Review

A Review of XAI Failures and Works on Its Influence on Security

Ishan Banerjee *

Posted Date: 9 July 2024

doi: 10.20944/preprints202407.0754.v1

Keywords: explainable AI; machine learning; neural networks; cycbersecurity

 doi:10.20944/preprints202407.0754.v1

Article

A Review of XAI Failures and Works on Its Influence on Security

Ishan Banerjee

Chennai Mathematical Institute; ishanb3.141@gmail.com

Abstract: The importance of XAI is undeniable in the current scenario and thus it's importance in considerable. Unfortunately there are a couple of scenarios where the current XAI models fail. Additionally, attackers seek vulnerabilities in these models. This review paper is intended to cover some of the important scenarios where the existing XAI models may fail and some progresses made which shows their vulnerabilities and how to secure them.

Keywords: explainable AI; machine learning; neural networks; cycbersecurity

---

1. Introduction

The applications of AI in cybersecurity is significant which ranges from identifying threats and breaches using various models like clustering, classification and regression. However, they come with certain limitations which increases the usage of XAI. This may include manipulating malware files to escape the detections of the ML based detections.

Over the years a close association is established between XAI and cybersecurity. In numerous aspects, XAI has proven to be beneficial for compensating the limitations of AI. Two secondary risks posed are false positives and false negatives. It becomes important to ensure the accuracy of the prediction made. Further deployment of standard AI models is expensive. Cybersecurity significantly relies on AI to create secure systems. However, attackers can manipulate the malware such that the data manipulation remains undetected. Therefore we need a model which is more reliable, accurate and justifiable than the ones involving standard models.

Considering their applications in fields like health-care, it is important to provide relevant explanations of predictions, accurately without consuming more time, the validity of the results XAI provides is of considerable significance. However, owing to it's white box nature, they are vulnerable to attacks where the attackers can manipulate with the internal workings of the model which may result in the wrong and faulty explanations. Hence it is increasingly important to take measures to ensure the safety against any adversarial attacks.

This aim of the report is to give an overview of scenarios where XAI can fail and the current progresses made in the field of security involved with XAI in itself.

2. When Does XAI Fail ?

This section is referenced from [2] and [3] where we get situations where XAI can fail and the models relevant are DNNs and beyond. [2] gives an account of five critical failures of XAI and [3] classifies the failures into two parts, which includes Model-specific failures and User-specific failures. In this section we have collectively summarized the failures mentioned in both the works.

2.1. Robustness

[Alvarez-Melis and Jaakkola, 2018] mentioned that XAI methods might not be robust towards small changes. Let's consider a simple experiment. Considering a data-set involving prediction of hear-attack with the influence of certain features. Minor perturbations lead to different explanations which involves varied feature importance for varying feature contributions. In another experiment in [2], some minor noise was added to an image of the handwritten number '3'. A simple model involving CNNs was correctly able to classify the images but LIME's explanations was pretty varied

 doi:10.20944/preprints202407.0754.v1

considering the positive and the negative influence of pixels. [Adebayo et al., 2018] studied the robustness of saliency maps and figured out that many advanced saliency maps were giving similar explanations even when the data was randomized. In some cases, saliency maps which were supposed to provide local explanations, provided edge detection, which was independent of the model. Selecting appropriate methods or optimizing for robustness, prove to helpful in these scenarios.

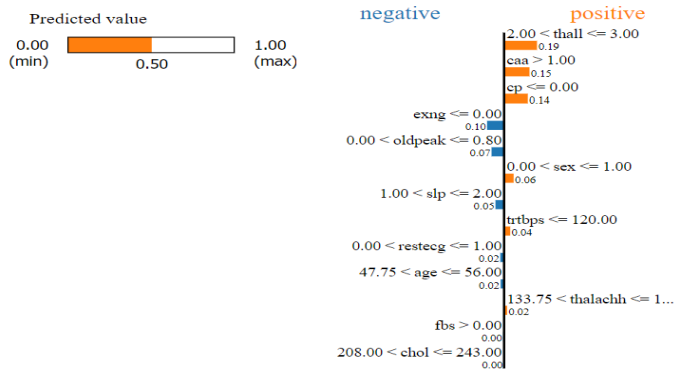

Figure 1. LIME explanation for a certain sample

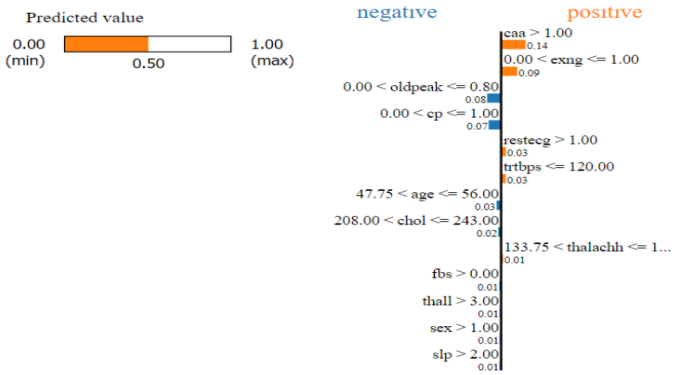

Figure 2. LIME explanation for the same sample as in Figure 1 on increasing each feature value by 0.1

2.2. Adversarial Attacks

Adversarial attacks lead to models giving nonsensical explanations, regardless of how good the model really is. Attackers can design imperceptible changes to the input which leads to the machine, misbehaving or misclassifying. An example would be to specifically distort a human picture such that the machine either misclassifies of provides wrong explanations.

2.3. Partial Explanations

Providing a small portion of the entire picture of what is happening, is another cause of problem. Providing partial or incomplete explanations of complex models leads to over-confidence. This can cause significant problems, for instance in the medical field. [2] gives an example where the AI correctly gives more attention to the tumors detected in the lung CT scan for the classification of benign and malignant but we don't get any idea about which portions of the tumors the AI took into consideration without a higher resolution picture.

 doi:10.20944/preprints202407.0754.v1

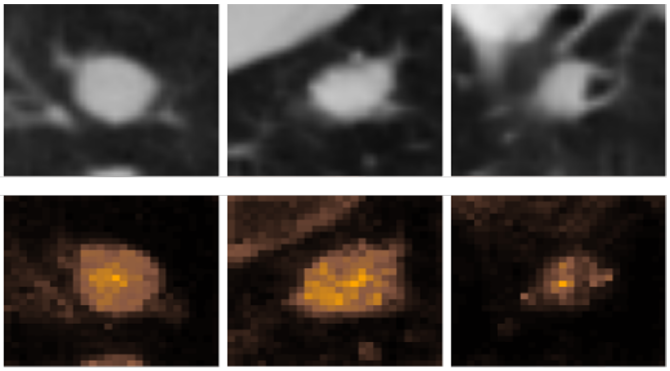

Figure 3. The first row shows the Lung CT scan and the second row shows that attention maps

2.4. Data and Concept Drift

Data and Concept drift involves the change of data distributions in the applicable domain. The model and its explanations remain fixed and hence isn't much reliable in certain cases. For instance consider an AI which is trained on data collected from African countries which end up in Asian hospitals. Additionally, new concepts might come up in the domain particularly. Hence it is required to consider them as well in the AI which was previously trained on old ideas.

2.5. Anthropomorphization

Another problem involves human's tendency to anthropomorphize XAI explanations, assuming a closer association between human and machine inferences. As mentioned in [2], observing a picture, humans perceive the features present in it, in a certain way but the machine may consider background pixels. This is problematic as seemingly plausible explanations are wrongly understood and can trick humans into believing the system.

2.6. Contradictory Explanations

This problem occurs when conflicting explanations are provided by either one or several systems. This can be classified into three cases. Contradiction between different pieces of information for the same explanation, contradiction between explanations provided by the same explainer, or contradiction between explanations by different explainers. This happens when complex effects like interactions of correlations between features are not taken into account.

2.7. Unstable Explanations

This problem has a close association with robustness. For example, consider the same data-set for heart-attack predictions (Figure 4). We can get different explanation for the exact same scenario. It occurs when for a relatively stable scenario, the explanations provided by the explainer are inconsistent. A probable cause of this is lack of robustness in the model. However other studies suggest that the lack of stability might be because of the complexity of the problem the model is dealing with.

2.8. Incompatible Explanations

This problem involves explanations provided by several systems which are not compatible with each other despite them being faithful. A high level cause of this lies in the method of providing explanations. For instance, consider LIME and SHAP. On the basis of different initial assumptions, the explanations they provide can vary. This can include varying top feature and mismatching order of importance. Interestingly, this is applicable to even global feature attribution methods. A different way of implementation of the system, or using different parameters for the same system, are potential reasons of this problem.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 July 2024 doi:10.20944/preprints202407.0754.v1

2.9. Mismatch

Mismatch occurs when the explanation provided by the system does not meet the expectations of the user. These problems occur as a certain explanation techniques do not consider the explanations the user seeks. This can cause mistrust and rejection of the model overall.

2.10. Counterintuitive Explanations

This problem occurs when the explanations provided doesn't match the perspective of an expert. Although this has some similarity with the mismatch problem, this case in particular refers to explanations which are contradictory to prior knowledge available. Considering prior knowledge in a domain which are acquired with time and experience, explanations provided by XAI may contradict those knowledge.

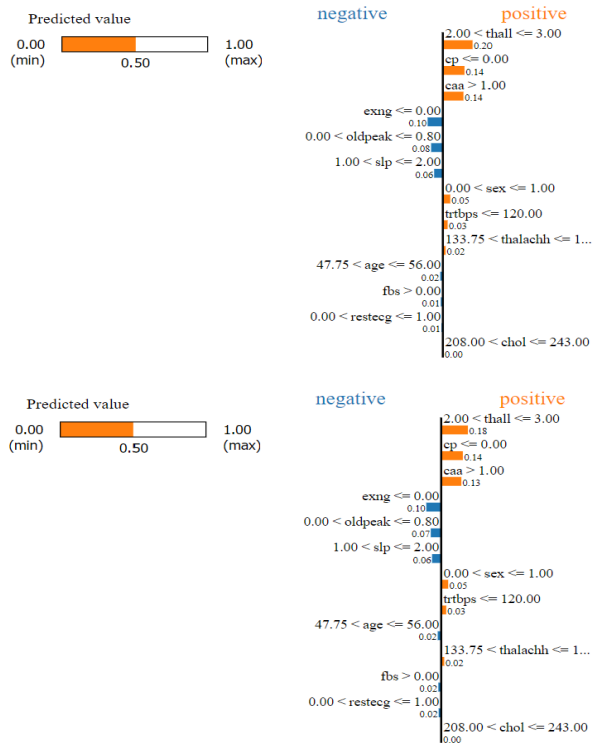

Figure 4. Different LIME explanations for a certain sample (Varying contributions of features)

3. Progresses

This section is dedicated towards some of the progresses made towards XAI security, which includes development of ideas towards a more secured XAI system or explicit vulnerabilities of already existing models.

3.1. Six Ws

Vigano and Magazzeni came up with the idea of explainable security (XSec). The method they mention is quite complicated as they involve a number of different stakeholders. They elaborated on six questions, and provided a review on how to secure the systems. The six questions are,

1. Who gives and receives the explanation?
2. What is explained?
3. When is an explanation given?
4. Where is the explanation given?
5. Why is XSec needed?
6. How to explain security?

 doi:10.20944/preprints202407.0754.v1

We will no elaborate on what the individual questions signify.

3.1.1. Who Gives and Receives the Explanation?

The characters who are considered in XSec includes, the designer of the system, the user, assuming them to be innocent non-experts, who can mistakes, making the system vulnerable, the attacker who tries to exploit the weak points of the system, the analyst of the system who analyses and tests the system and finally we have the defender who tries to defend the system. All of these roles involve someone to receive of provide explanations.

3.1.2. What is Explained?

it depends on the stakeholder, the type of explanation that is relevant. Developers need a detailed desiderata of the client so that they can realize the system in a secure and satisfactory way. For users, they need explanations which will establish their trust in the system and which can also help them in understanding on how the system can be used. For analysts, they will require system's specifications so that they can create models to analyze. Defenders need to have the knowledge of the vulnerabilities and related attacks and finally attackers will need visible vulnerabilities which they can exploit.

3.1.3. Where?

The authors consider four main cases that falls under this question.

1. Explanations are provided to the users as a part of the security policy.
2. Detaching explanations from the system and making it available elsewhere.
3. Considering a service where the users can interact with an expert system which provides explanations.
4. The best option which the authors consider is a 'security-explaining-carrying-system', although a considerable amount of work is required to ensure the it's safety.

3.1.4. When?

The authors mention when the explanations of security is required. They are required when the system is designed, implemented, deployed, used, analyzed, attacked, defended, modified and possibly even when decommissioned. These explanations are not only required during the runtime but also when the system is designed.

3.1.5. Why?

With the exception of attackers, all the other roles want the system to be secure. The explanations increases trust, confidence, transparency, usability, concrete usage, accountability, verifiability and testability.

3.1.6. How?

Depending on the intended audience, the explanations can be provided using natural language by an informal but structured spoken language, graphical language, involving explanation trees, attack trees, attack-defense trees, attack graphs, attack patterns, message-sequence charts, formal languages which includes proofs and plans or a gamification process where users learn about how to use the system.

3.2. Taxonomy of XAI and Black Box Attacks

Kuppa and Le-Khac in their paper presented a taxonomy of XAI in the security domain. Further they propose a novel black box attack which attacks the consistency, correctness and confidence of gradient based XAI methods.

 doi:10.20944/preprints202407.0754.v1

3.2.1. Taxonomy

The authors classify explainabilty space concerning the security domain into three main parts, X-PLAIN, XSP-PLAIN, XT-PLAIN. We will now discuss briefly about each of them.

1. X-PLAIN is regarding the explanations provided for the predictions given by the model. This includes, the static and interactive changes in explanations, local/global explanations, in-model/post-hoc explanations, surrogate models and visualizations of a model.
2. XSP-PLAIN includes confidential information such as features which are required to be protected, integrity properties of the data and the model ND privacy properties of the data and the model.
3. XT-PLAIN deals with the threat models considered. This includes correctness, consistency, transferability, confidence, fairness and privacy.

3.2.2. Proposed Black Box Attack

Consider a neural network which considers d features from the input and classifies them into k categories. Formally, $f; \mathbb{R}^d \rightarrow \mathbb{R}^k$. Consider the explanation map $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$, which assigns a value to each feature denoting their importance. Consider $g^t \in \mathbb{R}^d$ as a target explanation and $x \in \mathbb{R}^d$ as an input. I-attacks in particular, attacks the interpreter. In this case, we need a manipulated input, $a_{adv}x + \delta x$ such that it's explanation is very close to the target explanation, such that output of the classifier approximately, remains the same. In other words, $g(x_{adv})^t \approx g(x)^t$ and $f(x_{adv}) \approx f(x)$. In the case of CI-attack, both the model and the interpreter are attacked where $f(x_{adv}) \neq f(x)$ and $g(x_{adv})^t \approx g(x)^t$. Additionally, it is required that the perturbations made is small. A few more constraints must be taken into consideration. The model is black box, the perturbation made must be sparse and the perturbed input must be interpretable, which means that the instance must be close to the distribution of data considered for training the model.

Considering all these constraints, the attack is presented in two steps. The first one involves collecting n data points, using the model as a black box and train a surrogate model on the data and the output provided by the model. A problem in this case is the fact that the attacker is not aware of the distribution of data considered for training the model. To deal with that problem, we have Manifold Approximation Algorithm on the n data points which will give us the best piece wise spherical manifold or a subspace and a projection map p, mapping to the space such that mean square error $\sum_{i=1}^{n} ||x_i - p(x_i)||^2$ is minimized. This step will help us in dividing the data points into various data distributions. In a similar way, we can get explanation distributions. For the second step, we need to induce some minor distortions in the input distribution (z_a) to move the decision boundary to z_i and g_a to g_i, where a is the natural sample distribution and i represents the target distribution.

3.3. Interpretation of Neural Networks is Fragile

Authors of [4] introduce adversarial perturbation to neural network interpretation. They call the interpretation of neural network to be fragile if seemingly indistinguishable images with the same label, is given different interpretation.(See Figure 5) In this section we will talk about how the perturbations are made.

Consider a neural network, N a test data x_t and the interpretation $I(x_t, N)$. The goal is to make perturbations in the data such that they are imperceptible but change the interpretation. Formally this can be written as

$$\begin{aligned} &\arg\max_{\delta} D(I(x_t, N), I(x_t + \delta, N)) \\ &pred(x_t, N) = pred(x_t + \delta, N) \\ &\text{considering } ||\delta||_\infty < \epsilon \end{aligned}$$

The authors in [4] talk about three kinds of perturbations.

 doi:10.20944/preprints202407.0754.v1

The first one is random perturbations by $\pm\epsilon$ of each pixel. Considering this as a baseline, the authors compare other adversarial attacks which includes iterative attacks against feature importance and gradient sign attack against influence functions.

The former involves the attack against feature importance methods which involves taking series of steps in the direction, maximizing a differentiable dissimilarity function between the interpretation of original and perturbed input. We have three more classifications for the same. This involves perturbing the feature importance map by decreasing the importance of the top important features. Second attack, involving visual data, creates maximum spatial displacement of the center of mass of a picture and the third attack involves increasing the concentration of feature importance score of some predefined regions.

The latter attack does not rely on iterative approaches. The authors linearize the equation for influence function around the values of current inputs and parameters and constraining the L_∞ norm of the perturbation to ϵ, we get an optimal single step perturbation, which is,

$$\delta = \epsilon sign(\nabla_{x_t} I(z_i, z_t)) = -\epsilon sign(\nabla_{x_t} \nabla_\theta L(z_t, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_\theta L(z_i, \hat{\theta}))$$

The attack in this case will be to apply negative sign to the perturbation to decrease the influence of three training images which were the most influential for the original test image.

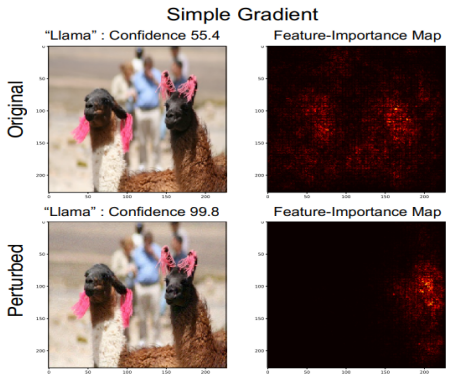

Figure 5. Seemingly indistinguishable pictures with the same label, but different interpretations on perturbations for the Simple Gradient model.

4. Fooling Neural Network Interpretations

Authors of [5] mention about two types of fooling which can be used to fool the state of the art saliency maps using data manipulation. They are Passive and Active fooling. This section is meant for discussing the two methods.

4.1. Preliminaries

Layer-wise relevance propagation (LRP) is a method that applies relevance propagation and generates a heatmap, showing the relevance of each pixel, including both the positive and the negatively impacting pixels.

Grad-CAM is an interpretation method that combines gradient activation and class activation maps to visualize the importance of each input. It's mainly used in CNN-based models.

SimpleGrad visualizes the gradients of prediction score with respect to the heat map for input. It shows the sensitivity of a prediction score with respect to small changes of input pixel.

Considering $D = \{(x_i, y_i)\}_{i=1}^n$ as supervised training set where $x_i \in \mathbb{R}^d$ represents the input data and $y_i \in \{1, \dots K\}$ represent the classification labels. Denoting w as a parameter for the neural network. The heatmap generated my an interpretation method I is denoted by

$$h_c^I(w) = I(x, c; w)$$

 doi:10.20944/preprints202407.0754.v1

where c is a certain class and $h_c^I(w) \in \mathbb{R}^{d_I}$. When $d_I = d$, jth value of the heatmap, $h_{c,j}^I(w)$ represents the score of the jth input x_j for the prediction class c.

4.2. Objective Function and Penalty Terms

The proposed the model requires the fine tuning of a pre-trained model with an objective function that combines ordinary classification loss and penalty terms for the interpretation results. The overall objective function which we need to minimize would be

$$\mathcal{L}(D, D_{fool}, I : w, w_0) = \mathcal{L}_C(D; w) + \lambda \mathcal{L}_F^I(D_{fool}; w, w_0)$$

Here $\mathcal{L}_C$ represents the ordinary cross entropy classification loss on the training data, w_0 is the parameter for the pre-trained model. $\mathcal{L}_F^I$ is the penalty term for D_{fool} and λ is a trade-off parameter.

4.3. Passive Fooling

Passive fooling involves making the interpretation method generate uninformative explanations. This includes location fooling, top-k fooling and center of mass fooling. These three attacks are already talked about in section 3.3.

4.4. Active Fooling

Making the interpretation method generate false explanations, falls under this fooling. The authors of [5] considered false explanations as swapping the explanations between two target classes. Unlike passive fooling, this method involves two datasets for computing the loss functions ($\mathcal{L}_C$ and $\mathcal{L}_F^I$ respectively).

5. More Works

Authors in [6] mention that the phenomenon of manipulating explanations arbitrarily by introducing visually hard to distinguish perturbations to the input that keeps the network's output constant, can be related to certain geometrical properties of neural networks. The large curvature in the network's decision function results in the in the unexpected vulnerability. Authors in [7] discussed a defense against adversarial attacks on heatmaps using multiple explanation methods which makes it robust against manipulations. The method is to take average of explanations. This is done by building and ensemble of explanation methods.

6. Conclusion

With the rise in the importance of Explainable AI, it becomes important to consider the vulnerabilities to secure the systems. This paper covered some of the advancements in the field in particular. We talked about the different scenarios where XAI can fail and some progresses which includes the importance of the six W's , the taxonomy of XAI and the black box attacks and the various ways interpretations of neural networks can be manipulated and also a potential way to defend such attacks.

References

1. Rawal, A., McCoy, J., Rawat, D.B., Sadler, B.M. and Amant, R.S., 2021. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. IEEE Transactions on Artificial Intelligence, 3(6), pp.852-866.
2. Chung, N.C., Chung, H., Lee, H., Chung, H., Brocki, L. and Dyer, G., 2024. False Sense of Security in Explainable Artificial Intelligence (XAI). arXiv preprint arXiv:2405.03820.
3. Bove, C., Laugel, T., Lesot, M.J., Tijus, C. and Detyniecki, M., 2024. Why do explanations fail? A typology and discussion on failures in XAI. arXiv preprint arXiv:2405.13474.
4. Ghorbani, A., Abid, A. and Zou, J., 2019, July. Interpretation of neural networks is fragile. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 3681-3688).

 doi:10.20944/preprints202407.0754.v1

5. Heo, J., Joo, S. and Moon, T., 2019. Fooling neural network interpretations via adversarial model manipulation. Advances in neural information processing systems, 32.
6. Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R. and Kessel, P., 2019. Explanations can be manipulated and geometry is to blame. Advances in neural information processing systems, 32.
7. Rieger, L. and Hansen, L.K., 2020. A simple defense against adversarial attacks on heatmap explanations. arXiv preprint arXiv:2007.06381.