Journal of Emerging Computer Technologies
Research Article
Received: *2022-10-25* | ***Reviewing:*** *2022-11-08 & 2022-12-13* | ***Accepted:*** *2022-12-13* | ***Online:*** *2022-12-14* | ***Issue Date:*** *2022-12-31*

Measuring The Robustness of AI Models Against Adversarial Attacks: Thyroid Ultrasound Images Case Study

Mustafa Ceyhan
Huawei Turkey R&D Center / Artificial Intelligence MsC, Mugla Sitki Kocman University
Istanbul, Türkiye
0000-0003-3268-6898
mustafac94@gmail.com

Enis Karaarslan
Department of Computer Engineering
Mugla Sitki Kocman University
Mugla, Türkiye
0000-0002-3595-8783
enis.karaarslan@mu.edu.tr

***Abstract*— The healthcare industry is looking for ways on using artificial intelligence effectively. Decision support systems use AI (Artificial Intelligence) models that diagnose cancer from radiology images. These models in such implementations are not perfect, and the attackers can use techniques to make the models give wrong predictions. It is necessary to measure the robustness of these models after an adversarial attack. The studies in the literature focus on models trained with images obtained from different regions (lung x-ray and skin dermoscopy images) and shooting techniques. This study focuses on thyroid ultrasound images as a use case. We trained these images with VGG19, Xception, ResNet50V2, and EfficientNetB2 CNN models. The aim is to make these models make false predictions. We used FGSM, BIM, and PGD techniques to generate adversarial images. The attack resulted in misprediction with 99%. Future work will focus on making these models more robust with adversarial training.**

Keywords*— *Adversarial Attack, CNN Models, Thyroid Ultrasound Images, Machine Learning, Deep Learning

I. INTRODUCTION

Today, applications of artificial intelligence such as object detection, voice recognition, recommendation systems, credit risk estimations, and disease diagnosis are widely used in many sectors such as health, finance, robotics, agriculture, marketing, and education. AI is increasingly used in the healthcare industry, especially in radiology (1). Applications developed using artificial intelligence in cancer diagnosis allow radiologists to make faster and more reliable diagnoses. Several models are used for different diseases and cancer types of the brain (2), skin (3), breast (4), lung (3) and thyroid (5) using different types of medical images. Models trained with these images usually have successful results. Researchers continue to try new techniques with new data and new models.

Artificial intelligence is a black box application that can predict a situation by learning from structured or unstructured data. In cases where mathematics and statistics cannot conclude, they can catch patterns in the data and reach the correct result with high accuracy rates. Classical machine learning algorithms have been used successfully and continue to be used for a long time. However, classical machine learning algorithms cannot achieve the desired success in some areas like image recognition or natural language processing. We can use deep learning to overcome this problem. Learning has become better by adding multiple hidden layers to artificial neural networks. We can achieve deeper learning with this method and capture the patterns between the inputs better. CNN models, which are different versions of neural networks, are used for image recognition and classification, but these models are open to adversarial attacks. The attackers can use several attack methods to fool artificial intelligence models. These can cause the models to give false results. This can have unacceptable outcomes in the healthcare industry.

In this study, we applied attack techniques to evaluate thyroid ultrasound images. This data type was chosen as the literature lacked using attack techniques against a model trained with thyroid ultrasound images. We are using a two-output model developed for the diagnosis of thyroid cancer. This model returns whether the patient has cancer based on the input image. In this study, we tested VGG19, Xception, ResNet50V2, and EfficientNetB2 CNN models against the adversarial images generated using FGSM, BIM, and PGD techniques on an ultrasound image dataset. The major contribution of this paper is to show that artificial intelligence models trained for thyroid diagnosis using thyroid ultrasound images have a vulnerability to adversarial images.

We give a literature survey and present a brief background on Convolutional Neural Networks models for medical diagnosis and Adversarial Attacks to these models in section 2. In section 3, the method is given. The implementation is presented in section 4. The discussion is given in section 5. Finally, we conclude in section 6 with a hint of future works.

II. FUNDAMENTALS

A. Literature Survey

Finlayson et. al trained three models with three different datasets in their study (3). These are fundoscopy images for diabetes disease, x-ray images for pneumothorax disease, and dermoscopy images for skin cancer. They have performed successful white-box PGD attacks on the trained models. Bortsova et al. conducted black box attacks on diabetes, pneumothorax, and pathology data and achieved successful results (6). Alexandra et al. used FGSM, and JSMA techniques on a model developed for brain and lung cancer diagnosis and achieved successful results (7).

Goodfellow et al. demonstrated that artificial intelligence models can be fooled by perturbing the input image with noise calculated by the FGSM algorithm (8). This

Cite (APA): Ceyhan, M., Karaarslan, E. (2022). Measuring The Robustness of AI Models Against Adversarial Attacks: Thyroid Ultrasound Images Case Study. *Journal of Emerging Computer Technologies*, 2(2), 42-47.
Volume:2, ***No***:2, ***Year:*** 2022, ***Pages:*** 42-47, December 2022, *Journal of Emerging Computer Technologies*

attack tries to calculate the noise in the direction of the gradient, which will increase the loss of the model in one step. After calculating the gradient, the minimum amount is added to the input image that makes the model cross the decision boundary. This amount of addition is adjusted with epsilon. In another study published in 2017 by Goodfellow et al., they introduced the BIM algorithm, which is the iterative version of the FGSM attack (9). This algorithm calculates minimum perturbation more effectively using multiple steps. At each step, the algorithm tries to find the minimum perturbation by gradually increasing the epsilon value. Madry et al. used the PGD method similar to the BIM algorithm in 2017 (10). Different algorithms in the literature will perturb the input data to make it wrongly guessed. Among them, Jacobian-based Saliency Map Attack (11), DeepFool (12), Carlini & Wagner Attack (13), One Pixel Attack (14), and Adversarial Patch (15) are the leading effective attack techniques. Most of these studies have been tried on Image-Net (16), MNIST (17), and CIFAR (18) datasets and have shown successful results.

B. Convolutional Neural Networks

Convolutional neural network (CNN) is the most well-established method among the numerous deep learning models (19). CNN can be applied to different fields such as image classification, object detection, time series, and natural language processing. In simple terms, the properties of the input data are extracted in the convolutional and pooling layers. These extracted features are turned into a vector, and the weights of the deep neural network are trained. The most important part here is the feature extraction part. The better the feature extraction part, the more accurately the parameters of dense layers can be calculated. Various techniques have been developed since the first feature extraction applications such as LeNet and AlexNet, VGG19, Resnet50V2, Xception, and EfficientNetB2 are the most widely used models.

VGG19 (20) is a classic CNN model. Since the LeNet model, more convolution and pooling layers have always been added to the feature extraction part to achieve better results. VGG19 is a continuation of this tradition and consists of 16 convolution and 5 pooling layers in total. It has too many parameters and a high volume because the parameter feature mitigation layer is not applied.

ResNet (21) architecture is a model developed as an alternative to traditional feature extraction architecture. Adding more convolution and pooling layers does not increase the accuracy values after a certain level. Adding more of these layers causes the gradient to vanish or explode after a while. A strategy called residual nets has been developed to prevent this and to get more efficient results in fewer layers.

Xception (22), is a model based on the Inception model. It uses an approach called depthwise separable convolutions. It consists of blocks that try to capture different features with different filters. The model is formed from the combination of these blocks.

EfficientNet (23), one of the most recently developed models, and its derivatives can be considered the best in its field now. EfficientNet model versions are among the most successful feature extraction algorithms. It scales the model in depth, width, and resolution to get a better model.

C. Adversarial Attacks

Adversarial attacks are attempts to fool artificial intelligence models. Barreno et al. listed these attacks and their types comprehensively for the first time in their study "Can Machine Learning be Secure" (24). Artificial intelligence security is discussed more comprehensively in the study called “The Security of Machine Learning” (25). Attacks can be classified into different sub-categories such as the attacker's impact, knowledge, and specificity (26).

*1) Attacks based on the influence of the attacker :*These attacks are: causative attacks, evasion attacks and exploratory attacks.

Causative attacks occur in the training part of the model. These attacks are also called poisoning attacks. Data is added to the dataset, which will cause the model parameters to be miscalculated during training. Data poisoning can be the swapping of labels of training data or specially crafted data. It is used in attacks against models that are constantly trained with new data coming from outside in real-world scenarios.

Evasion attacks are performed on a trained model. The attacker makes an attempt to fool the model with perturbed data, that is, with adversarial examples. The attack occurs by adding noise to the data. Incorrect predictions are targeted by adding various noises. However, the important point here is that adversarial examples cannot be noticed by the human eye. Gradient-based adversarial example generation algorithms are the most successful noise addition methods. The study of Goodfellow et al. (8) can be considered a pioneer in applying this attack to images. Adversarial examples can be formed quickly and cheaply with their FGSM algorithm.

Exploratory attacks are based on a trained model such as evasion. The purpose here is to gather information about the model. It can be used to launch another attack in the future based on the gathered information.

2) Attacks based on the attacker knowledge: These attacks are white box and black box attacks. In white box attacks, the attacker knows the model and its parameters. In black box attacks, the attacker does not know the model and parameters.

3) Attacks based on attacker specificity: These attacks are targeted and untargeted. In targeted attacks, the perturbed data is asked to correctly predict a selected class instead of the actual class. In untargeted attacks, the aim is simply to misclassify the model.

III. METHOD

Thyroid ultrasound images were used as a dataset. These images are inherently noisy data (27). Therefore, more complex and deep models are more successful for feature extraction. Convolutional neural networks are used to train the data. Then Adversarial attack algorithms are used to attack the model.

A. Dataset

We used a dataset of Thyroid Ultrasound images like Fig 1 from Kaggle for the training application (29). There are 3282 cancer-free images and 4006 cancer-containing images in the dataset. The dataset is formed of training, testing, and validation sections. These sections are preprocessed for each model for training.

B. Preprocessinn and Training

Convolutional Neural Networks (CNN) are used in this study. CNN model is used in different architectures for training. Four different models (VGG19, Resnet50V2, Xception, and EfficientNetB2) are used. The characteristics of the models are given in Table 1. EfficientNetB2 expects each input pixel to be in the normal value range of 0-255 (23). Other models scale from 0-1 (20-22). Data is fitted to each model during this preprocessing.

TABLE I. THE FEATURES OF THE MODELS

Model	**Size (MB)**	**Top-1 Acc.**	**Top-5 Acc.**	**Param.**	**Depth**
VGG 19	549	71.3%	90%	90M	19
ResNet50V2	98	76%	93%	93M	103
Xception	88	79%	94.5%	94.5M	81
EfficientNetB2	36	80.1%	94.9%	94.9M	186

C. Attack Algorithms

Three different attack algorithms were used for the attack. These are respectively FGSM, BIM, and PGD algorithms. The purpose of these attack algorithms is to perturb the input image. But the important thing here is that the perturbed image cannot be distinguished by the human eye. False estimates can also be given by adding noises like Gaussian noise (28), but the human eye can detect these images. Gradient-based algorithms are the most suitable algorithms for calculating minimum noise.

The working principle is briefly as follows. During the training phase, many images are given to the model. A loss is calculated for each input. In backpropagation, the weights of the model are optimized with this calculated loss. The aim is to reduce the loss value of the model to the minimum value. Optimization is done with the gradient of the loss function. We try to reach the slope of the loss function close to zero. The aim is to maximize the loss function in the attack algorithms. For this, the gradient value is taken for each image pixel, and the mark function is used. Gradient values are the direction vector that shows how close they are to the correct class. A certain amount of these gradient values is added to the input image so that it crosses the correct class boundary.

In the FGSM algorithm (1) shown in Equation (1), the amount of perturbation is found in one step. The signed gradient is multiplied by a certain epsilon value and added to the input image. It is an untargeted attack. The goal is to make another class guess.

$$X^{adv} = X + \epsilon sign(\nabla_X J(\theta, X, y)) \quad (1)$$

X is the image sent to the model for prediction. y is the correct label of the image. J is the loss function that calculates the loss of the input. ∇ calculates the gradients of the input according to the loss function. Gradients extracted with the Sign function are signed. The values obtained with the sign function are added to the input image by a certain amount of ϵ. Even if the image seems unchanged when viewed with the human eye, the image has changed mathematically. When the model is asked to predict with the adversarial example produced by the added noise, it maximizes the loss function, and the model predicts the input incorrectly.

In the BIM algorithm shown in Equation (2), the amount of perturbation is found by increasing the epsilon value in each iteration. It is slower than FGSM, but more stable adversarial images can be obtained.

$$X_0^{adv} = X,\ X_0^{adv} = Clipx,\epsilon\{X_N^{adv} + \alpha sign(\nabla_X J(\theta, X_N^{adv}, y))\} \quad (2)$$

The PGD algorithm is the iterative version of the FGSM algorithm. It is a different version of the BIM algorithm. Unlike BIM, it uses random values at each iteration to find the best perturbation.

IV. IMPLEMENTATION

The implementation steps are as in Fig 2 and each step is explained one by one.

TABLE II. TRAINING PARAMETERS

Training Parameters	**Selected Parameters**
Loss Function:	Categorical Cross-Entropy
Optimizer:	Adamax
Learning Rate:	0.001
Batch-size:	30

We selected four different CNN models for model training. The selected models were VGG19, ResNet50V2, Xception, and EfficientNetB2, respectively. Keras library is used. We used transfer learning with models whose weights were pre-trained with image-net. We fine-tuned the dense layer according to the new model. The output layer is set to its new two classes. The model is trained with new data. Training parameters are as in the Table 2.

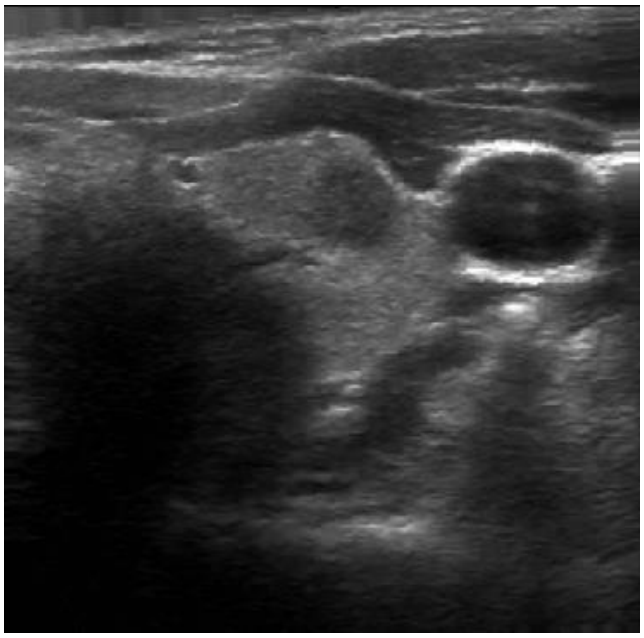


FIG 1. THYROID ULTRASOUND IMAGE

As a result of the training, the test accuracy values were 0.93 for EfficientNetB2, 0.89 for Xception, 0.83 for Resnet50v2, and 0.72 for VGG19. The EfficientNetB2 model with the highest depth achieved the best accuracy. The Xception and ResNet50V2 model have similar parameter numbers and depth values. Although ResNet50V2 is slightly deeper than the Xception model in depth, it did not give a better result. This may be due to different model architectures. ResNetV2 uses deep residual networks and Xception uses depth-wise separable convolutions. VGG19 is the oldest model of all. It has traditional architecture and a large number of parameters. It is also heavy in size, but it is a shallow model. Therefore, we cannot say that it is very successful in complex images.

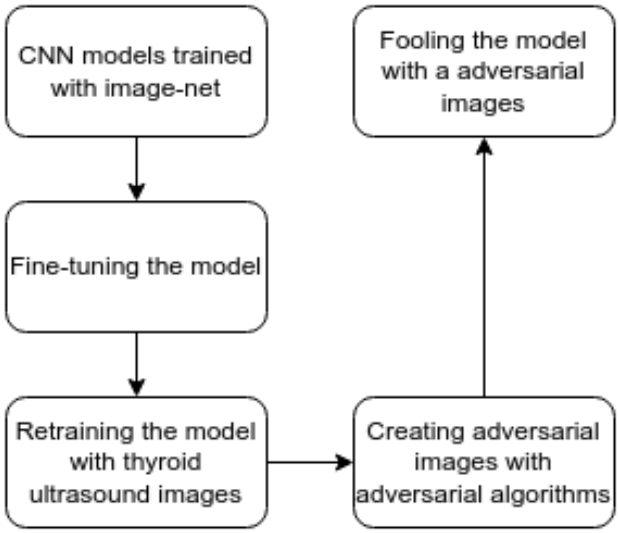


FIG 2. IMPLEMENTATION STEPS

We used the IBM Adversarial Robustness Toolbox library for implementing attacks (30). White-box targeted and white-box untargeted evasion attacks were performed on the trained models. These attacks are FGSM, BIM, and PGD attacks, respectively. The FGSM white box is an untargeted attack. BIM and PGD are white-box targeted attacks. The results are shown in Figs 3-6 and Table 3 and discussed in the next section.

V. RESULT AND DISCUSSION

In the graphs shown in Fig 3-4-5, each line represents the prediction accuracy of a model at different epsilon values. The positive part of the y-axis of the graph indicates that the model is classifying correctly. The corresponding values are the accuracy rate. The negative part of the y-axis of the graph is the false class prediction accuracy of the model deceived by the Adversarial image.

The incorrect prediction accuracy rates of the models at different epsilon values with the FGSM technique are shown in Fig 3. Applying the FGSM attack to EfficientNetB2, Xception, and ResNet50V2 models added 0.01 epsilon-generated noise to the input image. The adversarial images created with the added noise caused the models to predict incorrectly with an accuracy of 99%. Epsilon 0.01 was sufficient for BIM and PGD iterative attacks against these models.

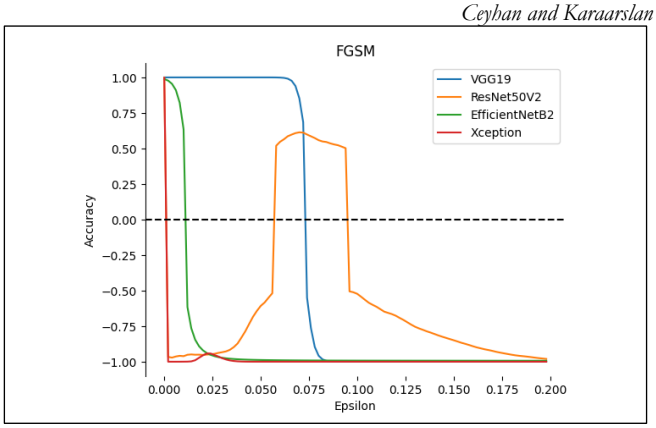


FIG 3. MISCLASSIFICATION ACCURACY RATES OF 100 DIFFERENT EPSILON VALUES BETWEEN 0-0.2 ON FOUR DIFFERENT CNN MODELS USING THE FGSM TECHNIQUE.

As can be seen in Fig. 3, different outputs were obtained in the ResNet50V2 model as a result of the perturbation made by the FGSM technique. The model predicted incorrectly with the epsilon values mentioned earlier and passed the decision boundary. However, we saw that it again came to the correct prediction region in some intervals. The model, which made an incorrect prediction at epsilon values between 0.04 and 0.06, started to make an accurate prediction between 0.06 and 0.08 epsilon values again. It continued to predict incorrectly at 0.08 and higher values. Images were perturbed with epsilon values between 0 and 16 to examine whether models at higher epsilon values obtained a similar result.

As shown in Fig 4, the situation seen for ResNet50V2 was also seen for EfficientNetB2 in the examination with epsilon values of 0-16. In Xception and VGG19, it was observed that the accuracy of false predictions increased in direct proportion to the increase in epsilon.

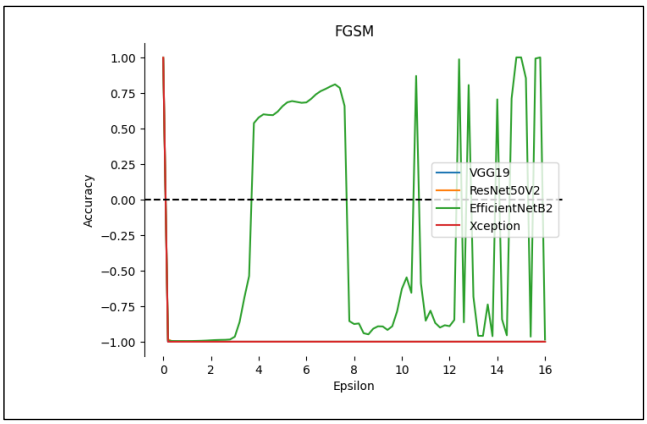


FIG 4. MISCLASSIFICATION ACCURACY RATES OF 100 DIFFERENT EPSILON VALUES BETWEEN 0-16 ON FOUR DIFFERENT CNN MODELS USING THE FGSM TECHNIQUE. (THE BLUE AND ORANGE LINES ARE BELOW THE RED LINE, AS THEY HAVE SIMILAR VALUES.)

As seen in Fig 5, these deviations in FGSM do not exist in BIM and PGD, which are the iterative methods of obtaining adversarial images. The accuracy rate of incorrect estimation increased with the increase in the epsilon value of iterative methods. This experiment proves that iterative methods are more powerful and stable techniques.

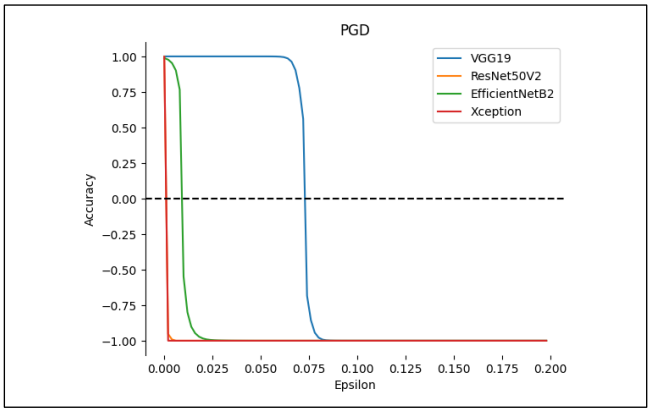


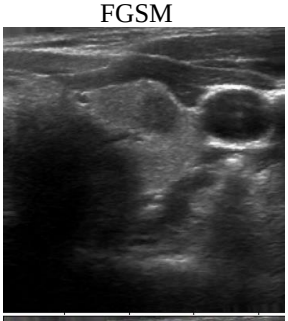
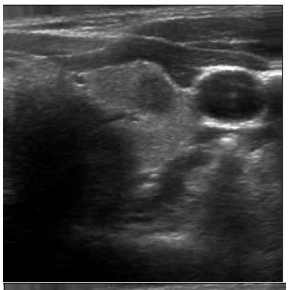
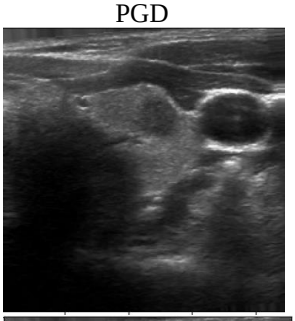
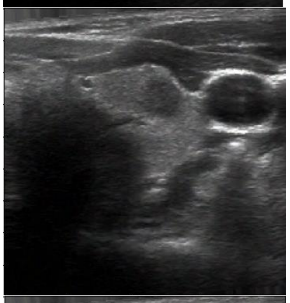
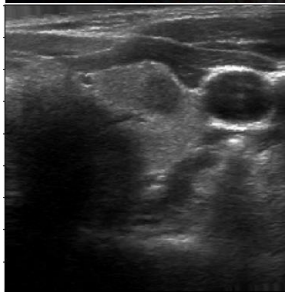
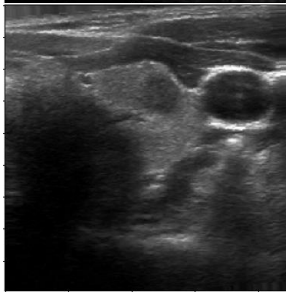
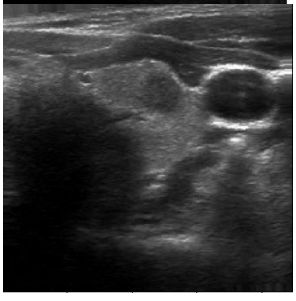
FIG 5. MISCLASSIFICATION ACCURACY RATES OF 100 DIFFERENT EPSILON VALUES BETWEEN 0-0.2 ON FOUR DIFFERENT CNN MODELS USING THE PGD TECHNIQUE. (THE ORANGE LINE IS BELOW THE RED LINE, AS THEY HAVE SIMILAR VALUES.)

Another interesting result is that the VGG19 model, which has the highest number of parameters but the lowest depth, is more difficult to fool than other models. EfficientNetB2 was the second strongest model, which was hard to fool. Xception and ResNet50V2 were the most easily fooled models. These may have different causes, such as depth and model architecture. As depth increases and model architectures change, it can cause blind spots on models to increase.

The images mostly gave good results with different epsilon values. There can always be exceptions. Some of the images can fool different models with different amounts of perturbation.

The creation process of adversarial images is shown in Fig 6, and all the hostile images generated are shown in Table 3. The original image in the table has thyroid disease. When the prediction is made with trained models, cancer can be detected. After adding noise with attack algorithms, %99 percent of non-cancerous prediction was provided. The remarkable point is that the added noises cannot be distinguished by the human eye.

TABLE III. ADVERSARIAL IMAGES PRODUCED WITH DIFFERENT MODELS AND ATTACK ALGORITHMS

	FGSM	BIM	PGD
VGG19			
RESNET50V2			
XCEPTİON			
EFFİCİENTNETB2			

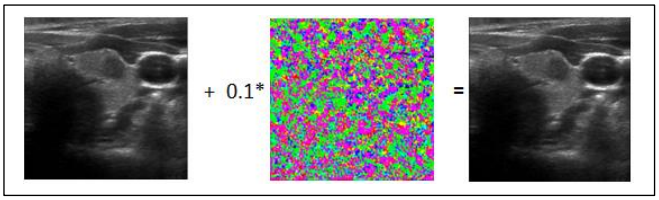


FIG 6. ADVERSARIAL IMAGE GENERATION FROM THE EFFICIENTNETB2 MODEL WITH THE FGSM ALGORITHM. (FORMULA 1)

VI. CONCLUSION

In the study, we tried attack algorithms on thyroid ultrasound images on the widely used models. The models incorrectly predicted with early every perturbed image. Artificial intelligence models trained from thyroid ultrasound images are successfully fooled. The possibility of cheating each model with different amounts of perturbation values is shown. These weaknesses form a problem that needs to be fixed. Artificial intelligence models developed for diagnosis need to be made stronger against attacks. In the continuation of the study, we aim to work on more robust models against these attacks with adversarial learning.

REFERENCES

[1] A. Hosny, C.Parmar, J.Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts "Artificial intelligence in radiology." *Nature Reviews Cancer*, Aug., pp. 500-510, 2018

[2] G. S. Tandel, M. Biswas, O. G. Kakde, A. Tiwari, H. S. Suri, M. Turk et al., "A review on a deep learning perspective in brain cancer classification," *Cancers*, vol. 11, no. 1, p. 111, 2019.

[3] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[4] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, p. 2199, 2017.

[5] F. Abdolali, A. Shahroudnejad, S. Amiri, A. Rakkunedeth Hareendranathan, J. L. Jaremko et al. "A systematic review on the role of artificial intelligence in sonographic diagnosis of thyroid cancer: Past, present and future," *Frontiers in Biomedical Technologies*, 2021.

[6] G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg et al. "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Medical Image Analysis*, vol. 73, p. 102141, 2021.

[7] A. Vatian, N. Gusarova, N. Dobrenko, S. Dudorov, N. Nigmatullin, A. Shalyto et al. "Impact of adversarial examples on the efficiency of interpretation and use of information from high-tech medical images," *2019 24th Conference of Open Innovations Association (FRUCT)*, 2019.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples." 2014 [Online]. Available: http://arxiv.org/abs/1412.6572

[9] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world." 2016 [Online]. Available: http://arxiv.org/abs/1607.02533

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," 2017 [Online]. Available: http://arxiv.org/abs/1706.06083

[11] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016.

[12] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool Deep Neural Networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *2017 IEEE Symposium on Security and Privacy (SP)*, 2017.

[14] J. Su, D. V. Vargas, and S. Kouichi, "One pixel attack for fooling deep neural networks." 2017 [Online]. Available: http://arxiv.org/abs/1710.08864

[15] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial Patch," 2017 [Online]. Available: http://arxiv.org/abs/1712.09665

[16] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[17] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "High-Performance Neural Networks for Visual Object Classification," *CoRR*, vol. abs/1102.0183, 2011 [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1102.html#abs-1102-0183

[18] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," pp. 32--33, 2009 [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[19] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[20] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." 2014 [Online]. Available: http://arxiv.org/abs/1409.1556

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *Computer Vision – ECCV 2016*, pp. 630–645, 2016.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.," in *ICML*, 2019, vol. 97, pp. 6105–6114 [Online]. Available: http://dblp.uni-trier.de/db/conf/icml/icml2019.html

[24] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," *Proceedings of the 2006 ACM Symposium on Information, computer and communications security - ASIACCS '06*, 2006.

[25] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of Machine Learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[26] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. Leung, "A survey on security threats and defensive techniques of Machine Learning: A data driven view," *IEEE Access*, vol. 6, pp. 12103–12117, 2018.

[27] D. T. Nguyen, J. K. Kang, T. D. Pham, G. Batchuluun, and K. R. Park, "Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence," *Sensors*, vol. 20, no. 7, p. 1822, 2020.

[28] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013 [Online]. Available: http://arxiv.org/abs/1312.6199

[29] T. zen, "Thyroid for pretraining," *Kaggle*, 27-Aug-2021. [Online]. Available: https://www.kaggle.com/tingzen/thyroid-for-pretraining. [Accessed: 08-Nov-2022].

[30] IBM, "Adversarial Robustness Toolbox," *Adversarial Robustness Toolbox 1.12.1 documentation*. [Online]. Available: https://adversarial-robustness-toolbox.readthedocs.io/en/latest/index.html. [Accessed: 08-Nov-2022].