Online at: https://researchtrendsjournal.com **ISSN No:** 2584-282X

INTERNATIONAL JOURNAL OF TRENDS IN EMERGING RESEARCH AND DEVELOPMENT

Volume 2; Issue 4; 2024; Page No. 109-113

Received: 18-04-2024
Accepted: 29-05-2024

Wavelet transform based defending mechanism against adversarial iris attacks

[1]Mahalle Sheetal Anil and [2]Dr. Kaushal Kumar

[1]Research Scholar, Sunrise University, Alwar, Rajasthan, India
[2]Assistant Professor, Sunrise University, Alwar, Rajasthan, India

DOI: https://doi.org/10.5281/zenodo.14618701

Corresponding Author: Mahalle Sheetal Anil

Abstract

This study explores the role of innovative green technologies in enhancing customer engagement, focusing on how businesses can leverage environmentally friendly practices to build stronger connections with their customers. With increasing awareness of environmental issues, customers are increasingly drawn to brands that demonstrate a commitment to sustainability. The findings reveal that customers not only appreciate but actively seek out brands that prioritize environmental responsibility, leading to increased customer satisfaction, brand loyalty, and advocacy. Moreover, the adoption of green technologies is shown to enhance the overall brand image and create a competitive advantage in the market. The research concludes that innovative green technologies are not only beneficial for the environment but also serve as a powerful tool for customer engagement. Companies that embrace these technologies are better positioned to meet the evolving expectations of eco-conscious consumers, thereby driving long-term business success.

Keywords: Wavelet, transform, based, mechanism, against, adversarial and iris attacks

Introduction

Because of its effectiveness in object identification, picture classification, semantic segmentation, and object tracking, Convolutional Neural Networks (CNNs) see extensive application in crucial domains including autonomous driving and face recognition. Yet, new studies have shown that adversarial assaults may exploit Convolutional Neural Networks (CNNs) by manipulating pictures in very subtle ways that humans would have a hard time detecting.

An example of an antagonistic picture is one that is disturbed. When used in production, hostile examples may severely damage real-time supervised learning models. Therefore, it is an important yet difficult issue to make CNNs more resistant to hostile cases. Approaches that are model-specific and those that are model-agnostic have been defined as two types of defence against adversarial instances. In order to standardize the parameters of a certain model, model-specific mechanisms are created. Both adversarial training and parameter smoothing may be used as training methods for these devices. Since model-specific approaches rely on deep learning, many think that defensive mechanisms based on deep networks are the only solution to the problem of deep learning systems' susceptibility.

Defenses tailored to individual models, on the other hand, are more effective against targeted assaults. However, due to the need of individually training a model to withstand each potential assault, these methods are time-consuming. This becomes much more difficult when dealing with massive datasets. One need just computes the gradient of a trained defence model to launch an assault. Methods that are independent of the model in question pre-process the input data using tools such as High-level representation Guided Denoiser (HGD), JPEG compression, random scaling, etc., in order to remove or minimise adverse perturbations. Modifying the classification model's structure or retraining it for each and every kind of attack is not necessary when using model-agnostic approaches. Input reconstruction is one kind of defence that falls under the umbrella of model agnostic strategies. To properly categorise hostile cases, it is necessary to first turn them into benign ones. Several picture

transform methods, including Discrete Sine Transform and Discrete Wavelet Transform, are used to replicate the malicious instances. The classification models then make use of these reconstructed pictures.

Literature Review

Nathalie Baracaldo *et al*. (2018) [1] Machine learning (ML) in the cloud and on the edge is becoming more important due to the usage of learned models in several Internet of Things (IoT) applications, such as industrial automation and environmental sensing. Nevertheless, there are distinct security concerns associated with ML in IoT settings. Specifically, by interfering with the measurements taken by sensors, attackers might alter the training data. Poisoning attacks create targeted misclassification or improper behaviour, implant "backdoors" and "neural trojans," and drastically reduce overall performance. We provide an approach that is based on newly created tamper-free provenance frameworks and employs contextual information about the training set's data points' origin and transformation to detect harmful data. Using a reliable test data set is not necessary for our method to function. With the suggested method and trustworthy provenance data, poisoning threats in IoT settings may be identified and prevented.

Deepak Upreti *et al*. (2023) [2] The significance of data mining in industrial engineering has grown as the value of processing power and storage capacity has risen. The field of industrial engineering has recently seen remarkable progress because to AI and ML. One method of machine learning that aims to solve the problem of data privacy in distributed computing systems and their data storage applications is federated learning. We have examined the effectiveness of Tolpegin's proposed defense technique and extended the work of Tolpegin *et al*. about data poisoning concerns in federated learning systems. Following that, we evaluated the efficacy of several clustering methods, such as K-means, PCA, KPCA, and UMAP. In comparison to PCA, KPCA, and K-means, UMAP provides better performance in avoiding data-poisoning attacks, according to the results.

The focus of this study is on developing methods to make machine learning models, and image classification algorithms in particular, more resistant to malicious assaults. Improving the models' robustness is the main goal, and evolutionary algorithms are used to optimise bit plane slicing configurations. Findings show that 5-bit depth representation models are more resilient, with 98.21% FGSM attack accuracies and 92.98% Deep Fool assault accuracies. The significance of bit plane slicing for altering detail levels to preserve algorithmic integrity under adversarial situations is shown by these findings. A large recovery was found, demonstrating the usefulness of the optimised defence tactics, even if performance dropped significantly owing to hostile alterations, with accuracy decreasing from 90.32% to 11.69%. Findings like this lend credence to the concept that sophisticated defensive mechanisms based on genetic algorithms and dynamic bit plane slicing need further research to make machine learning models more resilient to hostile assaults that are always evolving.

Aleksander Madry *et al*. (2018) [3] The susceptibility of neural networks to adversarial instances, or inputs that closely resemble real data but are wrongly identified by the network, has been recently shown in research. We tackle this issue by applying robust optimisation principles to the study of neural networks' adversarial resilience. This method gives us a holistic perspective on a lot of previous research on the subject. Our ability to discover trustworthy and, to some extent, universal approaches to training and attacking neural networks is further enhanced by its principled character. In instance, they provide a specific security promise that would shield against a clearly defined group of enemies. With these techniques, we can teach networks to be far more resilient against many forms of adversarial assaults. A natural security guarantee, according to them, should be robustness against a first-order opponent. We believe that building fully resistant deep learning models requires first making them immune to assaults from these particular categories of adversaries.

Wanman Li, *et al*. (2022) [4] Researchers in the field of adversarial machine learning have sought solutions to the problem of hostile samples, which is encountered by many machine learning systems. Because Support Vector Machines (SVMs) are so useful and widely used, this article will first explain how an evasion attack might affect SVM classification, and then it will provide a way to protect against it. Evasion attacks use support vector machines' (SVMs') classification surfaces to repeatedly identify small perturbations that fool nonlinear classifiers. Our proposed vulnerability function is a special case for evaluating the SVM classifiers' susceptibility to attack. In order to protect SVMs with Gaussian kernels against evasion attacks, we propose a kernel optimization-based defence method that makes use of this vulnerability function. It turns out that our defence strategy works well on benchmark datasets, and our kernel optimisation methodology makes the SVM classifier much more reliable.

Experimental Evaluations

In the defensive strategies that we have discussed, we do not modify the classification model; rather, our goal is to predict the adversarial iris data by removing changes in the input data. With the help of very small perturbations, adversarial Iris images can be created. Using wavelet transformation, we will break down the modified iris picture into its component wavelet components in the next step. We investigate the middle and high frequency band wavelet characteristics. Afterwards, the U-Net model is used to de-noise the adversarial samples and re-create the picture as the original. Figure 1 shows the general layout of the suggested goal.

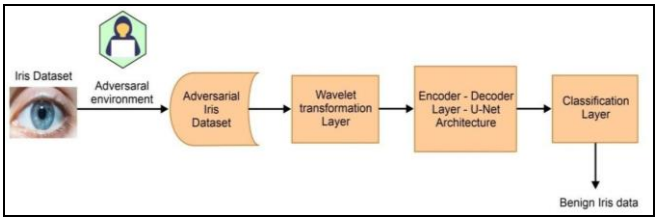

Fig 1: Proposed Architecture Diagram

Prior to being fed into the U-Net model, the adversarial iris input pictures undergo preprocessing and scaling. The output is sent to integrated layers, which combine

convolution and robust normalization, from the encoder's three layers-convolutional¡ strong normalization, and dropout layers. A convolutional layer receives the output and uses it to reconstruct the picture while filtering out any unwanted noise. The U-net architecture makes use of the significant and useful procedure known as robust normalisation. When compared to other Normalization methods, it generates good accuracy on common datasets. The adversarial distortions introduced by the changed example must be eliminated, hence the classifier must be designed to minimize reconstruction mistakes. Every batch's reconstruction error is calculated using Equation (4.14). In Equation 4.14, $x^{(i)}$ is an actual input and $\tilde{x}^{(i)}$ reconstructed input.

$$reconstruction_{err} = \left\| \tilde{x}_i^{(i)} - x_i^{(i)} \right\|_2$$

Fig 2: Shows just one instance of an encoder layer.

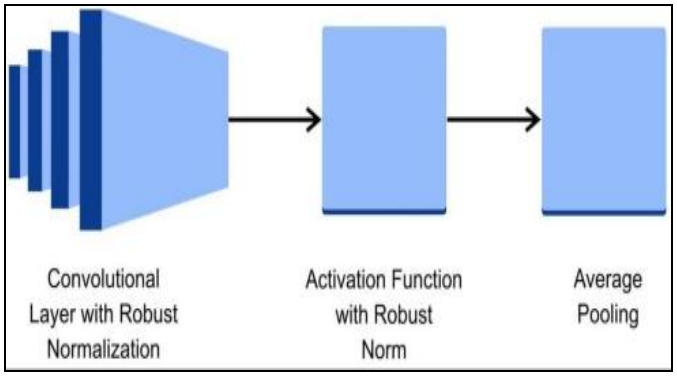

Fig 3: Integrated encoder layer

Dataset Description

This suggested study makes use of two publicly accessible benchmark iris datasets. I. Databases of IITD iris images Section ii) CASIA-Iris-Interval. The images in these datasets are captured under various conditions such as pupil dilation, occlusion of the eyelids/eyelashes, minor shadow of the eyelids, and so on. Both datasets' specifications are summarized in Table 1.

Utilizing the adversarial dataset that has been produced by the Deepfool, Fast Gradient Sign Method (FGSM), and iterative Gradient Sign Method (iGSM) techniques, the experiment is subsequently carried out. The details of the adversarial examples that result from the adversarial attacks on the IITD iris data are presented in Table 2.

Table 1: Dataset details

Dataset	Number of subjects	Number of images	Images Size	Image format	Number of classes
IITD	224	1120	320x240	BMP	224
Casia Iris Interval	249	2639	320x280	JPEG	395

Table 2: Adversarial Dataset

Adversarial attack name	Total number of Images	Number of Labels
Fast Gradient sign method	50000	1000
Iterative Gradient Sign Method	20000	1000
Deep fool	21080	1000

Results

For the development of adversarial iris examples, perturbations are included in the IITD Dataset. Our goal in using Fast Discrete Wavelet Transform, or FDWT, is to eliminate the pictures' high frequency components. Using U-Net model, denoised versions of the iris can be recreated. By reducing the reconstruction error, the significant iris image features are restored. The DNN based Iris recognition model uses this set of reconstructed images as its training dataset. Table 3 illustrates the model's performance both prior to and following the attack. Before the FGSM attack, the Deep CNN model classified IITD's iris dataset with 98 percent accuracy, but afterward it fell to 90.24 percent accuracy. When an iGSM attack is carried out, the level of accuracy falls from 97 percent to 86 percent. The accuracy is decreased to 93 percent while using the Deepfool attack, whereas it was 98 percent before the attack. Research has shown that classification models' accuracy takes a major hit when adversarial methods like FGSM, iGSM, and Deepfool are used. Figure 4 shows the model's accuracy.

Table 3: Performance of the classification model – accuracy

Performance measure	Attack 1	Attack 2	Attack 3
Original accuracy	98	97	98
Accuracy after attack	90	86	93

Attack 1 – FGSM, Attack 2- iGSM, Attack 3- Deepfool

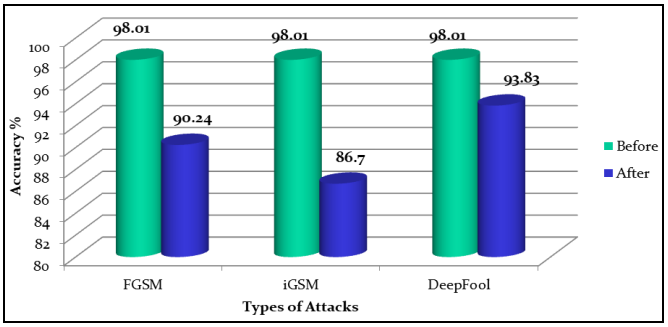

Fig 4: Model's accuracy

Table 4 compares the suggested model to the state-of-the-art model as of right now. The other classification metrics also included in the same table. To measure how well the model worked, we utilize metrics like F1score, Precision, and Recall. Compared to the state-of-the-art models, the proposed technique performs better in terms of accuracy and performance. The first five defensive techniques listed in the Table 4 are based on adversarial training. The denoising process is used in the last two procedures, number 6 and 7, respectively.

Table 4: Comparison of proposed defensive technique with other techniques

S. No	Defensive techniques	Attack 1	Attack 2	Attack 3	Attack 1	Attack 2	Attack 3	Attack 1	Attack 2	Attack 3	Attack 1	Attack 2	Attack 3
		Precision			Recall			Accuracy			F1 Score		
1	Defensive Technique –1	35	30	42	37	32	45	39	34	45.5	0.36	0.31	0.43
2	Defensive Technique –2	33	31	40	36	34	41	38	35	44	0.34	0.32	0.40
3	Defensive Technique –3	38	38	51	38	40	55	40	42	57	0.38	0.39	0.53
4	Defensive Technique –4	42	42	46	44	45	49	45	48	51	0.43	0.43	0.47
5	Defensive Technique –5	52	51	55	54	52	60	57	53	61	0.53	0.51	0.57
6	Defensive Technique –6	80	71	80	80	76	82	82	78	84	0.80	0.73	0.81
7	Our Method	88	82	91	90	85	92	92	86.9	95	0.89	0.83	0.91

Attack 1- FGSM, Attack 2- iGSM, Attack 3- Deepfool

Defensive Technique –1: Adversarial Training, Defensive Technique –2: Ensemble Adversarial Training, Defensive Technique –3: Function Transformation, Defensive Technique -4: robust deep learning model, Defensive Technique –5: two- pronged defense, Defensive Technique –6: Autoencoder. +wavelet

The sixth method in Table 4 denoises adversarial images using a single auto encoder without considering the frequency level. Our defense method analyses the middle and low frequency wavelet components. When denoising images, the U-net architecture is used. The suggested approach provides better results based on this approach. a visual representation of the performance analysis of several adversarial assaults with respect to different performance metrics.

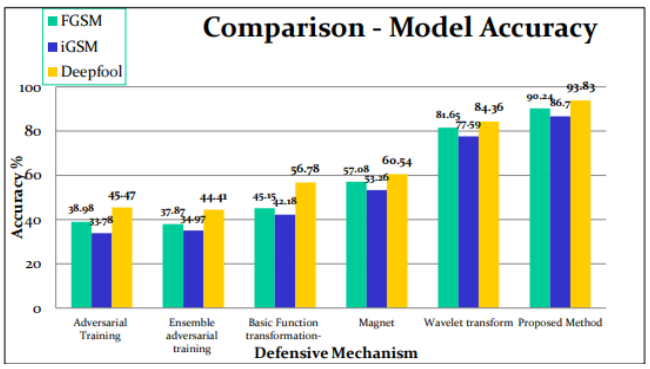

Fig 5: Comparison–Proposed Model accuracy with state of art method on various adversarial attacks

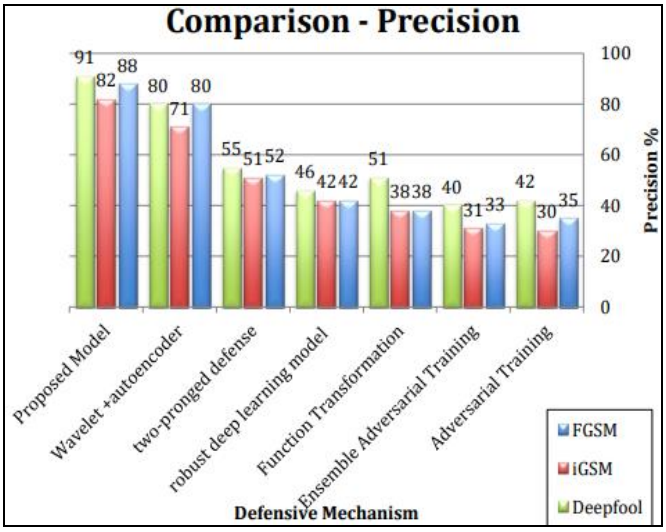

Fig 6: Comparison–Proposed Model precision with state of art method on various adversarial attacks

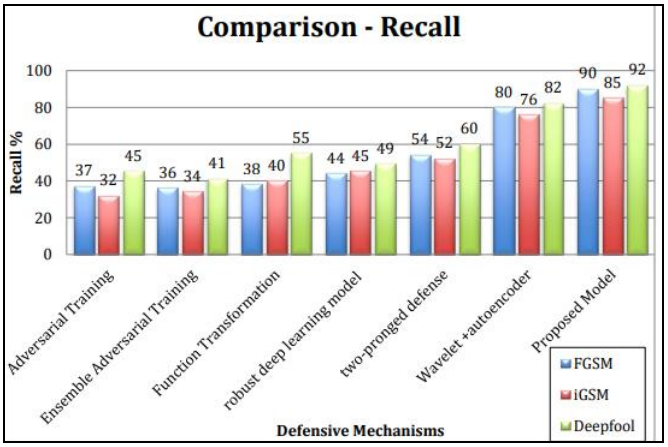

Fig 7: Comparison –Proposed Model Recall with state of art method on various adversarial attacks

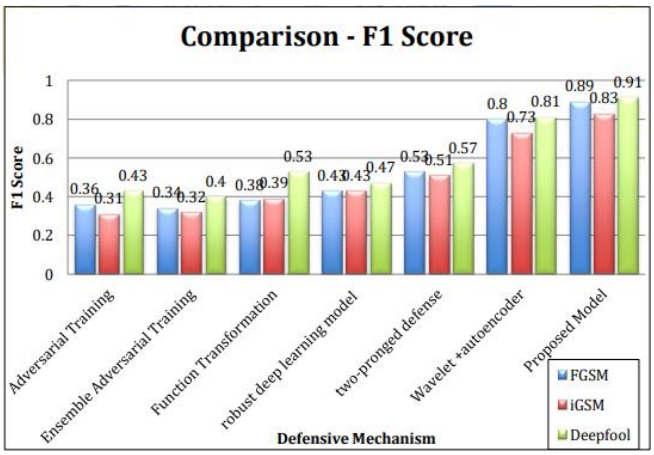

Fig 8: Comparison–Proposed Model F1 score with state of art method on various adversarial attacks

Conclusions

Several adversarial assaults, including as FGSM, Deepfool, and iGSM, are used to evaluate the proposed method for classifying hostile iris photos. With an average accuracy rate of 94%, the IITD generates noteworthy results as a standard iris image database. Following iris reconstruction, a pre-trained Convolutional Neural Network model (VGG 16) is used to extract key features, which are further classed using Multiclass Statistical Analysis. The categorization is carried out by use of a Support Vector Machine that has been trained using the Particle Swarm Optimization approach (PSO-SVM). Experiments conducted on the industry-standard IITD iris dataset often provide statistically significant results with a 95.8% success rate. the Optimised-Curie algorithm has been tried and proven to safeguard the

SVM classifier, it ought to be feasible to use the same approach to safeguard additional classifiers like Multilayer Perceptron and Naïve Base classifier, among others. The IITD iris dataset is used to test the U-Wavenet classifier and the Curve -CNN-PSVM model. For future evaluations on other iris datasets and biometric identification tasks, we intend to use several pre-trained CNN models. It is possible to test the proposed system with different adversarial tactics in order to build a more flexible defense framework.

References

1. Baracaldo N, Chen B, Ludwig H, Safavi A, Zhang R. Detecting Poisoning Attacks on Machine Learning in IoT Environments. In: Proceedings of the IEEE International Conference on Internet of Things (ICIOT); c2018. doi:10.1109/ICIOT.2018.00015.
2. Upreti D, *et al*. Defending Against Label-Flipping Attacks in Federated Learning Systems with UMAP. c2023. DOI: https://doi.org/10.21203/rs.3.rs-1984301/v1.
3. Madry A, *et al*. Towards Deep Learning Models Resistant To Adversarial Attacks. In: Proceedings of the International Conference on Learning Representations (ICLR); c2018.
4. Li W, *et al*. Kernel-based Adversarial Attacks and Defenses on Support Vector Classification. Digital Communications and Networks. 2022;8(6):492-497. doi:10.1016/j.dcan.2022.01.008.
5. Vassilev A. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. 2023. DOI: https://doi.org/10.6028/NIST.AI.100-2e2023.
6. Steinhardt J, *et al*. Certified Defenses for Data Poisoning Attacks. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS); c2017.
7. Sadeghi K, Banerjee A, Gupta SKS. A System-Driven Taxonomy of Attacks and Defenses in Adversarial Machine Learning. IEEE Transactions on Emerging Topics in Computational Intelligence. 2020 Aug;4(4):450-467. doi:10.1109/TETCI.2020.2968933.
8. Damiani E, *et al*. Data Partitioning and Compensation Techniques for Secure Training of Machine Learning Models. Academic Year 2020/2021.
9. Meenakshi K, Maragatham G. A Self Supervised Defending Mechanism Against Adversarial Iris Attacks Based on Wavelet Transform. International Journal of Advanced Computer Science and Applications. 2021;12(2):1-6. doi:10.14569/IJACSA.2021.0120270.
10. Yuan X, *et al*. Adversarial Examples: Attacks and Defenses for Deep Learning. arXiv:1712.07107v3 [cs.LG]. 7 Jul 2018.
11. Jmila H, Ibn Khedher M. Adversarial Machine Learning for Network Intrusion Detection: A Comparative Study. Computer Networks. 2022;214:109073. doi:10.1016/j.comnet.2022.109073.
12. Chen X, Li S, Huang H. Adversarial Attack and Defense on Deep Neural Network-Based Voice Processing Systems: An Overview. Applied Sciences. 2021;11(18):8450. doi:10.3390/app11188450.
13. Li S, Wang J, Wang Y, Zhou G, Zhao Y. EIFDAA: Evaluation of an IDS with Function-Discarding Adversarial Attacks in the IIoT. Heliyon. 2023;9(2):e13520. doi:10.1016/j.heliyon.2023.e13520.
14. Zhu Y, Wang M, Yin X, Zhang J, Meijering E, Hu J. Deep Learning in Diverse Intelligent Sensor-Based Systems. Sensors. 2023;23(1):62. doi:10.3390/s23010062.