IPTEK The Journal of Technology and Science, 35(2), 2024 (e/pISSN:2088-2033, 0853-4098)

DOI: 10.12962/j20882033.v35i2.19630 | Received 28 December, 2023; Revised 28 December, 2023; Accepted 11 June, 2024

ORIGINAL RESEARCH

ADVERSARIAL TRAINING FOR ROBUST DEFENSE IN CNN MODELS FOR LUNG AND COLON HISTOPATHOLOGICAL IMAGES

Chilyatun Nisa[1] | Nanik Suciati*[2] | Anny Yuniarti[3]

[1][1] Departement of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia, Email: nchilyatun@gmail.com

[2][1] Departement of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia, Email: nanik@if.its.ac.id

[3][1] Departement of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia, Email: anny@if.its.ac.id

Correspondence

*Email: nanik@if.its.ac.id

Present Address

Gedung Teknik Informatika, Jl. Teknik Kimia, Surabaya 60111, Indonesia

Abstract

Cancer stands as the world's second-leading cause of death, arising from abnormal cell growth that invades the body's cells and tissues. Simultaneous occurrences of lung and colon cancer are not uncommon, with lung cancer often emerging as the second primary cancer in colon cancer patients. While Deep Learning (DL) approaches have shown promise in accurate cancer classification, recent studies highlight the susceptibility of DL models to perturbations in input images. Merely achieving accuracy is insufficient; models must demonstrate resilience against even the slightest perturbations by applying adversarial defence methods. This study aims to enhance the reliability of the Convolutional Neural Network (CNN) algorithm in the face of adversarial attacks by implementing adversarial training. Leveraging the LC25000 dataset and various pre-trained CNN models for classification, we employ adversarial attack methods such as Carlini and Wagner, DeepFool, and SaliencyMap alongside adversarial training for defence. Evaluation metrics include precision, recall, F1-score, accuracy. Our assessment involves scrutinizing adversarial attacks and defences on histopathology images related to lung and colon issues, representing a state-of-the-art endeavour. The results indicate a significant improvement in susceptibility to adversarial attacks on histopathological images of the lungs and colon, from 0% to 81%.

KEYWORDS:

adversarial attack and defense, convolutional neural network, histopathology, image classification, lung and colon cancer

1 | INTRODUCTION

Cancer is the second largest cause of death in the world. This disease is characterized by uncontrolled growth of abnormal cells and can attack and spread to cells and body tissues. In some cases, lung and colon cancer can develop simultaneously, and both are the most common types of cancer. Usually, lung cancer is the second primary type of cancer found in patients with colon cancer, so the development of lung cancer in patients with colon cancer is very important to pay attention to[1]. According to the Global Burden of Cancer report published by the World Health Organization (WHO), in 2020, there were 19,292,789 new cases and 9,958,133 deaths due to cancer worldwide. Lung cancer is in second place at 11.4% of total cases, while colon cancer is in fourth place at 10%. In Indonesia, there were 396,941 new cases and 234,511 deaths due to cancer, with lung cancer in fourth place at 8.8% and colon cancer in sixth place at 8.6% of total cases[2]. Cancer deaths are expected to increase to more than 13,100,000 by 2030. The International Agency for Research on Cancer (IARC) estimates that one in five men and one in five women worldwide will suffer from cancer. Then, one in eight men and one in eleven women who suffer from cancer will die[3].

Early lung and colon cancer detection is critical for effective treatment and increasing survival rates. This can be done using various digital imaging techniques in the medical field, such as computed tomography (CT) scans, sputum cytology, chest X-rays, magnetic resonance imaging (MRI), and microscopic histopathology images. There are various diagnostic procedures for detecting cancer by observing medical images based on samples, such as sputum cytology and tissue removal (biopsy). During the biopsy process, a pathologist will take tissue samples from human organs and then evaluate the resulting microscopic histopathological images to carry out a diagnosis so that they can determine the type and subtype of cancer. Histopathological images are widely used in predicting a patient's chance of recovery. Technological developments have brought changes in diagnosing diseases using machine learning (ML) and deep learning (DL). ML and DL algorithms can support the diagnosis process and save costs with accurate results in large data sets. In clinical practice, classifying histopathological images accurately is very important to support the diagnosis of a disease at the tissue level[4].

Previous studies have shown that the DL algorithm can accurately classify histopathological images of lung and colon cancer. One of the popular DL algorithms is convolutional neural network (CNN), which can be used to classify types of lung and colon cancer based on histopathological images with high accuracy, as has been done by[5] that evaluated classification accuracy with CNN architecture to detect lung cancer tissue on the LC25000 dataset. Recent studies show deep neural networks (DNN) are highly vulnerable to adversarial attacks. An attacker who makes small changes in the form of perturbations to image samples undetectable by the human eye can significantly affect the performance of the DNN model. DNN models used in medical images are more vulnerable to attacks than those using natural images. Therefore, adversarial attacks are a major challenge in DL systems in medical imaging[6] [7]. Other research identifies adversarial attacks and instability of decision results as challenges in artificial intelligence and digital pathology, thereby creating fundamental problems that will be the focus of future research, including the potential for misrepresentation of facts in disease diagnosis[8]. Adversarial attacks are categorized based on the access model. In a white- box attack, the attacker has direct access to the target model parameters, while in a black-box attack, the attacker does not have access to the model parameters[9] [10]. Various defence mechanisms have been proposed to counter attacks. One of them is adversarial training, by adding adversarial samples to the training data to increase the model's resistance to attacks. Adversarial attack and defence mechanisms have performed better on DNN models with natural image datasets such as CIFAR-10. However, its performance on medical images is still not optimal due to the lack of medical images with quality labels[11].

There have been several previous studies regarding adversarial attacks and defences in medical imaging. For example, research conducted by[12] proposed a classification of diabetic retinopathy using data from a public dataset known as DR Fundus. They also use perturbed data generated through applying FGSM adversarial attacks while maintaining accuracy by applying a defence mechanism, namely adversarial training, to the DarkNet-53 model. Another study using two types of datasets, namely NIH Chest X-Ray and AREDS, was carried out by[13]. They used adversarial attack and defence methods, PGD and Sparsity Denoising on the ResNet50 model. This research succeeded in maintaining model accuracy when various attacks occurred. The latest research involves breast cancer tissue imaging, known as BreakHis, and was conducted by[14]. They use the DenseNet121 model and apply FGSM adversarial attacks and defences with the adversarial training.

However, until now, there has been no research related to adversarial attack and defence in the CNN model for classifying lung and colon cancer using histopathological images. Therefore, the author proposes research to improve the performance of the CNN classification model with adversarial attack and defence using the LC25000 and Chaoyang datasets. This research

uses three CNN models, namely GoogLeNet, ShuffleNetV2, and ResNet18, then applies an adversarial attack mechanism with a white-box attack approach, namely CW, DeepFool, and SaliencyMap attacks and a defence mechanism using adversarial training. Performance evaluation was carried out on the CNN classification model, the CNN model after an adversarial attack was carried out, and the CNN model after a defence mechanism was carried out. This research aims to improve the model's reliability and accuracy in classifying lung and colon cancer types on histopathological images. An update of this research is that We evaluate and analyze the adversarial attacks and defences on lung and colon cancer histopathology images, which is considered a state-of-the-art effort.

2 | PREVIOUS RESEARCHES

Several previous studies have reviewed adversarial attacks and defences in medical image data. However, the number of studies is still relatively small compared to case studies on natural data such as MNIST and CIFAR10. This research focuses on case studies of medical image data, especially with the application of white-box attacks. In this context, three previous studies that comply with the criteria set. The first study was conducted by [12]. They used the DR Fundus dataset and the DarkNet-53 classification model. This research adopts FGSM and adversarial training as adversarial attack and defence methods. The results show an accuracy of 99.90% for the normal model, 0% when the model is attacked, and 92% when the model defends. Subsequent research conducted by [13] used two datasets, namely NIH Chest X-Ray and AREDS, with the ResNet50-D classification model. This experiment used the PGD and sparsity denoising methods as adversarial attack and defence methods. The results show an accuracy of 91.94% on the normal model, 45.68% when the model is attacked, and 82.36% when the model survives on the first data. In the second data, accuracy reached 84.84% on the normal model, 28.92% when the model was attacked, and 46.57% when the model defended. In the same context, this research also uses the ResNet50-A-D model in a test scenario, with accuracy results of 92.96% on the normal model, 87.20% when the model is attacked, and 92.54% when the model survives on the first data. Meanwhile, in the second data, accuracy reached 81.93% for the normal model, 48.66% when the model was attacked, and 74.97% when the model survived.

The latest research was conducted by [14]. This research stands out because it is the only one that uses the same type of data as this research: histopathology medical image data with the name BreakHis dataset. The classification model used is CNN with pre-trained DenseNet121. The adversarial attack and defence methods applied are FGSM and adversarial training. The test results show an accuracy of 98.72% for the normal model, 10.99% when the model is attacked, and 96.70% when the model survives. This research contributes to understanding adversarial attacks and defences in histopathological medical image data.

3 | METHOD

The framework applied in this research consists of three stages: image classification model, adversarial attack, and adversarial defence, as presented in Figure 1. The first stage involves collecting lung and colon histopathology image data from a public dataset called LC25000. Next, data preprocessing was carried out by resizing all image sizes to 255 x 255, increasing the brightness level in the image randomly by 0.05%, and normalizing it by converting the image into a tensor so that each pixel in each image is in the value range 0– 255 and normalizing each tensor with a mean and standard deviation of 0.5. In addition, data splitting was carried out, where the processed data was divided into two sets, namely training data and test data, with percentages of 70% and 30%, respectively. This data set becomes the output of the data preprocessing stage. The process continues with the model training stage. This research uses three types of CNN classification models: GoogLeNet, ShuffleNetv2, and ResNet18. After model training, model testing, results evaluation, and performance analysis of the resulting models are carried out.

The next stage is an adversarial attack, where interference will be added to each set of training data and test data using predetermined attack methods, namely CW, DeepFool, and SaliencyMap attacks. The result of this process is disturbed training data and test data. Next, testing is carried out on the same classification model as before using disturbed test data. Results and model performance analysis are evaluated when the model receives input from disturbed data. The final stage is adversarial defence, where adversarial training is used. At this stage, the classification model is retrained with perturbed training data. Thus, the model knowledge will be updated, and the model can be considered robust. Next, the robust model is tested with normal test data to prove that the model can still classify normal data well. In addition, testing was carried out with disturbed test data to

show that the robust model-maintained accuracy from the previous stage when the model experienced attacks in the form of disturbed input data. Next, a performance analysis of the three

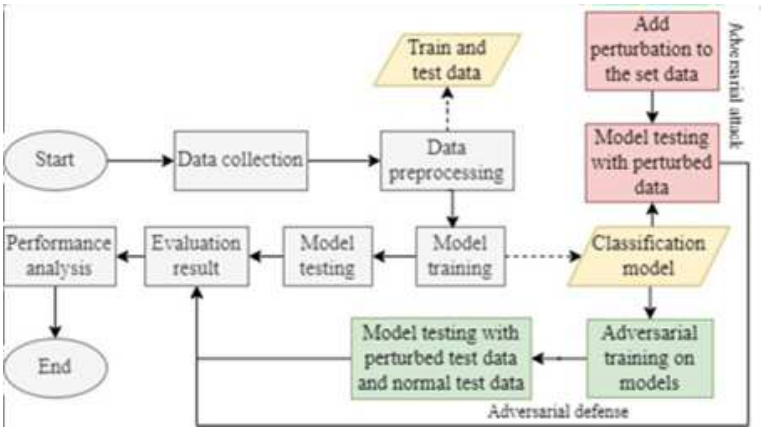

FIGURE 1 The framework applied in this research

stages were carried out for the accuracy obtained from each stage to prove that the adversarial attack and defence method could increase the resilience of the classification model to attacks.

3.1 | Dataset Description

In this research, a dataset of histopathology images depicting lung and colon cancer, designated LC25000 and released in 2019. The dataset, curated by Andrew A. Borkowski and team, encompasses 25,000 color histopathology images representing five distinct tissue types found in the lungs and colon. The identified tissue types encompass colon adenocarcinoma, benign colon tissue, lung adenocarcinoma, benign lung tissue, and lung squamous cell carcinoma[15]. Colon adenocarcinoma stands out as the predominant form of colon cancer, constituting over 95% of reported cases. Its onset involves the transformation of a specific type of tissue growth known as an adenoma within the colon, progressing into cancer. Lung adenocarcinoma, more prevalent in women than men, contributes to approximately 40% of all lung cancers. Characterized by the development of cancerous cells in gland cells, it subsequently spreads to the alveoli in the lungs. It's crucial to note that not all tumors originating in the lungs and colon are cancerous; those that do not spread to other body parts are termed benign tumors. Although generally non-life-threatening, these tumors necessitate surgical removal and biopsy examination to rule out cancer. Lastly, lung squamous cell carcinoma, a subtype of small cell cancer, emerges in the airways of the lungs or bronchi. Ranking as the second most prevalent type of lung cancer, it constitutes about 30% of all cases[16]

The images within the LC25000 dataset were gathered at the James A. Haley Veterans Hospital in Tampa, Florida. The team of researchers procured 1,250 images of cancerous tissues (250 images for each tissue type) from pathology slides. Subsequently, image augmentation techniques were implemented, involving rotations and flips of the original image under various conditions, resulting in an expanded dataset comprising 25,000 images (5,000 images for each tissue type). Initially, the original images were sized at 1024×768 pixels, and after the application of augmentation techniques, they were cropped to a square format of 768×768 pixels. It is noteworthy that all images in the dataset adhere to the regulations outlined in the Health Insurance Portability and Accountability Act (HIPAA), are validated, and are available for unrestricted use[15]. Figure 2 illustrates examples of histopathology images from the five classes sourced from the LC25000 dataset.

3.2 | Data Preprocessing

Before classifying image data with a Deep Learning (DL) model, the first step that needs to be taken is image preprocessing[17][18]. In this research, several data preprocessing techniques involved the following steps. First, resizing is carried out, where the size of all images in the dataset is changed to 255×255 pixels. Next, a Brightness step is applied, which includes

randomly increasing the brightness level on each image by 0.05%, aiming to improve identification performance. Next is Normalization, where the image is converted into a tensor so that every pixel in each image is in the value range 0–255. After that, normalization was carried out for each tensor using a mean and standard deviation of 0.5. The final step is data splitting, where the preprocessed data is divided into three sets: training and test data. The distribution is carried out with 70% and 30% percentages, respectively. These steps ensure the data is ready to train and test image classification models using deep learning models.

3.3 | Classification Model

a. GoogLeNet
The GoogleNet (Inception) model is a convolutional neural network architecture designed for image classification. Developed by Google researchers, GoogleNet introduces several innovations to enhance network efficiency and performance. Notably, the model employs Inception modules featuring multiple convolution paths with different filter sizes, enabling simultaneous feature extraction at various spatial scales for recognizing complex patterns. Integrating 1x1 convolutions facilitates dimensionality reduction, allowing for increased network depth without substantially increasing computational load, making the model more efficient. Global average pooling replaces traditional fully connected layers, reducing parameters, preventing overfitting, and enhancing interpretability. GoogleNet excels in computational efficiency, providing a deep representation of image features and contributing to a better understanding of hierarchical structures. While

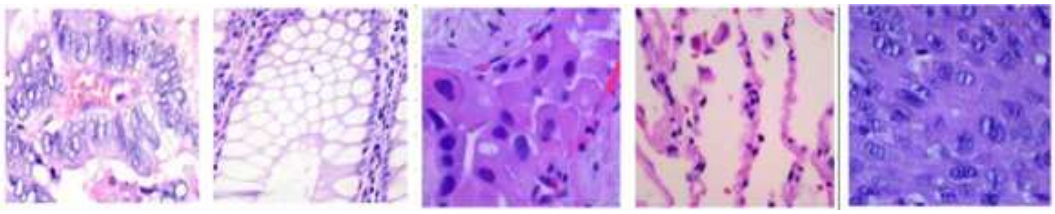

FIGURE 2 Samples of histopathology images from LC25000 dataset

subsequent architectures like ResNet and EfficientNet have emerged, GoogleNet remains a milestone in convolutional neural network design for image processing [19].

b. ShuffleNetV2
The ShuffleNetV2 model is designed to focus on practical guidelines for efficient convolutional neural network (CNN) architecture. One of its key features is the introduction of channel shuffling, a technique that facilitates efficient information exchange between channels, enhancing information flow without compromising computational efficiency. The model employs grouped pointwise convolution to reduce computational costs further, separating channels into groups to achieve efficiency without sacrificing performance. With a multi-path architecture, ShuffleNetV2 allows for parallel processing through different pathways, contributing to model efficiency and scalability. Notably, the paper provides practical guidelines for designing efficient CNN architectures, offering valuable insights for researchers and practitioners. Despite its emphasis on efficiency, ShuffleNetV2 achieves state-of-the-art accuracy in image classification tasks. It is a compelling choice for applications where computational resources are constrained and efficiency is paramount. Overall, ShuffleNetV2's innovative design and its ability to balance accuracy and computational complexity position it as a significant contribution to the field of CNN architectures [20].

c. ResNet18
ResNet architecture, with ResNet18 being a notable variant. This model is distinguished by its 18 layers, employing deep residual blocks characterized by shortcut connections. These shortcuts effectively train very deep networks by mitigating the vanishing gradient problem. Including skip connections facilitates a more direct flow of gradients, addressing the degradation problem observed in deep networks. ResNet18's innovative design simplifies training, allowing for learning identity mappings and contributing to ease of optimization. The model demonstrates improved accuracy on various image

recognition tasks compared to earlier architectures, showcasing its effectiveness in capturing intricate features and patterns. The scalability of ResNet18, driven by its residual block design, further positions it as a pivotal advancement in the field of image recognition, influencing subsequent architectures for their adaptability to more complex tasks [21].

3.4 | Adversarial Attack

In the DNN model, several security problems were found, such as adversarial attacks, representing most of these problems[19] [20] [21] [22]. In Figure 3 there is an adversarial attack workflow which explains how this method works. The attacker aims to create adversarial samples that can make incorrect classifications. In health, avoidance can be described as classifying benign cells as cancer cells or vice versa. This attack is carried out by adding perturbation or modifying image features. That will impact how the model views the input image.

a. CW attack
CW attack[23] is a significant adversarial method for exploring and exploiting neural network vulnerabilities. This attack method is known for its effectiveness in creating adversarial images. In essence, a CW attack formulates an objective function designed to minimize noise in the input data, ensuring that modified input results in misclassification by the targeted neural network. Importantly, this attack is flexible because it is not

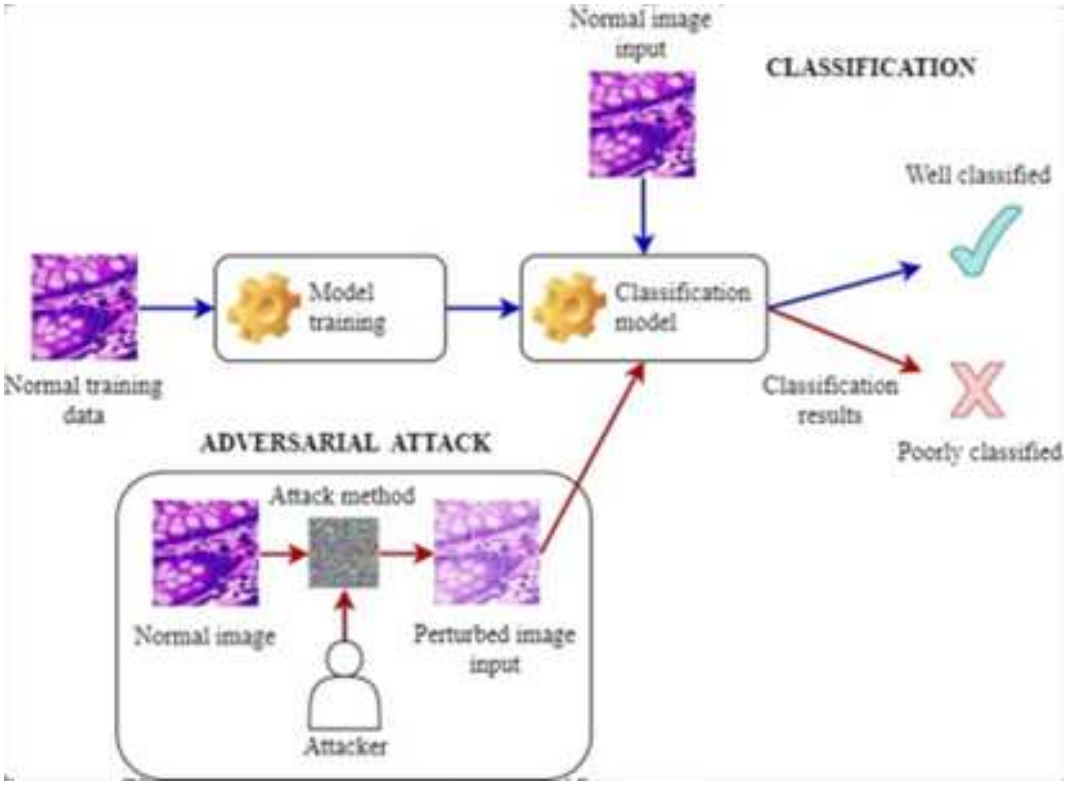

FIGURE 3 Adversarial attack workflow

strictly bound by a particular norm such as L2 or L ∞ , providing flexibility in generating effective adversarial interference. This attack uses an iterative optimization algorithm that systematically adjusts the input data to minimize the objective function, looking for the smallest noise that causes misclassification. An additional parameter, referred to as c, is introduced to control the magnitude of the perturbation, and its adaptive adjustment during optimization allows for achieving the desired level of adversarial perturbation. The CW attack considers the misclassification confidence level, attempting to cause misclassification.

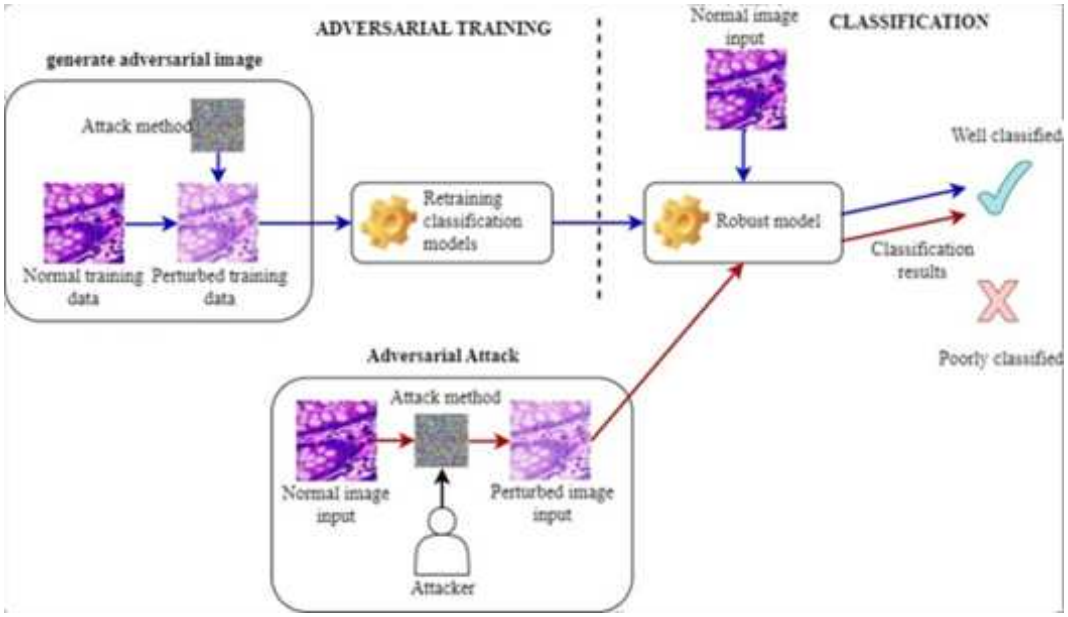

FIGURE 4 Adversarial training workflow

b. DeepFool attack
DeepFool is an undirected adversarial attack method designed to produce changes to input data to trick neural networks in their classification. This attack operates by finding the minimum perturbation necessary to fool the neural network's classification results, focusing on minimizing the L2 norm of that perturbation. DeepFool adopts a linear approximation approach, with the initial assumption that the decision boundary of the neural network is linear. Through an iterative process, the attack refines this approximation and computes the optimal perturbation to cross the decision boundary in the linearized space. Iterative steps are carried out until the true adversarial image is found, which ultimately results in misclassification by the neural network. Simplicity, analytical derivation, and iterative optimization make DeepFool an effective method for generating minimal perturbations that can successfully mislead neural networks in their classification[24].

c. SaliencyMap attack
SaliencyMap is a technique used in adversarial attacks that leverages salient maps—visualizations highlighting crucial regions or features in input data that a deep learning model draws attention to when making predictions. In the context of adversarial attacks, this method involves creating a salience map for a given input, identifying regions of influence that, if altered, could cause a change in the model's predictions. The attacker then creates visually invisible changes targeted at those vulnerable regions. Next, these altered inputs, called adversarial examples, are evaluated on the model to see whether they produce misclassifications or achieve the attacker's desired results. The success of a SaliencyMap attack depends on factors such as model architecture, dataset characteristics, and robustness of model training. This approach allows attackers to exploit model attention mechanisms, highlighting the complex challenges of maintaining the security of deep learning models in adversarial environments[25].

3.5 | Adversarial Defense

The author carried out data modifications to mitigate the attacks described previously, namely a countermeasure to attacks by changing the data or its features. Several studies mentioned previously have presented a method often used called adversarial training[26]. Adversarial training is a method for adding adversarial samples to the dataset but with the correct labels before retraining the model[27]. Retraining means using the previous normal model and doing more epochs but with an adversarial dataset. Thus, the model will learn modified features and become more robust when faced with adversarial samples or perturbed images. An adversarial training workflow is provided to make it easier to understand this method, which can be seen in Figure 4 .

3.6 | Evaluation Metric

Evaluation metrics serve as crucial benchmarks for measuring the quality of machine learning models, offering insights into the performance of trained deep learning algorithms on novel, unseen data. The landscape of evaluation metrics is diverse, providing varied tools to assess model performance. Employing multiple evaluation metrics is highly recommended, given that a model may excel in one metric while underperforming in another. Thus, the judicious utilization of evaluation metrics becomes essential in accurately determining the precision and optimization of the resulting model[28]. This section furnishes a concise explanation and relevant formulas for the evaluation metrics utilized in the research.

In the context of model predictions, correctly identifying the positive class is designated as True Positive (TP). Conversely, an erroneous prediction of the positive class is termed False Positive (FP). Similarly, accurate predictions of the negative class are denoted as True Negative (TN), while incorrect predictions of the negative class are referred to as False Negative (FN). To illustrate, in an image containing cancer cells, successfully classifying the cancerous region is categorized as TP. However, if the model erroneously classifies a cancer cell as non-cancerous, it is classified as FP[29].

Conversely, in scenarios where the image does not contain any cancer cells, and the model accurately predicts the absence of cancer cells, it is labeled as True Negative (TN). Nevertheless, if the model incorrectly predicts the presence of cancer cells in the absence of any, it is denoted as False Negative (FN). Subsequently, a succinct explanation will follow, elucidating the formulas associated with evaluation metrics[30].

Metrics for accuracy, which gauge the proportion of correct predictions in relation to the total number of assessed samples, as depicted in Equation 1 provided below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

Precision is employed to assess the precision of positive predictions among the entire set of predicted observations within the positive class, as delineated in Equation 2 provided below:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall serves as a metric to quantify the correctness of classifying positive observations, as expressed in Equation 3:

$$\text{Recall} = \frac{TP}{TP + FP} \tag{3}$$

The F1 Score involves assessing the weighted average of precision and recall, as expressed in the following Equation 4:

$$\text{F1 Score} = 2x\frac{Precisionx Recall}{Precision + Recall} \tag{4}$$

4 | RESULT AND CONCLUSION

In this section, we present the results of the experiments we have conducted. Table 1 shows the test results of three different pre-trained models, namely GoogLeNet, ShuffleNetV2, and ResNet18, after training for ten epochs in classifying lung and colon histopathology images. Model testing uses evaluation metrics of precision, recall, F1-score, and accuracy for each label in the test data. The test results show that the three models achieve the same level of accuracy, namely 99%. However, when evaluating the average value of the other three evaluation metrics, the ResNet18 model emerged as the best.

Table 2 illustrates the results of applying the classification model using the adversarial attack and defence method. Of the three adversarial attack methods used, the SaliencyMap attack shows superiority in deceiving the model, decreasing accuracy by up to 0%. Followed by the CW attack with a decrease rate of 0.01% and the DeepFool attack with an accuracy decrease range of 0.30% to 0.50%. Previously, the accuracy of each classification model was 99%. Furthermore, after undergoing retraining

TABLE 1 The classification model performance results (%)

Model	Label	Metric			
		Precision	Recall	F1-Score	Accuracy
GoogLeNet	colon aca	1	0.99	0.99	0.99
	colon_n	1	1	1	
	lung_aca	0.97	0.98	0.97	
	lung_n	1	1	1	
	lung_scc	0.98	0.97	0.97	
	colon_aca	1	1	1	
	colon_n	1	1	1	
ShuffleNetV2	lung_aca	0.98	0.98	0.98	0.99
	lung_n	1	1	1	
	lung_scc	0.98	0.98	0.98	
	colon_aca	1	1	1	
	colon_n	1	1	1	
ResNet18	lung_aca	0.97	0.98	0.98	0.99
	lung_n	1	1	1	
	lung_scc	0.98	0.97	0.87	

TABLE 2 The model performance results with adversarial attack and defense (%)

Model	Normal Accuracy	Attack Method	Normal Model Robust nodel	Accuracy on Normal Data on Normal Data	Robust Model on Perturbed Data
GoogLeNet	0.99	CW	0.01	0.98	0.81
		DeepFool	0.4	0.98	0.76
		SaliencyMap	0	0.97	0.59
ShuffleNetV2	0.99	CW	0.01	0.99	0.63
		DeepFool	0.54	0.99	0.73
		SaliencyMap	0	0.98	0.5
ResNet18	0.99	CW	0.01	0.98	0.55
		DeepFool	0.38	0.99	0.73
		SaliencyMap	0	0.99	0.79

using disturbed training data, or what is known as adversarial training, the model becomes more robust because its knowledge is updated. The immune-enhanced models are tested using normal test data to prove their ability to classify the data well. The ShuffleNetV2 and ResNet18 models obtained the best accuracy defence results on normal data, reaching 99%. Next, the enhanced robustness models were tested using perturbed test data, which is the main focus of this research. That will prove that a model attacked with disturbed data can still maintain its accuracy. The best results on the models tested with disturbed data were seen in the GoogLeNet model, which managed to maintain accuracy from the normal level of 99%, experienced a decrease in accuracy when attacked to 0%, and was successfully maintained at 81%.

Table 3 compares the performance results of the image classification model that applies the adversarial attack and defence method with previous research on medical image case studies. This comparison cannot be made directly due to differences in the dataset and case studies used with previous research. Therefore, this comparison is designed to demonstrate that our research makes an innovative contribution, illuminating a previously unexplored case study. All studies listed in Table 3 use white-box attacks, and most studies apply adversarial training for defence methods. The only exception was one study that used a different defence method: sparsity denoising.

TABLE 3 The comparison results with previous study (%)

Study	Datasets DR	Classification Method	Adv. Method FGSM/Adv.	Normal Acc	Attacked Acc	Defended Acc
[12]	Fundus Chest X-Ray	DarkNet-53	Training	99.90%	0%	92%
		ResNet50-A-D		91.94%	45.68%	82.36%
[15]	Chest X-Ray AREDS	ResNet50-A-D	PGD/ Sparsity Denoise	92.96%	87.20%	92.54%
		ResNet50-D		84.84%	28.92%	46.57%
	BreakHis	ResNet50-A-D	PGSM/ Adv. Training	81.93%	48.66%	74.97%
		DenseNet21		98.72%	10.99%	96.70%
[14]	This Research	GooleLeNet	CW / Adv. Training	0.01	0.98	0.81
			DeepFool /Adv. Training	0.4	0.98	0.76
			SaliencyMap / Adv. Training	0	0.97	0.59
		ShuffleNetV2	CW/ Adv. Training	0.01	0.99	0.63
			Deep Fool Adv. Training	0.54	0.99	0.73
			Saliency Map Adv. Training	0	0.98	0.5
		ResNet18	CW/ Adv. Training	0.01	0.98	0.55
			Deep Fool Adv. Training	0.38	0.99	0.73
			Saliency Map Adv. Training	0	0.99	0.79

5 | CONCLUSION

In this research, we have successfully implemented robustness for our trained models to classify lung and colon cancer histopathology data in the LC25000 dataset by performing adversarial training methods. Before adversarial training, the models were unable to predict perturbed input correctly, resulting in a decrease in the normal model's accuracy to 0% for all our pretrained models. Our most significant improvement is increasing the model accuracy on perturbed images from 0.01 to 0.81.

In detail, we have performed various types of adversarial methods, such as CW, DeepFool, and SaliencyMap, and implemented adversarial defences using adversarial training to reduce the impact of adversarial attacks. Experimental results show that SaliencyMap is a very effective attack method, reducing model accuracy from 99% to 0%. Additionally, increased accuracy was achieved after applying the best adversarial training on normal data for the ShuffleNetV2 and ResNet18 models, each reaching 99%. Meanwhile, the GoogLeNet model achieved the highest accuracy rate of 81% for improved accuracy on perturbed data, previously attacked by perturbed images generated by the CW method, resulting in low accuracy, which is 1%.

For the next improvement, other defense mechanisms might be added during the enhancement of the models, such as implementing Defensive Distillation, Interval Bound Propagation, Defense GAN, or other well- proven defense mechanism techniques.

CREDIT

Conceptualization, Material preparation, Methodology, Data collection, Formal analysis, and writing the original draft were conducted by Chilyatun Nisa'. The Writing review and editing were carried out by Nanik Suciati, Anny Yuniarti, and Chilyatun Nisa'. The supervision was provided by Nanik Suciati and Anny Yuniarti. All authors have reviewed and approved the final version of the manuscript for publication.

References

1. Kurishima K, Miyazaki K, Watanabe H, Shiozawa T, Ishikawa H, Satoh H, et al. Lung cancer patients with synchronous colon cancer. Molecular and Clinical Oncology 2017;8:137–140. https://www.spandidos-publications.com/10.3892/mco.2017.1471.

2. IARC. Global Cancer Observatory (GLOBOCAN). International Agency for Research Cancer 2019;http://gco.iarc.fr/today/home.

3. Dr dr Terawan Agus Putranto SR. Profil Kesehatan Indonesia Tahun 2019. Kementerian Kesehatan Republik Indonesia 2017;file:///C:/Users/DRPM-ITS/Downloads/Profil-Kesehatan-Indonesia-2019.pdf.

4. Chehade AH, Abdallah N, Marion JM, Oueidat M, Chauvet P. Lung and colon cancer classification using medical imaging: a feature engineering approach. Scientific Paper 2022;45:729–746. https://link.springer.com/article/10.1007/s13246-022-01139-x.

5. Hatuwal BK, Thapa HC. Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images. International Journal of Computer Trends and Technology 2019;68:21–24. https://ijcttjournal.org/archives/ijctt-v68i10p104.

6. Ma X, Niu Y, Gu L, Wang Y, Zhao Y, Bailey J, et al. Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images. Pattern Recognition 2021;110:107332. https://www.scopus.com/authid/detail.uri?authorId=54956194300.

7. Diyasa IGSM, Wahid RR, Amiruddin BP. Lung Cancer Detection Using Convolutional Neural Network on Histopathological Images. 2021 International Conference on e-Health and Bioengineering (EHB) 2021;p. 1–4. https://ieeexplore.ieee.org/document/9657589.

8. Sipola T, Puuska S, Kokkonen T. Model Fooling Attacks Against Medical Imaging: A Short Survey. 2021 International Conference on e-Health and Bioengineering (EHB) 2021;46:215–224. https://isij.eu/article/model-fooling-attacks-against-medical-imaging-short-survey.

9. Thangaraju A, Merkel C. Exploring Adversarial Attacks and Defenses in Deep Learning. 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) 2022;https://ieeexplore.ieee.org/document/9865841.

10. Wei C, Sun R, Li P, Wei J. Analysis of the no-sign adversarial attack on the covid chest x-ray classification. 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) 2022;p. 73–79. https://ieeexplore.ieee.org/document/9887371.

11. Kaviani S, Han K, Sohn J. Analysis of the no-sign adversarial attack on the covid chest x-ray classification. Expert Systems with Applications 2022;198:116815. https://ieeexplore.ieee.org/document/9887371.

12. Lal S, Rehman SU, Shah JH, Meraj T, Rauf HT, Damaševičius R, et al. Adversarial Attack and Defence through Adversarial Training and Feature Fusion for Diabetic Retinopathy Recognition. MDPI 2021;21:3922. https://www.mdpi.com/1424-8220/21/11/3922.

13. Shi X, Peng Y, Chen Q, Keenan T, Thavikulwat AT, Lee S, et al. FRNet: A Feature-Rich CNN Architecture to Defend Against Adversarial Attacks. Pattern Recognition 2021;132:108923. https://ieeexplore.ieee.org/document/10431780.

14. Li Y, Liu S. Adversarial attack and defense in breast cancer deep learning systems. Bioengineering 2021;10:973. https://ieeexplore.ieee.org/document/10431780.

15. Anjum S, Ahmed I, Asif M, Aljuaid H, Alturise F, Ghadi YY, et al. Lung Cancer Classification in Histopathology Images Using Multiresolution Efficient Nets. Computational Intelligence and Neuroscience 2023;https://onlinelibrary.wiley.com/doi/10.1155/2023/7282944.

16. Sikder MM, A NN, Bairagi A, A AKAM. A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework. Sensors MDPI 2021;21:748. https://www.mdpi.com/1424-8220/21/3/748.

17. He W, Liu T, Han Y, Ming W, Du J, Liu Y, et al. A review: The detection of cancer cells in histopathology based on machine vision. Computers in Biology and Medicine 2022;https://www.sciencedirect.com/science/article/pii/S0010482522004280?via%3Dihub.

18. S V, M S, S K. A new complete color normalization method for H and E stained histopatholgical images. Applied Intelligence 2021;51:7735–7748. https://link.springer.com/article/10.1007/s10489-021-02231-7.

19. C S, W L, Y J, P S, S R, D A, et al. Going deeper with convolutions. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2015;https://www.scopus.com/authid/detail.uri?authorId=8764903400.

20. N M, X Z, H Z, J S. Practical guidelines for efficient cnn architecture design. MDPI 2018;https://link.springer.com/chapter/10.1007/978-3-030-01264-9_8.

21. M S, Z G. Deep Residual Learning for Image Recognition: A Survey. Applied Sciences (Switzerland) 2022;https://www.mdpi.com/2076-3417/12/18/8972.

22. Y C, M Z, J L, X K. Adversarial attacks and defenses in image classification: A practical perspective. 2022 7th International Conference on Image, Vision and Computing (ICIVC) 2022;p. 424–430. https://ieeexplore.ieee.org/document/9886997.

23. N C, D W. Towards evaluating the robustness of neural networks. Proceedings - IEEE Symposium on Security and Privacy (2017) 2017;p. 39–57. https://ieeexplore.ieee.org/document/7958570.

24. M MDS, A F, P F. DeepFool: A simple and accurate method to fool deep neural networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016) 2016;https://ieeexplore.ieee.org/document/7780651.

25. N P, P M, S J, M F, B CZ, A S. The Limitations of Deep Learning in Adversarial Settings. 2016 IEEE European Symposium on Security and Privacy (EuroS and P) 2015;https://ieeexplore.ieee.org/document/7467366.

26. Z H, S RA, D F. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019;https://ieeexplore.ieee.org/document/8954187.

27. J W, C W, Q L, C L, C W, J L. Adversarial attacks and defenses in deep learning for image recognition: A survey. Neurocomputing (2022) 2022;p. 162–181. https://doi.org/10.1016/j.neucom.2022.09.004.

28. Dalianis, H. Evaluation metrics and evaluation. Dalam H. Dalianis, Clinical Text Mining. Springer International Publishing 2018;p. 45–53. https://doi.org/10.1007/978-3-319-78503-5_6.

29. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Scientific Reports 2018;12:45–53. https://doi.org/10.1038/s41598-022-09954-8.

30. Pacal I, Karaboga D, Basturk A, Akay B, Nalbantoglu U. A comprehensive review of deep learning in colon cancer. Computers in Biology and Medicine 2020;126. https://doi.org/10.1016/j.compbiomed.2020.104003.

How to cite this article: Chilyatun N., Nanik S., Anny Y., ADVERSARIAL TRAINING FOR ROBUST DEFENSE IN CNN MODELS FOR LUNG AND COLON HISTOPATHOLOGICAL IMAGES, *IPTEK The Journal of Technology and Science*