IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125186

eXplainable and reliable againstadversarial machine learning

Prof. Bhavya R A, Gopika T S, Anusha J

Department of CSE, SJC Institute of Technology, Chickballapur, India

Abstract— Machine learning models are increasingly being integrated into critical decision-making processes across various domains. However, these models are susceptible to adversarial attacks, where malicious actors deliberately manipulate input data to deceive the models and induce incorrect predictions. In this paper, we present an overview of state-of-the-art techniques that aim to enhance the explainability and reliability of machine learning models in the face of adversarial attacks. We begin by discussing the fundamental concepts and motivations behind adversarial machine learning, emphasizing the need for models that can provide explanations for their predictions while maintaining robustness.

Keywords— Explainability,Reliability,Adversarial attacksRobustness.

I. INTRODUCTION

Machine learning algorithms have achieved remarkable success in various domains, including image recognition, natural language processing, and decision-making systems. However, their vulnerability to adversarial attacks poses a significant challenge to their reliability and trustworthiness. Adversarial attacks involve manipulating input data in a subtle manner to deceive the machine learningmodel and induce incorrect predictions.

We explore explainability techniques that enable machine learning models to provide transparent and interpretable insights into their decision-making processes. These techniques include feature importance analysis, rule extraction, and attention mechanisms, which help users understand the factors influencing model predictions. By providing explanations, these models increase user trust and allow for the identification of potential vulnerabilities.

A. eXplainable AI:

The main motivation behind Explainable AI is to address the "black box" nature of complex machine learning models. Explainable AI techniques aim to bridge the gap between the internal workings of machine learning models and human understanding. By providing explanations, XAI enables users to gain insights into the factors influencing model predictions, understand the rationale behind decisions, detect potential biases or vulnerabilities, and ensure fairness and accountability in AI systems. Relaiable AI:

Reliable AI focuses on developing machine learning models and algorithms that consistently produce accurate and dependable results in a wide range of scenarios. By ensuring trust, safety, consistency, and resilience, reliable AI enables the adoption of AI technologies in critical domains and enhances their overall effectiveness and impact. Material handling.

B. Adversarial Machine Learning:

The main objective of adversarial attacks is to cause the machine learning model to make incorrect predictions or decisions, leading to potential security breaches, privacy violations, or system failures. Adversarial attacks can be categorized into different types, such as evasion attacks, where adversaries manipulate input samples to evade detection or mislead the model, and poisoning attacks, where adversaries inject malicious data during the training phase to compromise the model's performance.

II. ADVERSARIAL ATTACKS CONSIDERED

A. Fast gradient sign method:

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125186

In FGSM, the goal is to add imperceptible perturbations to the input data in order to cause misclassification by the targeted model. The attack utilizes the gradients of the model's loss function with respect to the input data to determine the direction in which the perturbations should be added. The key idea is to take a single step in the direction of the gradient, but instead of using the actual gradient, the signs of the gradientare used to control the direction.

Step by step overview

- Compute the gradient of the model's loss functionwith respect to the input data.
- Determine the sign of the gradient at each input datapoint.
- Multiply the sign of the gradient by a small epsilonvalue to control the magnitude of the perturbation.
- Add the scaled perturbation to the original input data tocreate the adversarial example.
- The resulting adversarial example is slightly perturbed, but visually similar to the original input. However, it can cause the targeted model to misclassify the example.

B. Jacobian based saliency map

JSMA specifically targets models that use a neural network architecture, and its goal is to find the most salient features in the input that, when modified, will cause misclassification by the targeted model. The attack utilizes the Jacobian matrix, which represents the partial derivatives of the output class probabilities with respect to the input features.

Step by step overview

- Compute the Jacobian matrix for the targeted modelwith respect to the input features.
- Select a target class to which you want to misclassify the input example.
- Initialize a saliency map, which represents the importance of each input feature in influencing the target class probability.
- While the target class probability is below a certain threshold:

a. Update the saliency map by increasing the saliency of the features that have the highest positive entries in the Jacobian matrix.

b. Update the saliency map by decreasing the saliency of the features that have the highest negative entries in the Jacobian matrix.

- Modify the input features corresponding to thehighest saliency values to create the adversarial example.
- Repeat steps 4 and 5 until the target class probability exceeds the threshold or a maximum number of features have been modified.

C. Carlini wagner

The CW attack is designed to be more powerful and effective compared to previous attack methods, as it introduces a novel formulation that takes into account both the misclassification objective and the imperceptibility constraint. The attack is formulated as an optimization problem, where the goal is to minimize the perturbation while maximizing the misclassification.

One of the key features of the CW attack is its flexibility in handling different types of models, including deep neural networks. It does not rely on specific knowledge of the model architecture or parameters, making it applicable toblack-box scenarios where only input-output access to the model is available.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125186

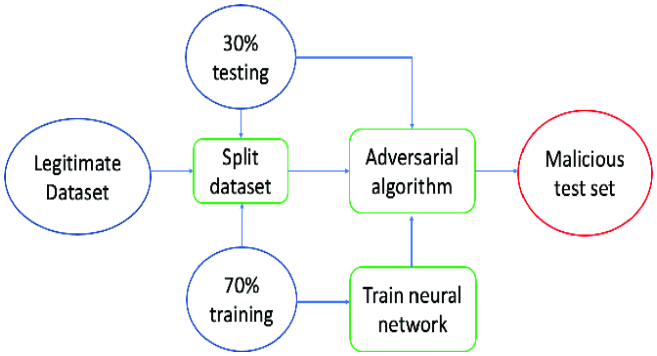

Fig: concept approach to executeadversarial machine learning attack.

III. WORK FLOW

1. Logic learning machine:

LLM is a rule-based supervised method. Inputs: dataset X ; number of features N^{FR};
Candidate perturbations $\Delta = (\delta 1, \ldots, \delta NFR), \delta j = (\delta sj, \delta tj), j = 1, \ldots, N^{FR.}$

Algorithm:

1. Apply LLM on X ;
2. Select features fj from feature ranking
3. Find [sj, tj] from value ranking
4. Define logical OR: $I = Uj=1^{N\ FR}$ [sj, tj];
5. Find hyper-rectangle: $P(\Delta) = Uj=1^{N\ FR}$ [sj ± δsj · sj, tj ± δtj ·tj];
6. Find optimal perturbations Δ*.

1.a Reliability from Outside

As suggested by its name (‘‘ from outside’’), this method aims at finding the adversarial regions based on the opposite class to the target (which is the adversarial class, y=1, in our case). Hence, our focus is here on the LLM for the legitimate class, denoted with y=0. Start with performing the feature/value ranking base on legitimate class (y=0) .
Find Δ* as: Δ*=arg min Δ:N0=D0 V(P (Δ)).

1.b Reliability from Inside

This method performs the same search for adversarial regions with a conceptually similar approach to the previous method, except for that it starts with N FR most important features for the adversarial class, which is our target in this case. Start withthe attack class (y=1) feature/value ranking.
Find Δ* as: Δ*=arg max Δ:N0=0 V(P (Δ)).

1.c LLM with 0% error

Differently from the two previous methods, in this algorithm we are interested in joining a number of entire rules, instead ofsingle intervals, thus giving rise to a new predictorr̂ with morecomplex geometry. Our goal is always to have zero false positive rate (FPR=0):).

2. SAFE SVDD

It is a machine learning technique that is used for single-class classification and outlier detection. The idea of SVDD is tofind a set of support vectors that defines a boundary around data.

2.a zeroFPRSVDD :

The zeroFPRSVDD algorithm performs successive iterationsof the SVDD on the target initial region, found with a preliminary SVDD, until there are no more negative pointsinside it. The convergence is achieved when a fixed number ofiterations is reached or when the condition on FPR is satisfied. **Algoithm**:

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125186

dataset X×Y is divided in training set Xtr×Ytr and test setXts×Yts. A threshold of ε is set

1. SVDD-cross-validation on Xtr×Ytr
2. [a,R2]=SVDD(Xtr,Ytr,C−1,C+1, σ)
3. Test SVDD on Xts×Yts
4. maxiter=1000;5. i=1;
6. while (i<maxiter)

6.1 .Xtri,Ytri = 4(Xts,Yts);
6.2. SVDD-cross-validation on Xtri×Ytri
6.3. [ai, R2i]=SVDD(Xtri,Ytri,C−1,C+1, σ)
6.4. Test SVDD on Xts×Yts
6.5. if(FPR< ε)
6.5.1. return [a∗, R∗2] = [ai, R2i];
6.5. end7 . i = i + 1;

End

2.b Explainable SVDD:

It refers to an extension of the standard SVDD algorithm that incorporates interpretability into the model. It allows the user to gain insight into how the model makes decisions and which features are important in identifying anomalies.

Algorithm :

Get a ∗ , R ∗ from ZeroFPRSVDD algorithm. Fix ε.

1. Sample uniformly a new dataset Xnew s.t. xi ∈ Xnew

⇐⇒ | ||xi − a||2 − R 2 | < ε

2. Classify Xnew in Ynew through optimalZeroFPRSVDD (w.r.t. [a ∗ , R ∗2])
3. Solve a classification problem via LLM w.r.t.[Xnew,Ynew]
4. The LLM rules defines an explained ZeroFPRSVDDregion R
5. return :

IV. TESTS AND OBTAINED RESULTS

The extensive performance evaluation corroborates thereliability of the threat detection, which is otherwise impossible through canonical ML and shows that at least one of the proposed algorithms finds a decision boundary with a good trade-off between false positives and false negatives.

A. Canonical supervised learning and hyperparameter optimization

Canonical Supervised Learning refers to the standard and widely used approach in machine learning where a model is trained on labeled data to make predictions or classifications.In this framework, the model learns a mapping between the input features (independent variables) and the corresponding output labels (dependent variables). The goal is to find a function that accurately generalizes from the training data to make predictions on unseen data.

Hyperparameters are the configuration settings of a machine learning model that are not learned from the data but need to be set before the training process. Examples of hyperparameters include the learning rate, regularization strength, number of layers in a neural network, or the depth ofa decision tree. Hyperparameter optimization refers to the process of finding the optimal values for these hyperparameters to improve the model's performance.

As presented, we tested the ML algorithms on three different datasets: DNS tunneling, platooning and RUL estimation. The results obtained are reported by using metrics extracted from confusion matrices, in particular we decided to report false positive rate (FPR), true positive rate (TPR), false negative rate (FNR) and true negative rate (TNR). All results are shown in Table 1, divided by algorithm, attack and dataset.

B. Detection through XAI-driven reliable AI Detection through XAI-Driven Reliable AI refers to the useof explainable artificial intelligence (XAI) techniques toenhance the reliability and trustworthiness of AI-based

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125186

detection systems. XAI focuses on providing transparent andinterpretable explanations for AI models' decisions, enablingusers to understand and trust the output of these models. Byincorporating XAI into detection systems, we can improvetheir reliability, accountability, and ethical use.

XAI-driven reliable AI enables users to understand and trust the decisions made by AI models, leading to improved detection outcomes, reduced biases, and increased user adoption. For all test cases, the first step was the training of the default Logic Learning Machine (with 5% maximumerror allowed for each rule) on a 70% training set with a 30%test set (the same sets are used for all the detection algorithms.

		DNS				Platooning				RUL			
		FPR	TPR	TNR	FNR	FPR	TPR	TNR	FNR	FPR	TPR	TNR	FNR
Decision Tree	CW	0.50	1.00	0.50	0.00	0.43	0.55	0.57	0.45	0.00	0.00	1.00	1.00
	JSMA	0.25	1.00	0.75	0.00	0.23	0.84	0.77	0.16	0.00	1.00	1.00	0.00
	FGSM	0.50	1.00	0.50	0.00	0.13	0.85	0.87	0.15	0.0006	1.00	0.9994	0.00
Gradient boost	CW	0.48	1.00	0.52	0.00	0.44	0.59	0.56	0.41	0.00	0.00	1.00	1.00
	JSMA	0.50	1.00	0.50	0.00	0.33	0.88	0.67	0.12	0.00	1.00	1.00	0.00
	FGSM	0.03	0.36	0.97	0.64	0.12	0.92	0.88	0.08	0.00	1.00	1.00	0.00
KNN	CW	0.97	1.00	0.03	0.00	0.62	0.69	0.37	0.31	0.00	0.00	1.00	1.00
	JSMA	0.89	1.00	0.11	0.00	0.53	0.84	0.46	0.16	0.00	1.00	1.00	0.00
	FGSM	0.11	0.28	0.89	0.72	0.47	0.57	0.53	0.43	0.00	1.00	1.00	0.00
Logistic regression	CW	0.49	0.99	0.51	0.01	0.51	0.6	0.49	0.4	0.00	0.00	1.00	1.00
	JSMA	0.09	0.98	0.91	0.02	0.34	0.79	0.66	0.21	0.00	1.00	1.00	0.00
	FGSM	0.03	0.99	0.97	0.01	0.56	0.78	0.44	0.22	0.00	0.81	1.00	0.19
Random forest	CW	0.49	1.00	0.51	0.00	0.46	0.66	0.54	0.34	0.00	0.00	1.00	1.00
	JSMA	0.50	1.00	0.50	0.00	0.33	0.87	0.67	0.13	0.00	1.00	1.00	0.00
	FGSM	0.03	0.32	0.97	0.68	0.17	0.92	0.83	0.08	0.00	0.9992	1.00	0.0008
SVM	CW	0.39	0.65	0.61	0.35	0.63	0.85	0.37	0.15	0.00	0.00	1.00	1.00
	JSMA	0.09	0.98	0.91	0.02	0.24	0.69	0.76	0.31	0.00	1.00	1.00	0.00
	FGSM	0.15	0.95	0.85	0.05	0.69	0.89	0.31	0.11	0.00	0.00	1.00	1.00

Table1:performance statistics of canonical machine learning

		DNS		PLATOONING		RUL	
		FPR	TPR	FPR	TPR	FPR	TPR
Inside	CW	0.05±0.01	0.46±0.03	0.02±0.01	0.01±0.01	0.02±0.00	0.01±0.00
	JSMA	0.02±0.01	0.92±0.02	0.07±0.02	0.56±0.03	0.00±0.00	0.00±0.00
	FGSM	0.04±0.01	0.62±0.03	0.26±0.02	0.28±0.02	0.02±0.00	1.00±0.00
Outside	CW	0.00±0.00	0.00±0.01	0.00±0.00	0.00±0.00	0.00±0.00	0.06±0.01
	JSMA	0.00±0.00	0.24±0.03	0.00±0.00	0.26±0.03	0.03±0.00	0.81±0.02
	FGSM	0.00±0.00	0.24±0.03	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
LLM0%	CW	0.04±0.01	0.44±0.02	–	–	–	–
	JSMA	0.26±0.02	0.95±0.01	–	–	0.00±0.00	0.81±0.02
	FGSM	0.00±0.00	0.78±0.03	–	–	0.00±0.00	0.77±0.02

Table 2: Statistical Analysis :Reliable AI

		DNS		PLATOONING		RUL	
		FPR	TPR	FPR	TPR	FPR	TPR
zeroFPRSVDD	CW	0.00±0.00	0.34±0.05	0.09±0.02	0.15±0.03	0.10±0.05	0.10±0.04
	JSMA	0.10±0.01	0.80±0.02	0.09±0.02	0.55±0.08	0.00±0.00	0.99±0.01
	FGSM	0.14±0.02	0.28±0.01	0.10±0.01	0.15±0.02	0.00±0.00	0.99±0.02
eXplainableSVDD	CW	0.00±0.00	0.01±0.00	0.08±0.00	0.01±0.03	0.00±0.04	0.33±0.04
	JSMA	0.00±0.00	0.00±0.00	0.01±0.01	0.08±0.00	0.12±0.01	0.01±0.01
	FGSM	0.28±0.04	0.58±0.03	0.00±0.01	0.14±0.02	0.07±0.00	0.01±0.02

Table 3: Statistical Analysis on Reliable AI

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125186

		DNS				Platooning				RUL			
		FPR	TPR	TNR	FNR	FPR	TPR	TNR	FNR	FPR	TPR	TNR	FNR
Inside	CW	0.03	0.45	0.97	0.55	0.03	0.02	0.97	0.98	0.02	0.01	0.98	0.99
	JSMA	0.03	0.93	0.97	0.07	0.07	0.56	0.93	0.44	0.02	1.00	0.98	0.00
	FGSM	0.04	0.62	0.96	0.38	0.26	0.29	0.74	0.71	0.03	0.81	0.97	0.19
Outside	CW	0.00	0.01	1.00	0.99	0.01	0.00	0.99	1.00	0.00	0.00	1.00	1.00
	JSMA	0	0.72	1.00	0.28	0.01	0.26	0.99	0.74	0.00	0.06	1.00	0.94
	FGSM	0.00	0.25	1.00	0.75	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00
LLM0%	CW	0.04	0.44	0.96	0.56	-	-	-	-	-	-	-	-
	JSMA	0.47	0.50	0.53	0.50	-	-	-	-	0.00	0.81	1.00	0.19
	FGSM	0.39	0.42	0.61	0.58	-	-	-	-	0.00	0.77	1.00	0.23

Table 4: :XAI-Based Reliable Results

V. FUTURE SCOPE

The future scope of explainable and reliable approachesagainst adversarial machine learning holds significantpotential for advancing the security and trustworthiness of AI systems. The future of explainable and reliable approaches against adversarial machine learning lies in continuously pushing the boundaries of research, collaboration, and technological advancements. By combining explainability, reliability, and advanced defense techniques, we can build AI systems that are more resistant to adversarial attacks, provide transparent decision-making, and inspire trust in their outputs.

REFERENCES

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).doi:10.1145/2939672.2939778.

[2] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (SP). doi: 10.1109/SP.2017.49

[3] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[4] Lipton,Z.C.(2018).The mythos of model interpretability. In Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI). doi: 10.5555/3327757.3327795.

[5] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIA CCS '17). doi: 10.1145/3052973.3053009.

[6] Grosse, K., Papernot, N., Manoharan, P., Backes, M., & McDaniel, P. (2017). On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280.

[7] Xu, W., Evans, D., & Qi, Y. (2019). Feature squeezing: Detecting adversarial examples in deep neural networks. In Proceedings of the 28th USENIX Security Symposium (USENIX Security '19).

[8] Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey.IEEE Access, 6, 14410-14430. doi:10.1109/ACCESS.2018.2815072.

[9] Rony, J., Tasdizen, T., & Boult, T. E. (2019). Decoupling "when to update" from "how to update" in adversarial defense. In Proceedings of the IEEE International Conference on Computer Vision (ICCV). doi: 10.1109/ICCV.2019.00664.

[10] Wicker, M., Sultana, M., & Ray, S. (2020). Ensuring the reliability of adversarial examples for adversarial machine learning. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). doi: 10.1145/3375627.3375843