

An efficient convolutional neural network for adversarial training against adversarial attack

Srinivas A. Vaddadi¹, Sanjaikanth E. Vadakkethil Somanathan Pillai², Santosh Reddy Addula¹,
Rohith Vallabhaneni¹, Bhuvanesh Ananthan³

¹Department of Information Technology, University of the Cumberland, Williamsburg, United States

²School of Electrical Engineering and Computer Science, University of North Dakota, Grand Forks, United States

³Department of Electrical and Electronics Engineering, PSN College of Engineering and Technology, Tirunelveli, India

Article Info

Article history:

Received Mar 18, 2024

Revised Aug 19, 2024

Accepted Aug 26, 2024

Keywords:

Adversarial instances

Adversarial training models

Convolutional neural network

Deep learning

Defensive mechanism

Image manipulation

ABSTRACT

Convolutional neural networks (CNN) are widely used by researchers due to their extensive advantages over various applications. However, images are highly susceptible to malicious attacks using perturbations that are unrecognized even under human intervention. This causes significant security perils and challenges to CNN-related applications. In this article, an efficient adversarial training model against malevolent attacks is demonstrated. This model is highly robust to black-box malicious examples, it is processed with different malicious samples. Initially, malicious training models like fast gradient descent (FGS), recursive-FGSM (I-FGS), Deep-Fool, and Carlini and Wagner (CW) techniques are utilized that generate adversarial input by means of the CNN acknowledged to the attacker. In the experimentation process, the MNIST dataset comprising 60K and 10K training and testing grey-scale images are utilized. In the experimental section, the adversarial training model reduces the attack accuracy rate (ASR) by an average of 29.2% for different malicious inputs, when preserving the accuracy of 98.9% concerning actual images in the MNIST database. The simulation outcomes show the preeminence of the model against adversarial attacks.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Srinivas A. Vaddadi

Department of Information Technology, University of the Cumberland

6178 College Station Drive, Williamsburg, KY 40769, United States

Email: Vsad93@gmail.com

1. INTRODUCTION

Nowadays, pattern identification and visualization associated with computer science have gained much attention from researchers. Specifically, the convolutional neural network (CNN) obtains better performance in various applications like medical diagnosis, voice identification, and image processing [1]. Nevertheless, these models suffer from increased adversarial vulnerabilities. The security issues are divided into two groups namely probing attack and causal attack. The causal attacks diminish the performance of the CNN model by purposefully integrating adversarial samples at the training stage during the learning process [2]. Some of the common causal attacks are poisoning and backdoor attacks that are manipulated to create a particular effect. The probing attack is one of the evasive attacks that modifies the test data that are already been trained. This type of attack is realistic since there is no definition of the learning data that are being accessed [3]. The adversarial attack is one important example of a probing attack that integrates some noises into the actual data that can be easily identified by humans but by the recent deep learning (DL) models.

In such cases, it could cause errors when they are used in real-time applications like medical imaging, autonomous vehicles, and business applications due to unpredictable outcomes. Several existing studies have been utilized as a defensive method against malevolent attacks and they are of two types. One of these is data manipulation and the other is making DL models more robust over other techniques [4]. The data manipulation process minimizes the attack level with the use of noise from malicious instances by filtering or resizing the data. In contradiction to this, the models of making CNN networks more reliable by utilizing the distillation method, and malicious training process. The distillation approach utilized dual network models to resist the creation of adversarial instances. However, adversarial training schemes are more flexible by jointly scrutinizing the CNN model on malicious samples that are obtained from the traditional DL models [5]. From these, the adversarial training schemes are simpler and more robust compared to conventional studies. Based on traditional techniques, the malicious scrutinization is calibrated to maintain robust than other manual techniques in terms of hands-on defensive mechanisms.

But the conventional malicious scrutinizing model trains the final model with adversarial instances obtained using single attack models [6]. When the final model is scrutinized with adversarial instances that are obtained from several attacks rather than focusing on a single attack, will be more effective against vulnerable malicious attacks. Developing robust malicious perturbations under varying environments with different learning levels on the CNN model has been deliberated in various frameworks [7], [8]. In the same way, the design of a defensive mechanism for increasing the reliability of CNN contrary to malevolent perturbation has undergone two major processes: attack recognition, and attack retrieval methods. The developed study aims to classify the images as normal or malicious by training process and dropout units.

Motivation: the presence of advanced DL models has contemplated the performance above the human intervention level based on image processing tasks. Despite the several models, CNNs are highly subjected to adversarial attacks by injecting perturbation into the actual image. This perturbation is unknown to human perception and distracts the models from providing better performance. Nowadays, adversarial training schemes are playing an integral role in learning vulnerable attacks using practical defensive mechanisms. However, traditional adversarial scrutinizing techniques are much more effective as they can train inputs with single attacks. These kinds of major cons, motivate us to develop a robust DL-based adversarial scrutinizing model to define various malicious attacks accurately. The key contributions of the adversarial training method are conquered below:

- To introduce a diverse adversarial training model in the CNN for generating malicious samples from the actual input.
- To present several adversarial training models like fast gradient descent (FGS), iterative-FGSM (I-FGS), Deep-Fool, and Carlini and Wagner (CW) for creating malicious examples in the CNNs.
- To validate diverse adversarial training models by assessing different measures like accuracy, and attack accuracy rate by comparing them with conventional techniques.

The upcoming sections are organized as depicted below: section 2 outlays the section about related work, section 3 deliberates over the suggested approach, section 4 presents the results and discussion, and section 5 represents the conclusion of the developed framework.

2. RELATED WORKS

Mani *et al.* [9], put forth DL-based models for defending adversarial scrutiny over malicious attacks. Here, the ResNet-based image classification technique was demonstrated and investigated to defend against the GSM adversarial model. For the simulation process, the CIFAR-10 database was utilized in this study. In the resultant part, accuracy was analyzed and distinguished from other studies. However, this method was highly sensitive when scrutinizing with noisy perturbed data and increased the error during the testing process.

Lal *et al.* [10], defined the DL-based defensive adversarial training model for detecting diabetic retinopathy. Here, the defensive method against speckle noise attacks, malicious training, and feature fusion mechanism were analyzed to secure the retinopathy images effectively. Some of the feature extraction techniques like local binary pattern (LBP) and histogram of oriented gradients (HOG) were considered. Moreover, segmentation-based adversarial scrutinizing methods were also investigated. The perturbed data was generated via adversarial models like FGSM, and deep fools were trained on the actual images to obtain malicious instances. In the resultant part, accuracy was analyzed and distinguished from other studies. However, this method faces high black-box issues and reduces the interpretability of the outcome images.

Pal *et al.* [11] established the malicious defensive mechanism using the integrated model on voice samples. Here, the cross-entropy, feature scattering, and marginal losses were jointly used to investigate perturbed samples. Moreover, the deep neural network (DNN) model was utilized to identify the attack that occurred on adversarial training models. The simulation process was demonstrated on the Librispeech dataset

and accuracy-based computational performance was analyzed. Moreover, CW attacks and projected GSM attacks were also considered. However, this method was highly cost-effective and required a more reliable training process for identifying other harmful attacks.

Hashemi, and Mozaffari [12] introduced a CNN-based DL model for training the perturbed data using actual inputs. Here, the CNN with variational autoencoder (AE) model was used as the target that was known to the attacker. Various attacks on gradient-based techniques, distance function optimizer (DFO), transformative models (TM), and preprocessing techniques were investigated. Moreover, uniform noises, salt-pepper noises, gaussian noises, and Poisson noises were also considered for the attack classification process. In the resultant part, accuracy and receiver-operating characteristic curve (ROC) were analyzed and distinguished from other studies. However, harmful CW attacks failed to consider that causes high perturbation that was unknown to the human intervention.

Shen, and Robertson [13] presented a DNN-based network model for training adversarial instances over original inputs. Here, various kinds of malicious training attacks like FGSM, I_FGSM, R_FGSM, and PGD were considered for the training on the DNN model. For the simulation process, the publicly available MNIST, and GTSRB databases were utilized in this study. In the resultant part, success rate, and transfer rate were investigated and distinguished from other studies. But, this model was operative only when processing single attacks like FGSM during the training process.

The degree of access that the “attacker” has to information allows us to classify adversarial assaults. While an attacker in a white-box scenario can see and change the target model's parameters directly [14]-[16], a black-box scenario prevents them from doing so. Thus, in many practical cases, black-box assaults are preferable. Adversarial assaults would render CNNs useless and maybe mislead human doctors. Significantly, this flaw not only creates serious security concerns, but it also hinders the practical use of automated CNN-based systems, which is particularly problematic in the healthcare industry where precise diagnostic findings are critical for patient treatment [17], [18]. Although there is some written material on adversarial machine learning, the vast majority of it deals with real-world photographs [19]. Only a small number of studies have examined medical pictures in depth; nevertheless, those that have done so have shown that medical DL systems are vulnerable to adversarial assaults using either generic or image-specific methods [20], [21]. A large body of prior research links the high dimensionality of training data often comprising images with hundreds or even tens of thousands of pixels to the vulnerability to adversarial attacks [22], [23]. Table 1 provides the several potential research gaps and areas for further exploration.

Table 1. Potential research gaps and areas for further exploration

Area	Research gap	Opportunity
Robustness across diverse datasets	Current research may focus on specific datasets, lacking a comprehensive evaluation across diverse and real-world datasets.	Explore the robustness of the proposed CNN across various datasets, including those from different domains such as healthcare, finance, and autonomous driving, to ensure its effectiveness in diverse scenarios.
Efficiency and scalability	While the model is termed “efficient,” detailed analysis of its computational efficiency, scalability, and performance on large-scale datasets may be limited.	Conduct thorough evaluations of the model's computational requirements and scalability. Explore optimizations like model compression, pruning, and quantization to improve efficiency without sacrificing robustness.
Generalization to new attack types	Adversarial training methods might be tailored to specific types of adversarial attacks, potentially limiting their effectiveness against new or unseen attack strategies.	Develop adaptive adversarial training techniques that can generalize to a wide range of attack types, including novel and sophisticated attacks. Investigate the use of meta-learning and transfer learning to enhance generalization.
Explainability and interpretability	The interpretability of CNNs, especially when adversarial training is involved, remains a significant challenge. Understanding how and why the model resists adversarial attacks is crucial for trust and deployment.	Implement techniques to improve the interpretability and explainability of the adversarially trained CNN. Use visualization methods and model-agnostic explainability techniques to provide insights into the model's decision-making process.
Real-time application and deployment	Research may lack a focus on the practical deployment of adversarially trained CNNs in real-time systems where latency and resource constraints are critical.	Investigate the deployment challenges of adversarially trained models in real-time applications. Explore hardware accelerations, edge computing, and real-time inference optimizations to facilitate the practical use of these models in real-world environments.

2.1. Problem statement

It is encompassed that various issues have been observed like high time complexity, poor training, and focus mainly on a single vulnerable attack. Nowadays, DL models are highly effective in classifying adversarial attacks caused by different attack strategies. The DL models like CNN, DNN, and recurrent

neural network (RNN) models are considered as the alternative solution that provides fascinating performance over antagonistic attacks. Hence, this study introduced the known CNN model to classify multiple adversarial training attacks using malicious inputs. To our knowledge, the developed method outperforms a better performance by solving various issues caused by conventional schemes.

3. DEVELOPED METHOD

In this article, an efficient adversarial training model against malevolent attacks is demonstrated. This model is highly robust to black-box malicious examples, it is processed with different malicious samples. Initially, malicious training models like FGS, I-FGS, Deep-Fool, and CW techniques are utilized that generate adversarial input using the CNN model known to the attacker.

3.1. Adversarial training method

The considered adversarial scrutinized scheme has two phases: the creation of malicious cases, and how the scheme acquires these inputs. Initially, it produces several malevolent inputs and includes them as supplementary data to the CNN approach for enhancing the robustness against malicious attacks. The adversarial training models generate several adversarial instances via FGSM, R-FGSM, Deep-Fool, and CW techniques employing the CNN that is acknowledged to the assaulter. Based on this procedure, the reliability of the final approach over malicious occurrences can be enhanced.

The maneuver function of the CNN X_l is represented as, $F_l(u)$ and it is processed with the actual training database. Assume the CNN model X_l , the actual training data $u \in U$, with its equivalent classes $v \in V$, and final classes $v' \in V$, then the optimization issues are tackled by generating adversarial instances u' can be mathematically as (1).

$$u' : \underset{u'}{\operatorname{argmin}} l(u, u') \text{ s.t. } f_l(u') = v' \tag{1}$$

Here, $l(u, u')$ defines the distance between the actual instances u and perturbed instances u' , $\underset{u}{\operatorname{argmin}} f_l(u)$ represents that $f(u)$ converts to minimal based on the value of u . $f_l(u')$ indicates the CNN model that identifies the input value. For creating these u' , every adversarial instance is defined using FGSM, I-FGSM, Deep-Fool, and CW schemes.

3.1.1. FGSM scheme

This method can generate u' by utilizing X_∞ and it can be mathematically formulated as (2).

$$u' = u + \beta \times \operatorname{sign}(\Delta \operatorname{loss}_{f,k}(u)) \tag{2}$$

Here, f and k represents the obtained class and method process function respectively. Using the normal instance u , the gradient descent is improved by β parameter. While u' is generated by optimizers. This process is simple and provides better performance. Figure 1 illustrates the workflow of the adversarial scrutinizing model.

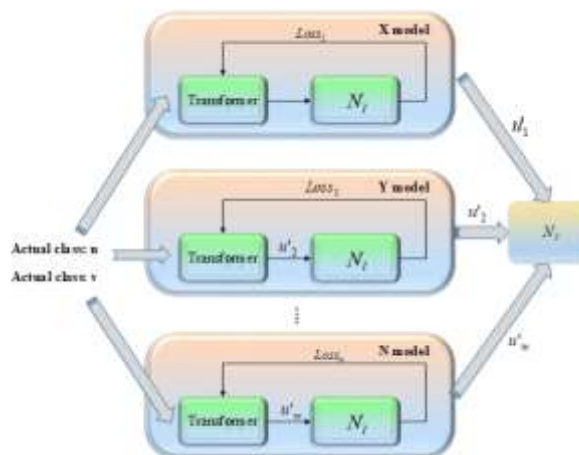


Figure 1. Workflow of the adversarial scrutinizing model

3.1.2. I_FGSM scheme

This model is the extension of the FGSM technique. Here, a smaller amount λ is eventually changed and clipped instead of changing β in each step. It can be mathematically interpreted as (3).

$$u'_a = u'_{a-1} - clip_{\beta} \left(\lambda \times sign \left(\Delta loss_{f,k}(u) \right) \right) \tag{3}$$

During each iteration, the I_FGSM generates an adversarial instance on a CNN model. When compared with FGSM, the I_FGSM technique has better prevention performance based on white-box attacks.

3.1.3. Deep-fool scheme

This approach generates an antagonistic instance with minimal falsification from the actual illustration. It creates u' via the linear approximation and due to the nonlinear DL models, this technique is more complex than FGSM.

3.1.4. Deep-fool scheme

This model is the Carlini attack that creates 100% adversarial examples by using various objective functions. It can be expressed as given (4).

$$D(u, u') + p \times f(u') \tag{4}$$

This model evaluates the corresponding binary p value to define a greater accuracy rate on attack instance. Moreover, it regulates the bout accuracy rate at the cost of increased falsification by fine-tuning self-assurance rate as depicted (5).

$$f(u') = max[\{Y_f(u')_a : a \neq c\}] - \{Y_f(u')_c\} \tag{5}$$

Here, c indicates the actual class, and $\{Y_f(\cdot)\}$ manipulates the preliminary SoftMax classification vector outcome. The adversarial instances obtained by every model are combined at the training set to further train the CNN model N_T . The process that takes place in the CNN model N_T is represented as, $f_k(u)$. The CNN model N_T is initially trained with the actual training database. Assume the adversarial instance $u' \in U$, the actual class $c \in C$ and the CNN classes $c' \in C$ with their particular label as the actual class c are formulated as given (6).

$$f_k(u') = c \tag{6}$$

Using this procedure, the CNN is accomplished with different antagonistic cases and finally, the reliability contrary to unwanted antagonistic instances is enhanced. Algorithm 1 indicates the antagonistic training progression.

Algorithm 1. Pseudocode for adversarial training method

```

Start:
Initialize various parameters like Actual training dataset  $u' \in U$ , Actual class  $v \in V$ ,
validation data  $k$ , CNN model  $N_l$ , malicious training models like fast gradient descent (FGS),
recursive-FGSM (I-FGS), Deep-Fool, and Carlini-Wagner (CW) technique.
Adversarial training models: ( $N_l, u$ , FGS, I_FGS, Deep-fool, CW)
 $u' \leftarrow$  Creation of adversarial cases ( $N_l, u$ , FGS)
 $u' \leftarrow$  Creation of adversarial cases ( $N_l, u$ , FGS)
 $u' \leftarrow$  Creation of adversarial cases ( $N_l, u$ , FGS)
 $u' \leftarrow$  Creation of adversarial cases ( $N_l, u$ , FGS)
Scrutinize the final model  $N_T \leftarrow (U, V) + (u', v)$ 
Store the accuracy of the final CNN model  $N_T(k)$ 
Return  $N_T$ 
Stop
    
```

4. RESULTS AND DISCUSSION

The developed method is processed and simulated by the MNIST [24] based handwritten representative database is utilized in this study. It consists of numerical digits that vary from 0 to 9 in the form of greyscale images. Here, a total of 60K training and 10K testing images are present. The images present in the database are in the size 28x28. Some existing models: without a baseline, and with baseline methods [25] are compared with the developed adversarial training model to prove its efficiency.

4.1. Assessment metrics

Attack success rate measures how effective an adversarial attack is in fooling a model by quantifying the proportion of successful attacks. Number of successful attacks represents how many adversarial attacks were able to fool the machine learning model, causing it to misclassify or behave unexpectedly. Accuracy is used to evaluate the overall performance of a classification model by determining the proportion of correctly classified instances (both positive and negative) out of the total number of instances.

$$\text{Attacksuccessrate}(\%) = \left(\frac{\text{Numberofsuccessfulattacks}}{\text{Totalattack}} \right) \times 100 \quad (7)$$

$$\text{Accuracy} = \frac{w+x}{w+x+y+z} \quad (8)$$

Here, w , x , y , and z indicates the true negative (TN), true positive(TP), false negative (FN), and false positive (FP) respectively.

4.2. Simulation analysis of developed method over conventional schemes

In this section, the performance achieved by the developed method is analyzed via graphical interpretation. Here, the attack success rate and accuracy are analyzed and compared with traditional malicious scrutinizing models. The detailed analysis of the obtained performance is depicted below. Figure 2 illustrates the attack success rate analysis over generated adversarial instances with different techniques.

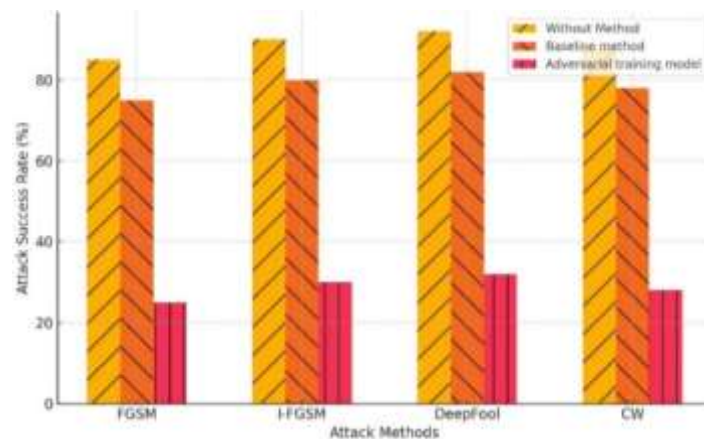


Figure 2. Attack success rate (ASR) analysis over generated adversarial instances with different techniques

While considering the without model as the final model does not perform any defensive mechanism over adversarial scrutinizing model. While considering the baseline model as a final model for training the adversarial samples, it can use only single type of attack model namely FGSM. While considering the developed training model as the final model, it can create various adversarial instances using FGS, I-FGS, Deep-Fool, and CW techniques. The attack success rate for the without method misclassified larger than 89.88% of the malicious instances. Though, the developed model minimizes the ASR to 32.88% and the mean computation is investigated to improve more than 44.79% distinguished from the baseline model. Hence, it is clear that the CNN model (final model) is highly vigorous against malicious attacks. Figure 3 depicts the accuracy analysis on testing with MNIST database. Even though, the additional adversarial instances are trained, the CNN model still remains the similar accuracy concerning the actual data. It is noted that the considered adversarial training scheme achieved almost similar performance when compared with the without model and baseline model during the testing process.

4.3. Outcome analysis

In this section, the outcomes achieved by different techniques for the CNN constructed on the MNIST database are analyzed. Figure 4 shows the outcomes obtained from each instance. Figure 4(a) to 4(e) indicates the original image, FGSM, I-FGSM, Deep-fool, and CW methods respectively. It is noted that FGSM, I-FGSM, and Deep-fool shows greater distortions compared to the CW model.

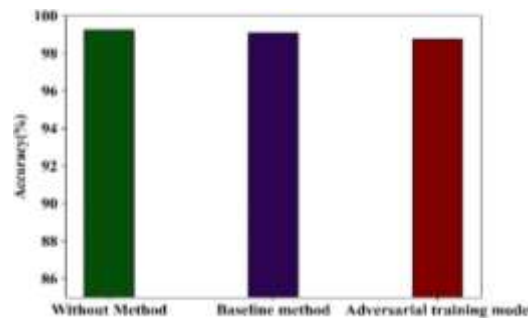


Figure 3. Accuracy analysis on testing with MNIST database

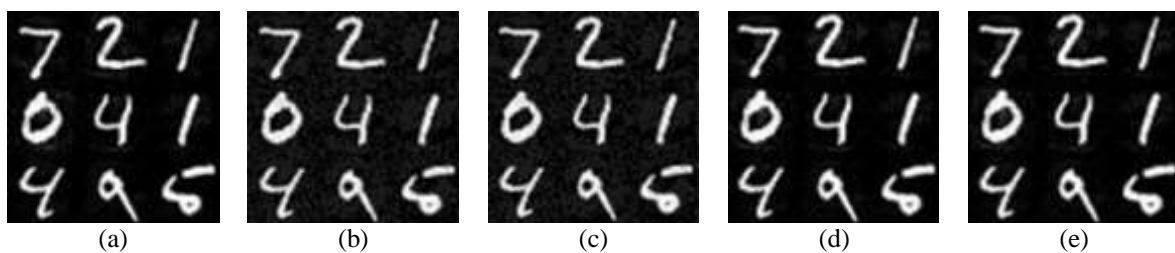


Figure 4. Outcomes obtained from each instance: (a) original image, (b) FGSM, (c) I-FGSM, (d) deep-fool, and (e) CW

5. CONCLUSION

The developed method demonstrated and investigated a robust adversarial training scheme for malicious classification attacks. In this study, the malicious instances are initially created using various approaches such as asI-FGSM, Deep-Fool, FGSM, and CW and then train the CNN model making it more reliable against unknown malevolent attacks. Outcomes obtained by these models are distorted to various degrees. It is concluded that among I-FGSM, Deep-Fool, FGSM, and CW models, the CW technique creates adversarial instances with lesser distortion. Moreover, the Adam optimizer (AO) is utilized to minimize the network complexities by considering cross entropy as the objective function. The experiments are carried out in the Python platform using the publicly available MNIST database. In the experimental section, the adversarial training model reduces the ASR by an average of 27.19% for different malicious inputs, while preserving the accuracy of 98.89% concerning actual images in the MNIST database. The adversarial scrutinizing model is highly useful in autonomous vehicular applications, and medical applications. However, there are many other adversarial scrutinizing mechanisms and it is failed to consider in this study. In future studies, more adversarial scrutinizing mechanisms will be introduced by considering larger datasets, and their performance is analyzed.

ACKNOWLEDGEMENTS

The Author with a deep sense of gratitude would thank the supervisor for his guidance and constant support rendered during this research.





REFERENCES

- [1] B. Chen, J. Yin, S. Chen, B. Chen, and X. Liu, "An adaptive model ensemble adversarial attack for boosting adversarial transferability," *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4489-4498. 2023, doi: 10.1109/iccv51070.2023.00414.
- [2] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25-45, 2021, doi: 10.1049/cit2.12028.
- [3] X. Jia, Y. Zhang, B. Wu, K. Ma, J. Wang, and X. Cao, "LAS-AT: adversarial training with learnable attack strategy," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13398-13408. 2022, doi: 10.1109/cvpr52688.2022.01304.
- [4] Y. Xiao, and C. M. Pun, "Improving adversarial attacks on deep neural networks via constricted gradient-based perturbations," *Information Sciences*, vol. 571, pp. 104-132, 2021, doi: 10.1016/j.ins.2021.04.033.
- [5] C. Zhang, X. Costa-Perez, and P. Patras, "Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms," *IEEE/ACM Transactions on Networking*, vol. 30, no. 3, pp. 1294-1311, 2022, doi: 10.1109/TNET.2021.3137084.





- [6] J. Gu, H. Zhao, V. Tresp, and P. H. Torr, "SEGP GD: an effective and efficient adversarial attack for evaluating and boosting segmentation robustness," *In European Conference on Computer Vision, Cham: Springer Nature Switzerland*, pp. 308-325, 2022, doi: 10.1007/978-3-031-19818-2_18.
- [7] M. R. Vemparala *et al.*, "Breakingbed: Breaking binary and efficient deep neural networks by adversarial attacks," *In Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys)*, Springer International Publishing, pp. 148-167, 2022, doi: 10.1007/978-3-030-82193-710.
- [8] S. Addepalli, and S. Jain, "Efficient and effective augmentation strategy for adversarial training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1488-1501, 2022, doi: 10.4018/ijswis.297038.
- [9] N. Mani, M. Moh, and T. S. Moh, "Defending deep learning models against adversarial attacks," *International Journal of Software Science and Computational Intelligence (IJSSCI)*, vol. 13, no. 1, pp. 72-89, 2021, doi: 0.4018/IJSSCI.2021010105.
- [10] S. Lal *et al.*, "Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition," *Sensors*, vol. 21, no. 11, pp. 3922, 2021, doi: 10.3390/s21113922.
- [11] M. Pal, A. Jati, R. Peri, C. C. Hsu, W. AbdAlmageed, and S. Narayanan, "Adversarial defense for deep speaker recognition using hybrid adversarial training," *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6164-6168, 2021, doi: 10.1109/ICASSP39728.2021.9414843.
- [12] A. S. Hashemi, and S. Mozaffari, "CNN adversarial attack mitigation using perturbed samples training," *Multimedia Tools and Applications*, vol. 80, pp. 22077-22095, 2021, doi: 10.1007/s11042-020-10379-6.
- [13] J. Shen, and N. Robertson, "BBAS: towards large scale effective ensemble adversarial attacks against deep neural network learning," *Information Sciences*, vol. 569, pp. 469-478, 2021, doi: 10.1016/j.ins.2020.11.026.
- [14] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 665-681.
- [15] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognition*, vol. 107, 2020, Art. no. 107332.
- [16] X. Li and D. Zhu, "Robust detection of adversarial attacks on medical images," in *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1154-1158.
- [17] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: investigating medical imaging networks using adversarial examples," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018, pp. 493-501.
- [18] 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.
- [19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.
- [20] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath, "Deep learning models for electrocardiograms are susceptible to adversarial attack," *Nature Medicine*, vol. 26, pp. 1-4, 2020.
- [21] T. K. Yoo and J. Y. Choi, "Outcomes of adversarial attacks on deep learning models for ophthalmology imaging domains," *JAMA Ophthalmology*, vol. 138, no. 11, pp. 1213-1215, 2020.
- [22] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: a survey," *arXiv preprint arXiv:1810.00069*, 2018.
- [23] C. Xie, Y. Wu, L. van der Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 501-509.
- [24] H. Khodabakhsh "MNIST dataset," *Kaggle*, 2018, <https://www.kaggle.com/datasets/hojjatk/mnist-dataset>
- [25] H. Kwon, and J. Lee, "Diversity adversarial training against adversarial attack on deep neural networks," *Symmetry*, vol. 13, no. 3, pp. 428, 2021, doi: 10.3390/sym13030428.

BIOGRAPHIES OF AUTHORS







Srinivas A. Vaddadi     is a dynamic and forward-thinking professional in the field of Cloud and DevSecOps. With a solid educational foundation in computer science, he embarked on a journey of continuous learning and professional growth. Their relentless pursuit of knowledge and commitment to staying at the forefront of industry advancements has earned them recognition as a thought leader in the Cloud and DevSecOps space. He can be contacted at email: Vsad93@gmail.com.







Sanjaikanth E. Vadakkethil Somanathan Pillai     holds an MS in Software Engineering from The University of Texas at Austin, Texas, USA, and a BE from the University of Calicut, Kerala, India. Currently pursuing a Ph.D. in Computer Science at the University of North Dakota, Grand Forks, North Dakota, USA, his research spans diverse areas such as mobile networks, network security, privacy, location-based services, and misinformation detection. He is a proud member of Sigma Xi, The Scientific Research Honor Society, underlining his commitment to advancing scientific knowledge and research excellence. He can be contacted at email: s.evadakkethil@und.edu.







Santosh Reddy Addula     is a Senior Member of the IEEE, holds a Master of Science in Information Technology from the University of the Cumberland in Kentucky, USA. With extensive experience in the IT industry, he has demonstrated expertise across multiple domains. Santosh is an innovator with a strong portfolio of patents and has significantly contributed to academic research through his articles as an author and co-author. Additionally, he serves as a reviewer for esteemed journals, reflecting his dedication to advancing knowledge and ensuring the quality of scholarly publications in his field. He can be contacted at email: santoshaddulait@gmail.com.



Rohith Vallabhaneni     is a dedicated worker with a strong work ethic in leading teams to solve organizational issues. He is capable of learning all aspects of information within a company and using the technical knowledge and business background to effectively analyze security measures to determine their effectiveness in order to strengthen the overall security posture. He has great work ethic and outstanding team leadership skills and seek to accomplish organizational goals, while growing in knowledge and experience. He can be contacted at email: rohit.vallabhaneni.2222@gmail.com.



Bhuvanesh Ananthan     received the B.E. degree in Electrical and Electronics Engineering from Anna University in 2012, M.Tech. in Power System Engineering from Kalasalingam University in 2014 and Ph.D. degree from Faculty of Electrical Engineering of Anna University in 2019. He has published more than 65 papers in reputed international journals, 25 papers in international conferences and 10 books. He is a life time member of International Society for Research and Development, International Association of Engineers. He can be contacted at email: bhuvanesh.ananthan@gmail.com.