➔ Regular Research Paper – NS

Manipulation of Artificial Intelligence in Image Based Data: Adversarial Examples Techniques

Emsal AYNACI ALTINAY
Suleyman Demirel Univeristy, Turkey
emsalaynaci@gmail.com

Utku KOSE
Suleyman Demirel Univeristy, Turkey
utkukose@sdu.edu.tr

Abstract

Artificial intelligence systems are widely used in all fields of life. While the solutions of artificial intelligence have had phenomenal success there is also a dangerous side employing efforts to design attack techniques against Artificial Intelligence and its sub-field, machine learning. Thanks to such techniques, intelligent systems can be fooled to cause misclassified output results. While artificial intelligence builds the future of humanity on intelligent systems, it also has future concerns as various applications from self-driving cars, disease detection to security will be done with autonomous intelligent systems without human beings. Moving from explanations, in this thesis study, artificial intelligence is manipulated by using adversarial examples techniques in image-based data. Adversarial examples are defined as training data that can deceive a machine learning technique into misleading it about the target problem, resulting in a failed or malicious intelligent system. Machine learning models are not robust to adversarial examples. This study shows how artificial intelligence systems are deceived by applying current adversarial example techniques. Obtained results show that the applied techniques provide sufficient attack success rates.

Keywords: *Artificial intelligence, machine learning, adversarial examples, artificial intelligence safety.*

1. INTRODUCTION

The human race's exposure to ever-evolving technology, its relationship with knowledge, and its approach to solving problems have changed its identity and even its sense of self. It should be suspected that there will be many abuses as well as good uses of this artificial intelligence age, where almost all sectors around the world are affected in some way. The concept of Artificial Intelligence (AI) was first used in 1955 as the engineering and science of making intelligent machines. In another definition Artificial intelligence is expressed as the ability of a computer-controlled machine to perform high-level tasks by learning usually based on human reasoning, problem solving, and making sense [1-5]. In short, artificial intelligence is systems that imitate humans and develop themselves iteratively. Artificial intelligence and its-subfield machine learning skills are developing daily and new examples are seen. These technologies have many useful applications ranging from machine translation to medical image analysis [6-9]. Many applications have been developed in these areas and still continue. However, less attention has historically been paid to ways in which AI can be misused [10-14]. With the rapid development of artificial intelligence, security problems have emerged. Adversarial examples pose security concerns as they can be used to attack machine learning systems even if the adversary does not have access to any information about the network structure.

As an attack or hacking method to fool Machine Learning techniques, the concept of adversarial examples was first introduced by Szegedy et al.. In early 2014, Goodfellow et al. showed that minimally modifying the inputs to machine learning models can lead to misclassification. In this context, deep neural networks can be fooled by making tiny modification in input data. Thus, a great speed has been gained in developing defensive and attack techniques that defeat every new defense technique, taking into account adversarial examples. The bad consequences of the existence of adversarial examples for human life cannot be ignored. While artificial intelligence can be used for good, a small mistake or manipulation against artificial intelligence systems can ruin human life. With the continuous advancement of intelligent solution approaches, methods, and techniques, there has always been some debate about the disadvantages and uncertain properties of intelligent systems. Elon Musk, the founder and CEO of the world-famous Tesla Motors and SpaceX companies, also said that a high-level intelligence can be created as a result of the combination of the human brain and artificial intelligence. On the other hand, he demonstrated the dangers of artificial intelligence and warned that people are at risk of being taken over by artificial intelligence in the upcoming years. As a result of the ongoing discussion about whether artificial intelligence is safe or dangerous for humanity in the future, the field of artificial intelligence safety explores the potentially dangerous aspects of artificial intelligence-based systems [15-17]. Artificial intelligence safety is a more complex field, based on a mathematical background and includes adversarial examples that work intensively on machine learning applications in particular.

Based on the explanations, the main aim of this study is to show that AI can be deceived by using adversarial techniques and harm human life when manipulated. Thanks to these techniques, intelligent systems can be fooled into causing directed consequences for failed outputs. It is designed to create a failed or maliciously manipulated intelligent system that can deceive a machine learning technique into misleading about the target problem by applying adversarial example techniques. Thanks to such techniques, AI applications can see things that don't exist and make unexpected mistakes even in real pictures. In order to fool AI applications, slight manipulations are made in the input data and it makes possible for system to make mistakes.

2. RELATED WORK

Recent studies have shown that deep neural networks are not robust to adversarial examples [18-20]. Sharif et al. demonstrated how adversarial examples can fool facial recognition systems [21]. Liu et al. propose new community-based approaches for transferable adversarial examples and show that adversarial examples created using these approaches can successfully attack the http URL, the black box image classification system [22]. Liang et al. proposed an algorithm for generating adversarial text examples using character-level CNN [23]. Hossein et al. showed that adding strategic punctuation marks with selected words can deceive the classifiers [24]. Papernot et al. stated that machine learning models are vulnerable to malicious input and carried out adversarial attack against deep neural network without knowing the internal structure of the model or the training data. They showed that targeted DNNs are misclassified using logistic regression substitutions to generate adversarial examples [25]. Samantha et al. proposed a new method to generate adversarial text examples by replacing the original samples. Experimental results performed by the authors on the IMDB movie review dataset for sentiment analysis and Twitter datasets for gender determination demonstrated the effectiveness of the method [26]. Ebrahimi et al. suggested using the gradient estimation method to rank hostile processes and a greedy search or beam search method to search for adversarial samples [27]. In another study by Carlini and Wagner, it was stated that speech recognition systems can be fooled by vocal adversarial examples [28]. Akhtar and Mian present a comprehensive study of adversarial attacks on deep learning in computer vision [29]. Ilyas et al. also demonstrates that their proposed method is effective against the ImageNet classifier and a targeted attack on a commercial classifier that overcomes limited query access, partial information, and other practical issues to break the Google Cloud Vision API [30]. Su et al. performed low-dimensional attack analysis by changing only one pixel of the images. In this work, the authors propose an innovative model based on differential evolution to generate one pixel of adversarial perturbations [31].

3. MATERIAL METHOD

3.1. One Pixel Attack

A deep neural network simply misclassifies the image by changing only one pixel and predicting the probability of the attack. In most cases, the attacker can cause the predetermined output to get the desired result by the neural network. This type of attack only requires probability values for each category inferred by the neural network. Firstly, a pixel is selected and changed to specific color then adversarial examples are created. Using the evolution algorithm called differential evolution, adversarial examples are created iteratively in order to minimize confidence value or probability of the classification of the neural network. Multiple trials based on pixels that can be changed in an image are made for each selected model. In order to manipulate one or more pixels in the image, the perturbation function is used and the perturbation pixels are divided into coordinates (x, y) and red (R), green (G) and blue (B) values. For the image perturbation function, image is taken as input and a copy of the modified image is made so that each newly identified pixel has RGB color. In this way, the pixels of the selected picture can be changed. For one pixel attack, the input images and the outputs of the model must be known. Several manipulated images are employed on a selected model, and a process is run that returns the probability output of the model's target class, one confidence value per image. This minimizes the confidence value of the correct classification category and distorts the image to maximize the possibilities of all other categories. After that, the most adversarial image that reduces the confidence of the network is output. Confidence value is significantly reduced when the attack is successfully executed. Finally, a new category is obtained that incorrectly classified by the system and this category has the highest classification confidence.

3.2. 3D Adversarial Image Attack

By 3D printing real objects that trick artificial intelligence, an attempt is made to produce three dimensional adversarial examples. Adversarial examples are created by using information specific to their three-dimensional shapes of texture. Adversarial examples that cause targeted misclassification under transformations such as blur, rotate, zoom, or translation are reliably produced. Expectation over Transformation (EOT) algorithm is used. EOT is a meta optimization process that can be combined with various optimization based attacks. In this way, a 2D image is extracted as input and then adversarial examples are created from this image on various rotation angles. For 3D, a simple manipulation on the shape of an object while still distinguishable by the human eye has misclassified the mesh model as something completely different from any viewing angle. 3D traffic input images are employed on different deep network models. Objects are shown to the classifier from different directions. EOT is also tested on 400 2D images, where each image belongs to a different image type and has an average adversarial score of 95%. In 3D simulations with 50 different objects divided into 58 categories. In addition, 3D traffic sign images are output which are classified differently in every aspect.

3.3. Noise Attack

Noise attack is a perceptually imperceptible and inconspicuously visible adversarial example technique with clear potential for malicious use in real life. Noise attack has also been developed in natural ways such as various impulse, contrast, elastic and blurs [32, 33]. Existing adversarial noise generation algorithms are divided into two categories as single perturbation and generalized perturbation. Single image perturbations are cases where a noise vector for each image is learned to fool any DNN classifier. Generalized perturbation is a single noise vector that can be used on multiple images to trick classification. These perturbations exist as noise by nature and are used to trick classification. Noise attack reliably causes machine learning models to misclassify their inputs. Figure 3 shows that how noise attack is applied. The left part contains the original images and the right part contains the modified images by adding carefully crafted noise. Although imperceptible to human eye, the images on the left are correctly classified by DNN classifier, while the images on the right are incorrectly classified. Eventually, deep neural networks are fooled. Worse still, the added seemingly harmless little noise often tricked the model into giving highly reliable predictions. To achieve this, values of the epsilon parameter are tried, which make small but sufficiently effective changes on the data. The epsilon value is chosen to trick the system by making it similar to the real original data.

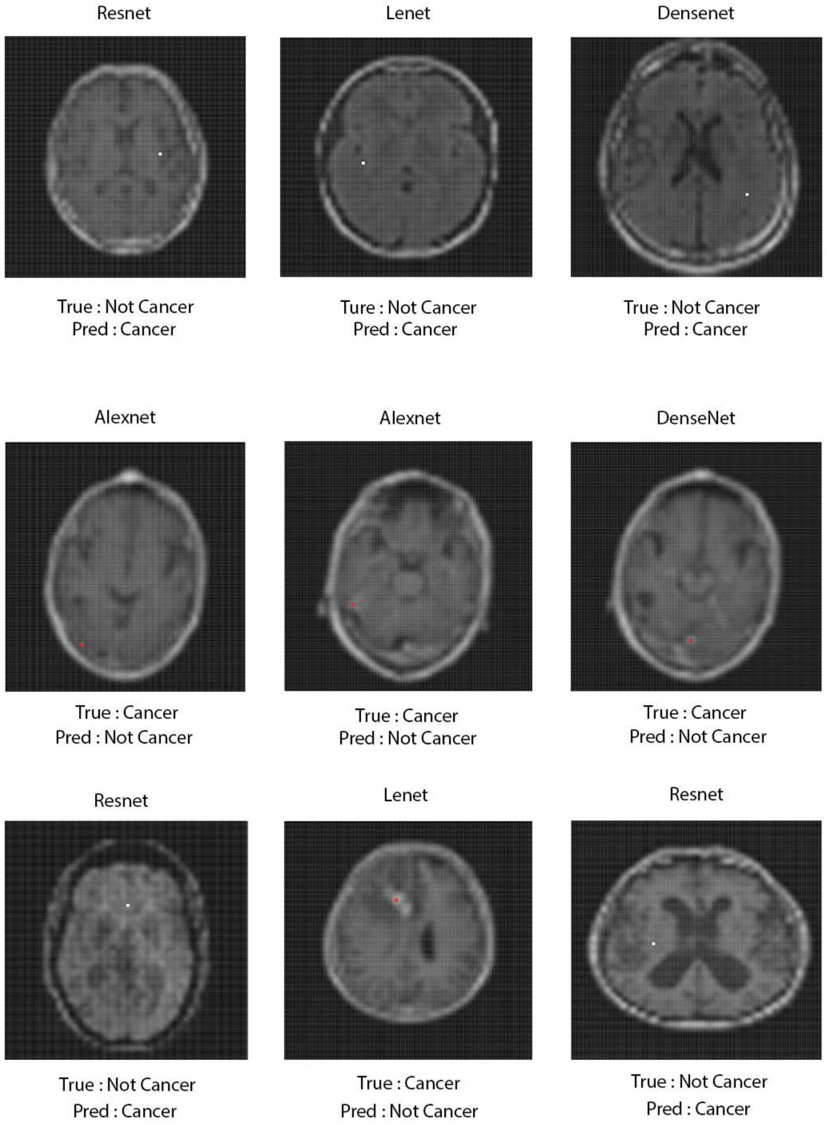

Figure 1. Successful pixel attack images applied on four different models.

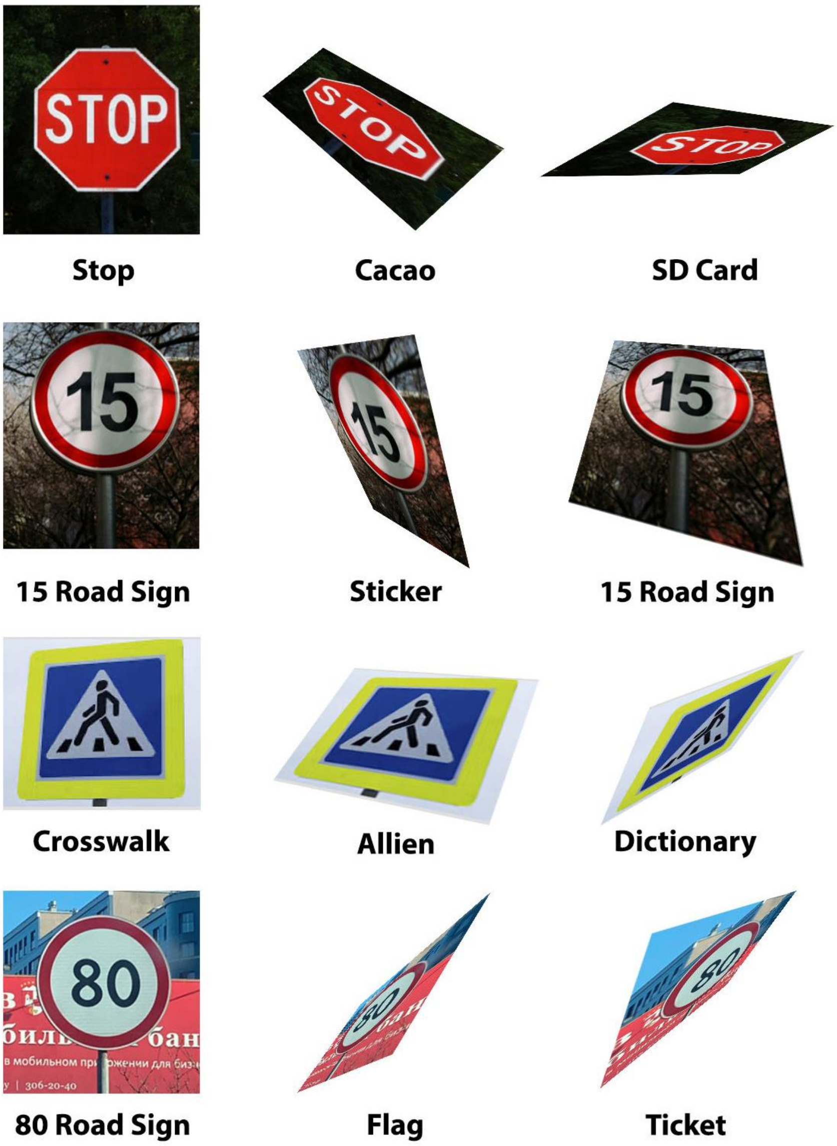

Figure 2. 3D adversarial image attack applied under Transformations.

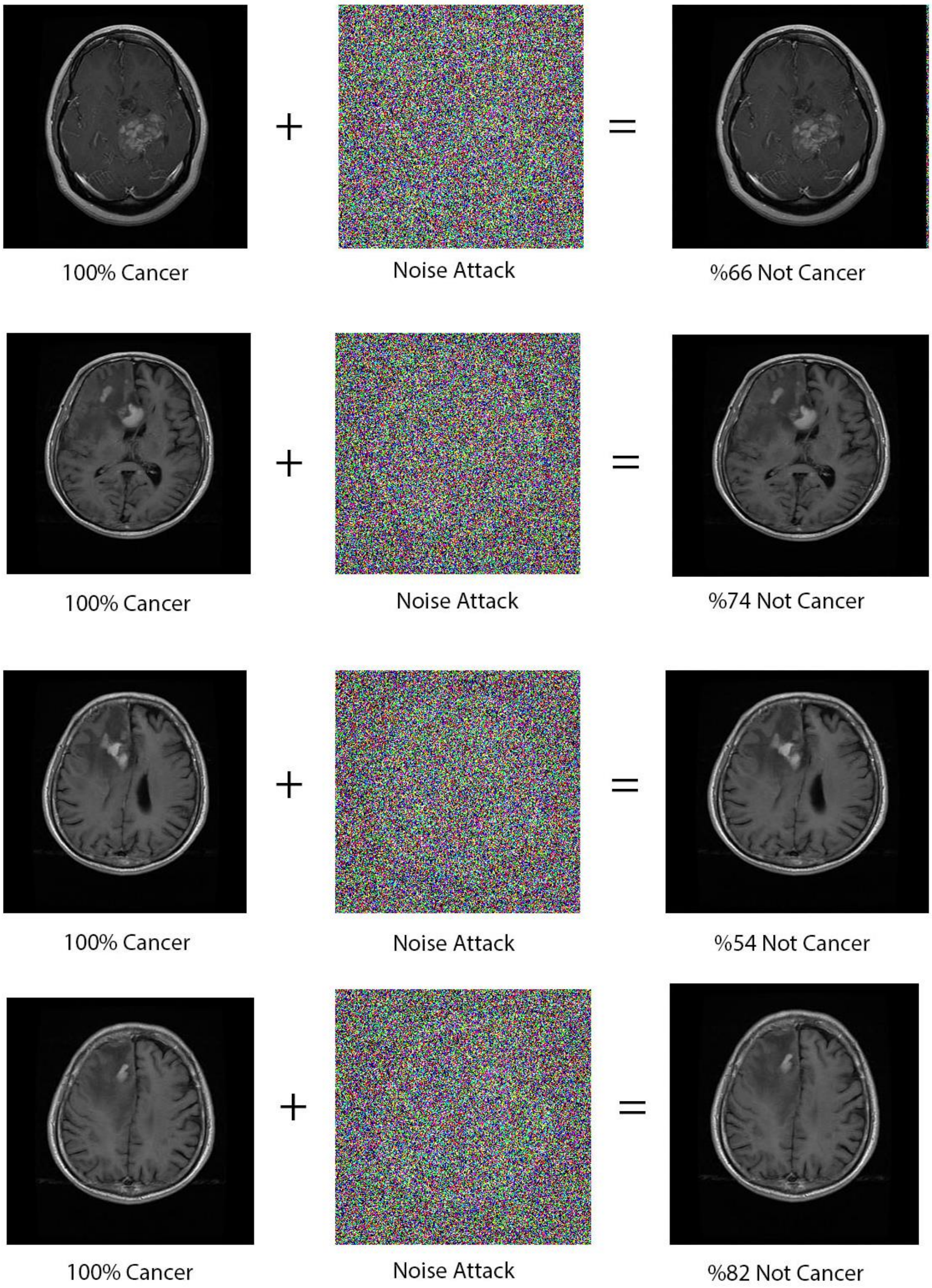

Figure 3. Noise data added to brain tumour images.

3.4. Adversarial Patch

Patches are obtained from random photos using a physical obstacle or algorithm in the photos taken. Rarely, any image with a patch or an unidentified object in the image is a form of attack used. By applying adversarial patch attack, the most practical threat model against computer vision systems in the real world is created. A digital sticker is placed on the object. Thus, machine learning models cannot define the main object and the classifier decides it is another object. This attack technique is different from the previous attacks Because there is no need to know which image is attacked while the attack is being created. At this stage, neural network models tried to trick without the input picture. Patches (stickers) are used in traffic signs and models are misleaded. A patch of any shape can be used. Basically, the intended patch with this process is placed on any image at the specified location, direction, and scale. During training, a random image is selected, patched using a different location, direction and scale each time, and this process is repeated many times. Because only the patch is trained and combined in a new scenario with a different image each time, the learning algorithm is forced to modify the subtraction to make the model predict the wrong class in various scenarios. The EOT algorithm is used to maximize the log probability of a new class, under the constraint that it does not deviate too much from the initial patch. Figure 4 shows the traffic sign images with the adversarial patch attack applied.

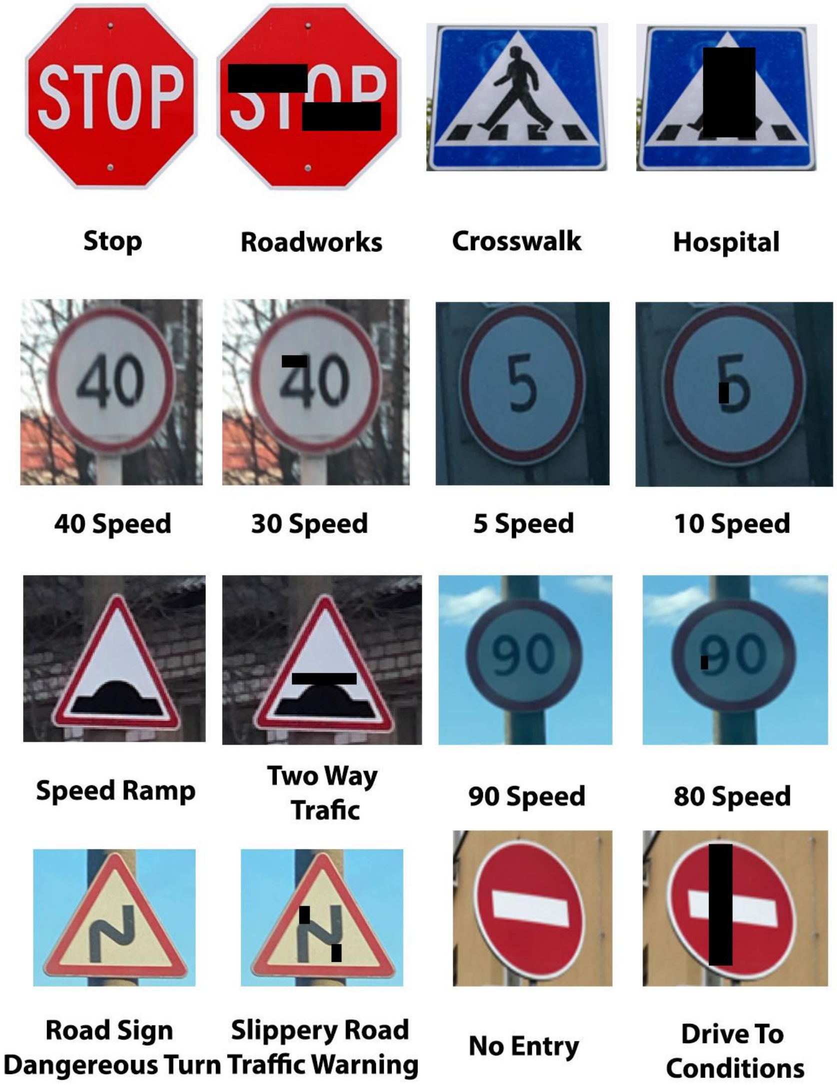

Figure 4. Adversarial patch attack on traffic sign images.

4. RESULTS AND CONCLUSION

Adversarial examples techniques have been applied that defeat the security of machine learning based system. Techniques have been employed to generate adversarial examples so that different types of deep neural network models were misguided in the classification process have a high attack success rate. In addition, optimization algorithms were used to increase the effectiveness of the attacks and perturbations were made on the images with the adversarial examples which are quite new in the literature. In this study, one pixel attack, 3D adversarial image attack, noise attack and adversarial patch techniques were applied on AlexNet, LeNet, DenseNet and ResNet models, respectively. The obtained results were evaluated regarding to the success rate evaluation criterion and the following results were obtained.

Thanks to a one pixel attack, 4 different types of deep neural network models trained on brain tumour dataset were fooled with an average attack success rate of 91%. One pixel attack results are given in Table 1. In 3D adversarial image attack, a 3D traffic sign images were extracted, which are classified as belonging to a different image type from every angle. Tested on 400 2D images with an average adversary score of 95%. With the epsilon value selected in the noise attack method, four different neural network models were 100% tricked in the direction that maximize the loss. Attack success rates are given versus epsilon value in Table 3. Finally, adversarial patches were created to attack four different models on the traffic sign dataset. This effective and powerful attack method has a high attack success rate, reducing the classification accuracy to an average of 17%. Patch attack success rate are given in Table 3.

Table 1. Four pixel attack results on four different network types.

Model	Test Accuracy	Pixel Number	Success Rate
AlexNet	%94	4	95%
LeNet	%90	4	93%
ResNet	%96	4	90%
DenseNet	%95	4	88%

Table 2. Noise attack results on four different models.

Model	Epsilon	Test Accuracy	Success Rate
AlexNet	0.25	16%	100%
LeNet	0.25	18%	100%
ResNet	0.25	20%	100%
DenseNet	0.25	23%	100%

Table 3. Accuracy and attack success rate of adversarial patch.

Model	Accuracy Value	Attack Success Rate
AlexNet	12%	88%
LeNet	15%	85%
ResNet	22%	78%
DenseNet	18%	82%

REFERENCES

[1] Nabiyev, V.V. (2016). Artificial Intelligence: Human-Computer Interaction. Seçkin Publishing, Ankara.

[2] Kose, U., & Koc, D. (Ed.). (2014). Artificial Intelligence applications in distance education. IGI Global.

[3] Pavaloiu, A., & Kose, U. (2017). Ethical artificial intelligence-an open question. arXiv preprint arXiv:1706.03021.

[4] Kose, U. (2017). Development of artificial intelligence based optimization algorithms. PhD. Thesis. Selcuk University, Institute of Natural Sciences, Dept. of Computer Engineering.

[5] Deperlioglu, O., Kose, U., Gupta, D., Khanna, A., & Sangaiah, A. K. (2020). Diagnosis of heart diseases by a secure Internet of Health Things system based on Autoencoder Deep Neural Network. Computer Communications, 162, 31-50.

[6] Neill, Daniel B. (2013). Using Artificial Intelligence to Improve Hospital Inpatient Care. IEE Computer Society.

[7] Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. (2017). Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol. 2017; 69:2657–64.

[8] Walton-Rivers, J., Williams, P.R., Bartle, R., Perez-Liebana D., Lucas, S.M. (2017). Evaluating and modelling hanabi-playing agents, 2017 IEEE Congress on Evolutionary Computation (CEC).

[9] Rajpurkar, P., Irvin, J., Zhu, K. (2017). CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. 05225

[10] Pan, Y.H. (2016). Heading toward artificial intelligence 2.0. Engineering, 2(4), 409–413. http://dx.doi.org/10.1016/J.ENG.2016.04.018.

[11] McFarland, M. (2017). Google uses AI to help diagnose breast cancer, Erişim Tarihi:1.04.2021.http://money.cnn.com/2017/03/03/technology/google-breast-cancer-ai/.

[12] Cheung, C.W., Tsang I.T., Wong, K.H. (2017). Robot Avatar: A Virtual Tourism Robot for People With Disabilities. International Journal of Computer Theory And Engineering, Singapore, (9)3, 229-234.

[13] Johnson, K.W., Soto, J.T., Glicksberg, B.S., Shameer, K., Miotto, R., Ali, M., Ashley, E., Dudley, J.T. (2018). Artificial Intelligence in Cardiology. Journal of the American College of Cardiology, 71(23), 2668-2679.

[14] Park SH & Han K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology; 286:800–809.

[15] Vassev, E. (2016). Safe artificial intelligence and formal methods. In International Symposium on Leveraging Applications of Formal Methods (pp. 704-713). Springer, Cham.

[16] Yampolskiy, R. V. (2016). Taxonomy of pathways to dangerous artificial intelligence. In Workshops at the Thirtieth AAAI Conference on Artificial Intelligence.

[17] Köse, U. (2018). Are we safe enough in the future of artificial intelligence? A discussion on machine ethics and artificial intelligence safety. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 9(2), 184-197.

[18] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. 6199.

[19] Goodfellow, I. J., Shlens, J., Szegedy, C. (2015). Explaining and harnessing adversarial examples. 6572.

[20] Hu, W., Tan, Y. (2017). Generating adversarial malware examples for black-box attacks based on GAN.05983.

[21] Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 24-28 October, Vienna, 1528-1540.

[22] Liu, Y., Chen, X., Liu, C., Song, D., (2016). Delving into transferable adversarial examples and black-box attacks. 02770.

[23] Liang, B., Li, H., Su, M., Bian, P., Li, X., Shi, W. (2017). Deep text classification can be fooled. 08006.

[24] Gümüş, F. (2019). Artificial Intelligence Applications, Effects and Future in Museums. Istanbul University, Institute of Social Sciences, MS Thesis, Istanbul. Hosseini, H., Kannan, S., Zhang, B., Poovendran, R., 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments.08138.

[25] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. (2017). Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia conference on computer and communications security (pp. 506-519). ACM.

[26] Samanta, S., Mehta, S. (2018). Generating Adversarial Text Samples. In Advances in Information Retrieval, Proceedings of the 40th European Conference on Information Retrieval Research, 26–29 March, Grenoble, 744-749.

[27] Ebrahimi, J., Lowd, D., Dou, D. (2018). On Adversarial Examples for Character-Level Neural Machine Translation. 09030.

[28] Carlini, N., Wagner, D. (2018). Audio adversarial examples: Targeted attacks on speech to-text. In 2018 IEEE Security and Privacy Workshops (SPW) (pp. 1-7). IEEE.

[29] Akhtar, N., Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 6, 14410-14430.

[30] Ilyas, A., Engstrom, L., Athalye, A., Lin, J. (2018). Black-box adversarial attacks with limited queries and information. arXiv preprint arXiv:1804.08598.

[31] Su J., Vargas, D.V., Sakurai K. (2017). One pixel attack for fooling deep neural networks, *CoRR*, 08864.

[32] Vasiljevic, I., Chakrabarti, A., & Shakhnarovich, G. (2016). Examining the impact of blur on recognition by convolutional networks. arXiv preprint arXiv:1611.05760.

[33] Zheng, S., Song, Y., Leung, T., & Goodfellow, I. (2016). Improving the robustness of deep neural networks via stability training. In Proceedings of the ieee conference on computer vision and pattern recognition (pp. 4480-4488).

JOMUDE
http://www.jomude.com