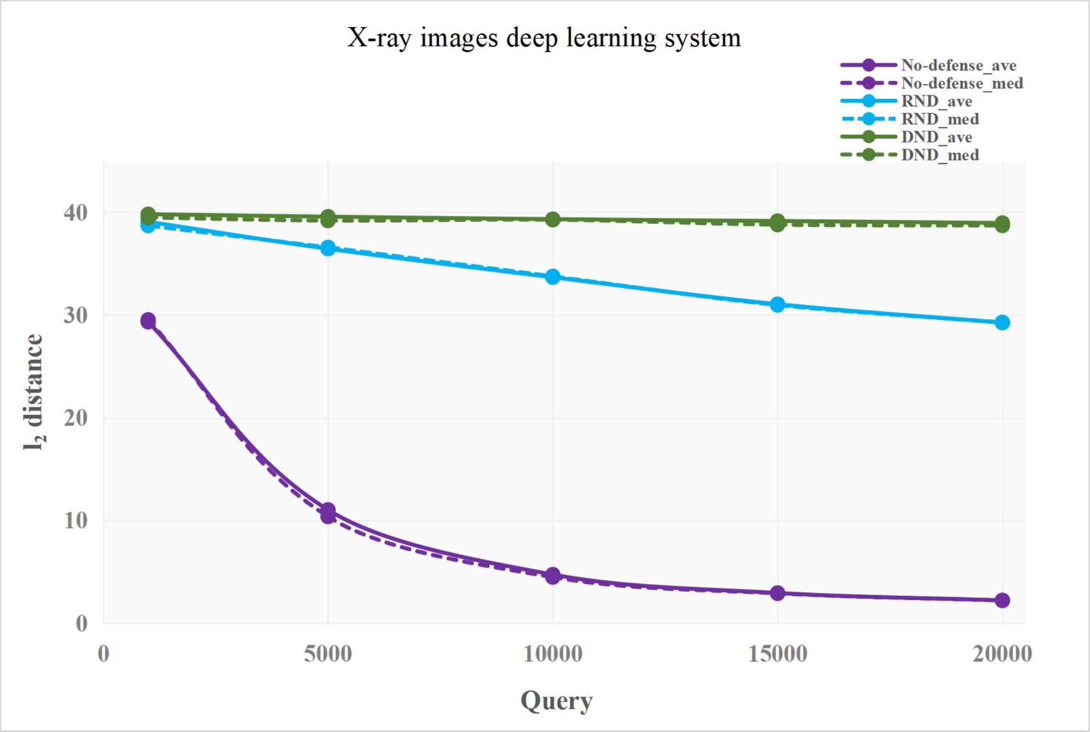


Figure 6-5 The plots of query counts and l_2 distances for X-ray image deep learning system

To better illustrate the differences between original and adversarial images under DND defense mechanisms, we conducted a comparison using a deep learning system based on chest X-ray images, varying the number of queries. Figure 6-6 shows the comparison between original and adversarial images generated with DND defense, evaluated across different query numbers. The figure indicates that as the number of queries increases, the noise in the adversarial images decreases, making them more similar to the original images, regardless of whether a defense is applied. However, with fewer queries, DND defense methods show significant noise and deviations from the original images compared to the no-defense condition. It suggests that the DND defense algorithm is more effective in mitigating the impact of black-box attacks, providing superior defense performance.

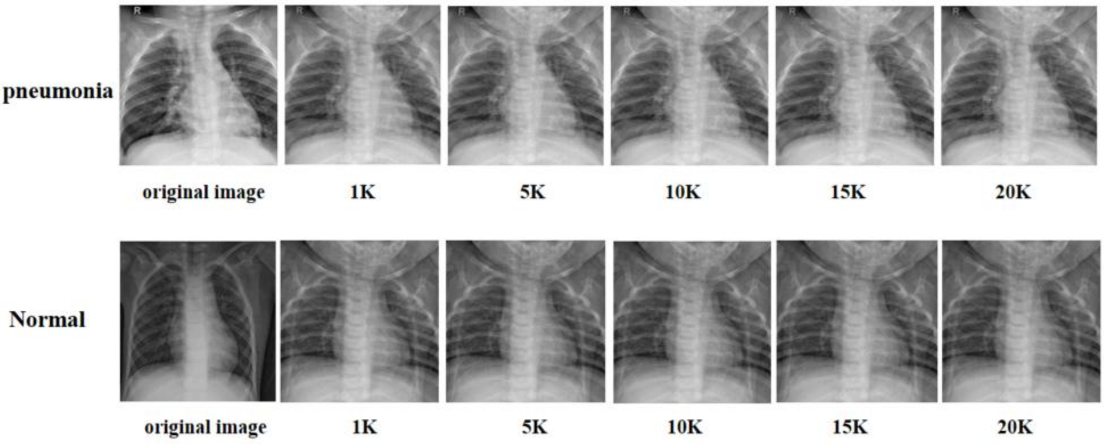


Figure 6-6 Comparison of adversarial images and original images of X-ray images under DND defense

These results emphasize the effectiveness of our proposed defense mechanism in controlling the perturbation size, especially in environments with limited query budgets. The DND defense algorithm allows black-box attack algorithms to apply larger perturbations to the model than other defense strategies, thus enhancing the robustness of medical image processing systems against adversarial attacks. This point highlights the advantages of the DND defense strategy in enhancing the model's ability to withstand black-box attacks.

6.6 Summary

In this study, we propose a lightweight defense scheme named Double Noise Defense (DND) for black box adversarial attacks on medical image deep learning systems. Compared to traditional defense methods such as RND and undefended medical image deep systems, our approach significantly reduces the success rate of query-based black box adversarial attacks.

Our approach provides a lightweight, practical, and effective means to harden deep learning systems against the threat of Black Brother attacks. Through experimental testing and evaluation, we demonstrate the effectiveness of DND in hindering the performance of query-based black-box attacks, thereby enhancing the robustness of deep learning models in medical image analysis tasks. Furthermore, comparative analysis with existing defense methods highlights the superiority of DND in protecting medical image processing systems from malicious attacks. This defense approach provides a strong guarantee for enhancing the security and reliability of medical image deep learning systems.

7 Conclusion and Future work

7.1 Conclusion

The integration of deep learning into medical diagnostics undeniably holds vast potential, with artificial intelligence (AI) technologies significantly accelerating advancements in medicine and healthcare. Their contributions to the rapid progression of medical care are substantial. However, the security and reliability of deep learning systems, particularly in the medical domain where human health is at stake, are matters of paramount concern.

In this paper, we take the deep neural learning system based on medical images as the research object, take the vulnerability of deep learning system as the research background, and mainly test the security and robustness of the deep system of medical images under the scenarios of white-box attack and black-box attack. Meanwhile, corresponding adversarial defense methods are proposed for white-box attack and black-box attack respectively.

Overall, white box attacks are useful in theoretical research and model testing because they can reveal specific weaknesses of a model. Black-box attacks, on the other hand, are more suitable for evaluating the security of a model in real-world applications because they do not rely on knowledge of the model's internals.

In studying the security of deep learning systems, it is important to understand the strengths and weaknesses of these attack methods to better design and implement defense strategies for models.

Undoubtedly, there remains a need for further research and development to effectively apply the medical image deep learning system in real-world scenarios. Adapting the model to real-world environments will require addressing various challenges and potential threats. Continual refinement and exploration of novel defensive mechanisms will be necessary to fortify the deep learning model’s defenses against increasingly sophisticated adversarial attacks. In future work, we will continue to study the security and reliability of the medical deep learning system. We also hope that more researchers will pay attention to the security of medical deep learning systems.

7.2 Future work

（1）Enhancing the transferability of adversarial samples

The transferability of adversarial examples refers to the effectiveness of adversarial perturbations across multiple models or datasets, proving the generalizability of adversarial attacks. In practical scenarios, especially in opaque black-box environments, it is crucial to enhance the transferability of adversarial samples. In this case, the structure and parameters of the model are unknown to the attacker. Therefore, developing adversarial samples that are broadly applicable to a wide range of models can maintain a high attack success rate without detailed knowledge of the target model.

（2）Enhancing the defense capabilities in medical image deep learning system

The defensive capability of a model is its ability to maintain accuracy and functionality in the face of various types of interference, especially interference involving adversarial attacks. Models with strong general defenses are more likely to perform well on unseen data or tasks and thus show better generalization. This is useful in real-world scenarios where the data is often very different from what was used during training. Only when models are safer and more reliable across a range of conditions are they more likely to be trusted and adopted by users and the industry. Trust is a key factor in the widespread adoption of AI technologies. By improving the general defense, the model is made more robust and reliable under various inputs and conditions. This is critical for healthcare deep learning systems where unexpected or malicious inputs can lead to erroneous outputs and potentially harmful results.

Bibliography

1. Dawson, J. W. (2007). The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Ingelligence, and Artificial Life plus The Secrets of Enigma, by Alan M. Turing (author) and B. Jack Copeland (editor).

2. Rajaraman, V. (2014). JohnMcCarthy—Father of artificial intelligence. Resonance, 19(3), 198–207. https://doi.org/10.1007/s12045-014-0027-9

3. Ertel, W. (2018). Introduction to Artificial Intelligence. Springer.

4. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

5. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

6. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

7. Pak, M., & Kim, S. (2017). A review of deep learning in image recognition. 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), 1–3. https://doi.org/10.1109/CAIPT.2017.8320684

8. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks (arXiv:1312.6199). arXiv. https://doi.org/10.48550/arXiv.1312.6199

9. Wu, Y., & Feng, J. (2018). Development and Application of Artificial Neural Network. Wireless Personal Communications, 102(2), 1645–1656. https://doi.org/10.1007/s11277-017-5224-x

10. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408. https://doi.org/10.1037/h0042519

11. Taud, H., & Mas, J. F. (2018). Multilayer Perceptron (MLP). M. T. Camacho Olmedo, M. Paegelow, J.-F. Mas, & F. Escobar, Geomatic Approaches for Modeling Land Change Scenarios (451–455). Springer International Publishing.

https://doi.org/10.1007/978-3-319-60801-3_27

12. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and Learning Systems, 33(12), 6999–7019. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2021.3084827

13. Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. IEEE Transactions on Neural Networks and Learning Systems, 33(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

14. Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. 2017 international conference on engineering and technology (ICET), 1–6. https://ieeexplore.ieee.org/abstract/document/8308186/

15. Gholamalinezhad, H., & Khosravi, H. (2020). Pooling Methods in Deep Neural Networks, a Review (arXiv:2009.07485). arXiv. https://doi.org/10.48550/arXiv.2009.07485

16. Sharma, S., Sharma, S., & Athaiya, A. (2020). ACTIVATION FUNCTIONS IN NEURAL NETWORKS. International Journal of Engineering Applied Sciences and Technology, 04(12), 310–316. https://doi.org/10.33564/IJEAST.2020.v04i12.054

17. Han, J., & Moraga, C. (1995). The influence of the sigmoid function parameters on the speed of backpropagation learning. J. Mira & F. Sandoval, From Natural to Artificial Neural Computation (195–201). Springer. https://doi.org/10.1007/3-540-59497-3_175

18. He, J., Li, L., Xu, J., & Zheng, C. (2020). ReLU Deep Neural Networks and Linear Finite Elements. Journal of Computational Mathematics, 38(3), 502–527. https://doi.org/10.4208/jcm.1901-m2018-0160

19. Xu, J., Li, Z., Du, B., Zhang, M., & Liu, J. (2020). Reluplex made more practical: Leaky ReLU. 2020 IEEE Symposium on Computers and communications (ISCC), 1–7. https://ieeexplore.ieee.org/abstract/document/9219587/

20. Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2022). Activation functions in deep learning: A comprehensive survey and benchmark. Neurocomputing, 503, 92–108. https://doi.org/10.1016/j.neucom.2022.06.111

21. Dunne, R. A., & Campbell, N. A. (1997). On The Pairing Of The Softmax Activation And Cross-Entropy Penalty Functions And The Derivation Of The

Softmax Activation Function.

22. Akhtar, N., & Mian, A. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. IEEE Access, 6, 14410–14430. IEEE Access. https://doi.org/10.1109/ACCESS.2018.2807385

23. Kevles, B. (1997). Naked to the Bone: Medical Imaging in the Twentieth Century. Rutgers University Press.

24. Duncan, J. S., & Ayache, N. (2000). Medical image analysis: Progress over two decades and the challenges ahead. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), 85–106. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.824822

25. Dhawan, A. P. (2011). Medical Image Analysis. John Wiley & Sons.

26. Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., & Summers, R. M. (2021). A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. Proceedings of the IEEE, 109(5), 820–838. Proceedings of the IEEE. https://doi.org/10.1109/JPROC.2021.3054390

27. Ou, X., Chen, X., Xu, X., Xie, L., Chen, X., Hong, Z., Bai, H., Liu, X., Chen, Q., Li, L., & Yang, H. (2021). Recent Development in X-Ray Imaging Technology: Future and Challenges. Research, 2021, 2021/9892152. https://doi.org/10.34133/2021/9892152

28. Buzug, T. M. (2011). Computed Tomography. R. Kramme, K.-P. Hoffmann, & R. S. Pozos, Springer Handbook of Medical Technology (311–342). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74658-4_16

29. Plewes, D. B., & Kucharczyk, W. (2012). Physics of MRI: A primer. Journal of Magnetic Resonance Imaging, 35(5), 1038–1054. https://doi.org/10.1002/jmri.23642

30. Chan, V., & Perlas, A. (2011). Basics of Ultrasound Imaging. S. N. Narouze, Atlas of Ultrasound-Guided Procedures in Interventional Pain Management (13–19). Springer. https://doi.org/10.1007/978-1-4419-1681-5_2

31. Jeans, W. D. (1990). The development and use of digital subtraction angiography. British Journal of Radiology, 63(747), 161–168. https://doi.org/10.1259/0007-1285-63-747-161

32. Zhang, R., Li, W., & Mo, T. (2018). Review of Deep Learning (arXiv:1804.01653). arXiv. https://doi.org/10.48550/arXiv.1804.01653

33. Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857

34. Sharma, A. K., Nandal, A., Dhaka, A., & Dixit, R. (2021). Medical Image Classification Techniques and Analysis Using Deep Learning Networks: A Review. 收入 R. Patgiri, A. Biswas, & P. Roy, Health Informatics: A Computational Perspective in Healthcare (233–258). Springer. https://doi.org/10.1007/978-981-15-9735-0_13

35. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., & Nandi, A. K. (2022). Medical image segmentation using deep learning: A survey. IET Image Processing, 16(5), 1243–1267. https://doi.org/10.1049/ipr2.12419

36. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

37. Hu, K., Su, Y., Wang, J., & Xu, Y. (2021). A review of COVID-19: A summary of the epidemic in Wuhan and other local areas in China. E3S Web of Conferences, 292, 03099.

38. Shi, Y., Wang, G., Cai, X.-P., Deng, J.-W., Zheng, L., Zhu, H.-H., Zheng, M., Yang, B., & Chen, Z. (2020). An overview of COVID-19. Journal of Zhejiang University. Science. B, 21(5), 343–360.

39. Wan, S., Xiang, Y., Fang, W., Zheng, Y., Li, B., Hu, Y., Lang, C., Huang, D., Sun, Q., Xiong, Y., Huang, X., Lv, J., Luo, Y., Shen, L., Yang, H., Huang, G., & Yang, R. (2020). Clinical features and treatment of COVID-19 patients in northeast Chongqing. Journal of Medical Virology, 92(7), 797–806. https://doi.org/10.1002/jmv.25783

40. Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W.-C., Wang, C.-B., & Bernardini, S. (2020). The COVID-19 pandemic. Critical Reviews in Clinical Laboratory Sciences, 57(6), 365–388. https://doi.org/10.1080/10408363.2020.1783198

41. Mizrahi, B., Shilo, S., Rossman, H., Kalkstein, N., Marcus, K., Barer, Y., Keshet, A., Shamir-Stein, N., Shalev, V., & Zohar, A. E. (2020). Longitudinal symptom dynamics of COVID-19 infection. Nature communications, 11(1), 6208.

42. Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., & Rothe, C. (2020). Virological assessment of hospitalized patients with COVID-2019. Nature, 581(7809), 465–469.

43. Fernandes, N. (2020). Economic effects of coronavirus outbreak (COVID-19) on the world economy.

44. Budd, J., Miller, B. S., Manning, E. M., Lampos, V., Zhuang, M., Edelstein, M., Rees, G., Emery, V. C., Stevens, M. M., & Keegan, N. (2020). Digital technologies in the public-health response to COVID-19. Nature medicine, 26(8), 1183–1192.

45. Gorbalenya, A. E., Baker, S. C., Baric, R. S., de Groot, R. J., Drosten, C., Gulyaeva, A. A., Haagmans, B. L., Lauber, C., Leontovich, A. M., Neuman, B. W., Penzar, D., Perlman, S., Poon, L. L. M., Samborskiy, D. V., Sidorov, I. A., Sola, I., Ziebuhr, J., & Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. (2020). The species Severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology, 5(4), 536–544. https://doi.org/10.1038/s41564-020-0695-z

46. Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., & Xia, L. (2020). Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology, 296(2), E32–E40. https://doi.org/10.1148/radiol.2020200642

47. Lee, E. Y. P., Ng, M.-Y., & Khong, P.-L. (2020). COVID-19 pneumonia: What has CT taught us? The Lancet Infectious Diseases, 20(4), 384–385. https://doi.org/10.1016/S1473-3099(20)30134-1

48. Wan, S., Li, M., Ye, Z., Yang, C., Cai, Q., Duan, S., & Song, B. (2020). CT Manifestations and Clinical Characteristics of 1115 Patients with Coronavirus Disease 2019 (COVID-19): A Systematic Review and Meta-analysis. Academic Radiology, 27(7), 910–921. https://doi.org/10.1016/j.acra.2020.04.033

49. Bao, C., Liu, X., Zhang, H., Li, Y., & Liu, J. (2020). Coronavirus Disease 2019 (COVID-19) CT Findings: A Systematic Review and Meta-analysis. Journal of the American College of Radiology, 17(6), 701–709. https://doi.org/10.1016/j.jacr.2020.03.006

50. Hani, C., Trieu, N. H., Saab, I., Dangeard, S., Bennani, S., Chassagnon, G., & Revel, M.-P. (2020). COVID-19 pneumonia: A review of typical CT findings and differential diagnosis. Diagnostic and Interventional Imaging, 101(5), 263–268.

https://doi.org/10.1016/j.diii.2020.03.014

51. Zhu, J., Zhong, Z., Li, H., Ji, P., Pang, J., Li, B., & Zhang, J. (2020). CT imaging features of 4121 patients with COVID-19: A meta-analysis. Journal of Medical Virology, 92(7), 891–902. https://doi.org/10.1002/jmv.25910

52. Ferlay, J., Colombet, M., Soerjomataram, I., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2021). Cancer statistics for the year 2020: An overview. International Journal of Cancer. https://doi.org/10.1002/ijc.33588

53. Sharma, G. N., Dave, R., Sanadya, J., Sharma, P., & Sharma, K. K. (2010). Various types and management of breast cancer: An overview. Journal of Advanced Pharmaceutical Technology & Research, 1(2), 109–126.

54. Akram, M., Iqbal, M., Daniyal, M., & Khan, A. U. (2017). Awareness and current knowledge of breast cancer. Biological Research, 50(1), 33. https://doi.org/10.1186/s40659-017-0140-9

55. Benson, J. R., Jatoi, I., Keisch, M., Esteva, F. J., Makris, A., & Jordan, V. C. (2009). Early breast cancer. The Lancet, 373(9673), 1463–1479. https://doi.org/10.1016/S0140-6736(09)60316-0

56. McDonald, E. S., Clark, A. S., Tchou, J., Zhang, P., & Freedman, G. M. (2016). Clinical Diagnosis and Management of Breast Cancer. Journal of Nuclear Medicine, 57(Supplement 1), 9S-16S. https://doi.org/10.2967/jnumed.115.157834

57. Valdora, F., Houssami, N., Rossi, F., Calabrese, M., & Tagliafico, A. S. (2018). Rapid review: Radiomics and breast cancer. Breast Cancer Research and Treatment, 169(2), 217–229. https://doi.org/10.1007/s10549-018-4675-4

58. Crivelli, P., Ledda, R. E., Parascandolo, N., Fara, A., Soro, D., & Conti, M. (2018). A New Challenge for Radiologists: Radiomics in Breast Cancer. BioMed Research International, 2018, e6120703. https://doi.org/10.1155/2018/6120703

59. Conti, A., Duggento, A., Indovina, I., Guerrisi, M., & Toschi, N. (2021). Radiomics in breast cancer classification and prediction. Seminars in Cancer Biology, 72, 238–250. https://doi.org/10.1016/j.semcancer.2020.04.002

60. Li, H., Zhu, Y., Burnside, E. S., Drukker, K., Hoadley, K. A., Fan, C., Conzen, S. D., Whitman, G. J., Sutton, E. J., Net, J. M., Ganott, M., Huang, E., Morris, E. A., Perou, C. M., Ji, Y., & Giger, M. L. (2016). MR Imaging Radiomics Signatures for Predicting the Risk of Breast Cancer Recurrence as Given by Research Versions of MammaPrint, Oncotype DX, and PAM50 Gene Assays. Radiology, 281(2), 382–391. https://doi.org/10.1148/radiol.2016152110

61. Ruuskanen, O., Lahti, E., Jennings, L. C., & Murdoch, D. R. (2011). Viral pneumonia. The Lancet, 377(9773), 1264–1275. https://doi.org/10.1016/S0140-6736(10)61459-6

62. Quinton, L. J., Walkey, A. J., & Mizgerd, J. P. (2018). Integrative Physiology of Pneumonia. Physiological Reviews, 98(3), 1417–1464. https://doi.org/10.1152/physrev.00032.2017

63. Grief, S. N., & Loza, J. K. (2018). Guidelines for the Evaluation and Treatment of Pneumonia. Primary Care: Clinics in Office Practice, 45(3), 485–503. https://doi.org/10.1016/j.pop.2018.04.001

64. Parveen, N., & Sathik, M. M. (2011). Detection of pneumonia in chest X-ray images. Journal of X-ray Science and Technology, 19(4), 423–428.

65. Yanase, J., & Triantaphyllou, E. (2019). A systematic survey of computer-aided diagnosis in medicine: Past and present developments. Expert Systems with Applications, 138, 112821. https://doi.org/10.1016/j.eswa.2019.112821

66. Jaiswal, A. K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A., & Rodrigues, J. J. P. C. (2019). Identifying pneumonia in chest X-rays: A deep learning approach. Measurement, 145, 511–518. https://doi.org/10.1016/j.measurement.2019.05.076

67. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

68. Zhang, R., Li, W., & Mo, T. (2018). Review of Deep Learning (arXiv:1804.01653). arXiv. https://doi.org/10.48550/arXiv.1804.01653

69. Kamath, U., Liu, J., & Whitaker, J. (2019). Deep Learning for NLP and Speech Recognition. Springer International Publishing. https://doi.org/10.1007/978-3-030-14596-5

70. Fourcade, A., & Khonsari, R. H. (2019). Deep learning in medical image analysis: A third eye for doctors. Journal of Stomatology, Oral and Maxillofacial Surgery, 120(4), 279–288. https://doi.org/10.1016/j.jormas.2019.06.002

71. Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep Learning for Medical Image Processing: Overview, Challenges and the Future. N. Dey, A. S. Ashour, & S. Borra, Classification in BioApps: Automation of Decision Making (323–350). Springer International Publishing. https://doi.org/10.1007/978-3-319-65981-7_12

72. Akhtar, N., & Mian, A. (2018). Threat of Adversarial Attacks on Deep

Learning in Computer Vision: A Survey. IEEE Access, 6, 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385

73. Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. CAAI Transactions on Intelligence Technology, 6(1), 25–45. https://doi.org/10.1049/cit2.12028

74. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples (arXiv:1412.6572). arXiv. http://arxiv.org/abs/1412.6572

75. Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236. https://arxiv.org/abs/1611.01236

76. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083. https://arxiv.org/abs/1706.06083

77. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting adversarial attacks with momentum. Proceedings of the IEEE conference on computer vision and pattern recognition, 9185–9193.

78. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.

79. Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M., & Song, D. (2019). Generating Adversarial Examples with Adversarial Networks (arXiv:1801.02610). arXiv. http://arxiv.org/abs/1801.02610

80. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2015). The Limitations of Deep Learning in Adversarial Settings (arXiv:1511.07528). arXiv. https://doi.org/10.48550/arXiv.1511.07528

81. Saputro, D. R. S., & Widyaningsih, P. (2017). Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR). 040009. https://doi.org/10.1063/1.4995124

82. Jandial, S., Mangla, P., Varshney, S., & Balasubramanian, V. (2019). Advgan++: Harnessing latent layers for adversary generation. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 0–0.

83. Deb, D., Zhang, J., & Jain, A. K. (2020). Advfaces: Adversarial face synthesis. 2020 IEEE International Joint Conference on Biometrics (IJCB), 1–10.

https://ieeexplore.ieee.org/abstract/document/9304898/

84. Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., & Xie, P. (2020). COVID-CT-Dataset: A CT Scan Dataset about COVID-19 (arXiv:2003.13865). arXiv. http://arxiv.org/abs/2003.13865

85. Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016). A Dataset for Breast Cancer Histopathological Image Classification. IEEE Transactions on Biomedical Engineering, 63(7), 1455–1462. IEEE Transactions on Biomedical Engineering. https://doi.org/10.1109/TBME.2015.2496264

86. Kermany, D., Zhang, K., & Goldbaum, M. (2018). Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley data, 2(2), 651.

87. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

88. Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. 4700–4708.

89. Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization (arXiv:1412.6980). arXiv. http://arxiv.org/abs/1412.6980

90. Zhou, Z.-H. (2021). Machine Learning. Springer Nature.

91. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

92. Brendel, W., Rauber, J., & Bethge, M. (2018). Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models (arXiv:1712.04248). arXiv. http://arxiv.org/abs/1712.04248

93. Cheng, M., Le, T., Chen, P.-Y., Yi, J., Zhang, H., & Hsieh, C.-J. (2018). Query-Efficient Hard-label Black-box Attack:An Optimization-based Approach (arXiv:1807.04457). arXiv. http://arxiv.org/abs/1807.04457

94. Chen, J., Jordan, M. I., & Wainwright, M. J. (2020). Hopskipjumpattack: A query-efficient decision-based attack. 2020 ieee symposium on security and privacy (sp), 1277–1294. https://ieeexplore.ieee.org/abstract/document/9152788/

95. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). Gradient Surgery for Multi-Task Learning (arXiv:2001.06782). arXiv. https://doi.org/10.48550/arXiv.2001.06782

96. Danielsson, P.-E. (1980). Euclidean distance mapping. Computer Graphics

and Image Processing, 14(3), 227–248. https://doi.org/10.1016/0146-664X(80)90054-4

97. Qin, Z., Fan, Y., Zha, H., & Wu, B. (2021). Random Noise Defense Against Query-Based Black-Box Attacks. Advances in Neural Information Processing Systems, 34, 7650–7663.

Publication List of the Author

I. SCI Journal Papers

[1] Yang Li and Shaoying Liu. “The Threat of Adversarial Attack on a COVID-19 CT Image-Based Deep Learning System”. Bioengineering, 2023, 10(2): 194.

[2] Yang Li and Shaoying Liu. “Adversarial Attack and Defense in Breast Cancer Deep Learning Systems”. Bioengineering, 2023, 10(8): 973.

[3] Dingbang Fang, Shaoying Liu, and Yang Li. “Cross-Project Transfer Learning on Lightweight Code Semantic Graphs for Defect Prediction”. International Journal of Software Engineering and Knowledge Engineering, 33(7), 2023: 1095-1117.

II. International Conference Papers

[4] Yang Li, Shaoying Liu, and Lisen Guo. “Testing the Security of Deep Learning Systems Based on Chest X-ray Images”. The 2024 13th International Conference on Software and Computer Applications (ICSCA’24). 2024, 198 – 203.

[5] Yang Li and Shaoying Liu. “Testing and Verifying the Security of COVID-19 CT Images Deep Learning System with Adversarial Attack”, International Workshop on Structured Object-Oriented Formal Language and Method. Cham: Springer LNCS 13854, 2022: 119-125.

Acknowledgments

As I complete my doctoral thesis, I wish to express my deepest gratitude to those who have supported and encouraged me throughout this journey.

First and foremost, I must thank my parents. Your love, support, and encouragement have been the solid foundation on which I stand today. When faced with challenges and difficulties, you were always there to give me strength and courage, reassuring me that no matter what obstacles come my way, I have unconditional support. Your sacrifices and relentless efforts have provided me with the opportunity to pursue my dreams, for which I will be forever grateful.

Secondly, I would like to wholeheartedly thank my supervisor Shaoying Liu. You have provided invaluable guidance academically and offered insights into life's paths. Your patience, wisdom, and passion have profoundly impacted me, deepening my understanding and love for my field of study. Your support and encouragement have been indispensable in the completion of this thesis.

Finally, I would like to thank myself. Throughout this extraordinary journey, faced with various challenges and pressures, I chose to persevere and work hard instead of giving up. These experiences have made me a more resilient and confident person. I am proud of the effort I have put in and the achievements I have made.