International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI),
Vol.14, No.1, February 2025

# ADVXAI IN MALWARE ANALYSIS FRAMEWORK: BALANCING EXPLAINABILITY WITH SECURITY

Darring White

School of Business and Technology, Marymount University, Virginia, USA

***ABSTRACT***

*With the increased use of Artificial Intelligence (AI) in malware analysis there is also an increased need to understand the decisions models make when identifying malicious artifacts. Explainable AI (XAI) becomes the answer to interpreting the decision-making process that AI malware analysis models use to determine malicious benign samples to gain trust that in a production environment, the system is able to catch malware. With any cyber innovation brings a new set of challenges and literature soon came out about XAI as a new attack vector. Adversarial XAI (AdvXAI) is a relatively new concept but with AI applications in many sectors, it is crucial to quickly respond to the attack surface that it creates. This paper seeks to conceptualize a theoretical framework focused on addressing AdvXAI in malware analysis in an effort to balance explainability with security. Following this framework, designing a machine with an AI malware detection and analysis model will ensure that it can effectively analyze malware, explain how it came to its decision, and be built securely to avoid adversarial attacks and manipulations. The framework focuses on choosing malware datasets to train the model, choosing the AI model, choosing an XAI technique, implementing AdvXAI defensive measures, and continually evaluating the model. This framework will significantly contribute to automated malware detection and XAI efforts allowing for secure systems that are resilient to adversarial attacks.*

***KEYWORDS***

*Artificial Intelligence (AI), Explainable AI (XAI), Adversarial XAI (AdvXAI), Machine Learning (ML), Malware Analysis, Cybersecurity, Security*

## 1. INTRODUCTION

Artificial Intelligence (AI) is accelerating rapidly in all facets of life where it can create efficiencies and add value. AI is being utilized in many areas such as education, healthcare, and cybersecurity. With the increased utilization of AI also comes various concerns. AI makes many decisions that users are not entirely aware of how the technology came to a specific conclusion. This is fine for many uses of AI, but in a high-risk business, stakeholders are very keen to know what the AI is thinking. This led DARPA, the US Defense Advanced Research Projects Agency, to come up with the Explainable AI (XAI) program. The goal of XAI is to provide models that humans can interpret in order to gain trust in the decisions AI models make [1]. There are many benefits gained from XAI including the legal right to explainability policies, trust in decision-making processes, fine-tuning gaps, avoiding bias, and more.

Researchers are using Machine Learning (ML) and Deep Learning (DL) models to analyze malware more effectively, but it is important to know how the models classify malware and how they came to their decision of whether it is malicious or benign. XAI assists in malware detection by providing a tool for auditing and validation. XAI is essential in malware analysis to clearly understand the reasons behind classification. More specifically, malware analysts can use XAI to improve ML and DL model interpretability by showing which aspects of a file led to its decision.

It allows an analyst to use their professional expertise to either agree or disagree with the output. If a model can discover previously unknown malware, it can provide an analyst with what led to the discovery. By providing insights, analysts can update and train the model as necessary to be more accurate. XAI also helps in other areas such as investigations, forensics, and threat intelligence as well. While these benefits are useful for analysts, they are equally beneficial for adversaries.

Adversarial XAI (AdvXAI) is a newer concept that explores the intersection of adversarial AI and XAI. Current studies showing examples of AdvXAI are in perturbations of images. Researchers can generate disruptions to cause the models to have different outputs that are invisible to the human eye. The model will misclassify or see the image as something else. Other such attacks allow adversaries to evaluate the XAI response to build evasions to AI models detecting malware. There are minimal studies that investigate all possible attack scenarios whether proven, logical, or theoretical, that fall under the AdvXAI concept.

With new adversarial attacks stemming from XAI in malware analysis, it is necessary to develop an AdvXAI malware analysis framework to balance explainability with security to ensure XAI does not create vulnerabilities.

## 2. LITERATURE REVIEW AND MOTIVATION

As mentioned, DARPA created the XAI program in 2016. In a broad agency announcement, the agency announced a funding opportunity for XAI soliciting research proposals [2]. In an effort to trust and manage AI systems, they desired new or innovative techniques and models to fully explain the decisions that systems make. They cited that AI systems were advancing and becoming more efficient, but partly due to new techniques in black-box environments, users were unsure of the underlying decisions. The goal was to support end-users in understanding and trusting AI systems more while also maintaining high levels of performance. Less than two years later, DARPA demonstrated initial implementations of explainability of AI systems.

DARPA may have revived the topic, but XAI traces back as early as AI itself with people needing to understand the decisions behind the technology. However, the concept, research, and application all have developed over time. Confalonieri et al. wrote about the history of XAI in 2020, highlighting trust and safety in AI decision-making across numerous applications [3]. Goebel et al. also wrote in 2018 about early XAI in the 1980s and its evolution over time, highlighting its use in medical applications and more [4]. Between the years of 2018 and 2020, research saw a spike in XAI topics. This researcher started to investigate XAI for malware analysis and found that most research papers began publishing starting in 2022. There are many discussions and challenges concerning XAI in malware analysis; a particular challenge is how XAI is vulnerable to adversarial attacks.

The researcher found that Kuppa and Le-Khac were among the earliest to mention AdvXAI; in 2021 their paper ‘Adversarial XAI Methods in Cybersecurity’ proposed a new attack leveraging XAI compromising confidentiality and privacy in classifiers [5]. In 2024, however, there has been a keen interest in AdvXAI. Key notable studies include a survey of adversarial attacks in XAI [6], explainable malware analysis challenges [7], XAI for malware hunting [8], and an anatomy of XAI adversarial attacks [9]. These articles succinctly address the list of possible AdvXAI attacks that threat actors can take advantage of, thus the need for a balance between explainability and security.

There are a number of recent existing defenses and proposed frameworks for AdvXAI. In 2022, MeTFA (Median Test for Feature Attribution) was introduced to address explainability as well as adversarial attacks against explainability in image classifications [10]. The framework was shown to have better explanations and the ability to defend against adversarial attacks. ATEX (Adversarial Training on Explanations) is an adversarial training methodology for the stability and robustness of explanations. ATEX has an application in defending against explainable manipulation. The original paper does not mention using ATEX for malware analysis, but there is a practical application for incorporating its methodology into this paper's framework. The Adversarial Observation framework was introduced in 2024 with the purpose "to facilitate the implementation of adversarial attacks and Explainable AI (XAI) techniques during machine learning model training" [11]. Again, this framework is targeted towards image classification and while there is a practical application in malware analysis, the framework is meant to train the model against attacks rather than attacks that occur from knowledge of the explanations. This framework will build upon these existing studies, techniques, and frameworks to outline a framework that incorporates the lifecycle from model selection, training, explanation, security, deployment, and continued evaluation. Unlike previous frameworks, this paper will focus specifically on malware analysis and preventing adversaries from using explainable methods to craft more advanced malware.

## 3. FRAMEWORK OBJECTIVE

The three main objectives for the framework are to have explainability balanced with secure robustness by design features to not leak vulnerable insights to adversaries while maintaining high levels of performance.

### 3.1. Explainability

For an XAI model to be effective, it must provide clear human-readable and understandable outputs.

### 3.2. Security

The XAI model must be secure from deliberate exploitation by adversaries that either assist them in deceiving, evading detection through leaked information, unauthorized access by the ml/dl model, or ways that mislead the model.

### 3.3. Robustness

Robustness refers to the performance of the model even when exposed to adversarial perturbations that attempt to degrade the performance of the system or trick it into making wrong decisions about malware samples between explainability and security.

## 4. FRAMEWORK COMPONENTS

Examples are given for the main components. Choosing specific component models or tools depends on various factors and goals for the organization in malware detection. Not all examples are listed, however, the ones chosen depict common uses. Multiple models can be used in different use cases the entity may have or in stages as deemed necessary. Multiple tools should be combined where feasible.

## 4.1. ML/DL Model Selection

The models listed are examples of effective models for malware analysis; there are various other ML/DL models available for different use cases.

### 4.1.1. Traditional ML Models

- Random Forests (RF). Highly accurate traditional model with the ability to ingest large datasets [12].
- Support Vector Machines (SVM). Ideal for smaller datasets [13].

### 4.1.2. DL Models

- Convolutional Neural Network (CNN). Black-box model that converts samples to images [8].
- Long Short-Term Memory (LSTM). Effective at analyzing sequential data [14].

## 4.2. XAI Technique

The two XAI tools described in this paper below are popular model-agnostic techniques, meaning that they can be implemented into any ML model. There are other model-specific methods such as DeepLIFT, Grad-CAM, and Integrated Gradients [15].

**Local Interpretable Model-Agnostic Explanations (LIME).** Focused on model prediction explanation for individual instances.

**SHapley Additive exPlanations (SHAP).** Utilizes game theory to provide a global understanding of the dataset within a model. Explains predictions by computing the attributes of each feature.

**Counter-Factual Statements.** While not an XAI technique on its own, counter-factual statements are important to implement to provide explainability by asking ‘what-if’ questions. These questions can answer why a malware sample was marked as benign or what would need to change for a sample to be considered malware.

## 4.3. Adversarial Security Defense Measures

These are highlighted known defense measures against AdvXAI. This is not an all-inclusive list as there are other tools such as ATEX that can be implemented in the framework. Multiple defense measures may be chosen dependent on factors such as goals, budget, skill, and integration.

**Secure by Design.** Ensure the framework is designed for security primarily. This includes the entire framework and the tools within so that attackers are less likely to find vulnerabilities that allow access to the ML/DL models and XAI technology.

**Robustness.** Robustness refers to AI models that may be less interpretable but are more robust to manipulation by adversaries. This concept is key to balancing out the amount of interpretability

that is necessary while not compromising security. XAI that provides too much explainability leaves information for attackers to better evade the detection models [5].

**Adversarial Training.** Create and inject adversarial samples into the learning datasets, those that would be used by an adversary, intended to trick the model so that it is able to learn and become resilient to manipulations [16].

**Gradient Masking.** Adding masking features that blur out important or sensitive areas that attackers will mostly use as an attack vector [17].

**Noise-Based Defense.** Adding noise or perturbations to models that hinder an adversary's ability to reverse-engineer or for adversarial perturbations to affect on the model [5].

**Feature Squeezing.** Utilizing compression techniques to convert values into smaller representations. Effectively squeezing out unnecessarily large input spaces [18].

**Security Audits.** Regular security assessments such as penetration tests that incorporate adversarial training techniques as well as broader threat actor tactics, techniques, and procedures (TTPs) that attempt to access model features.

## 5. FRAMEWORK WORKFLOW

### 5.1. Initial Data Collection

Identify the collection of malware files and benign files that will be used to train the ML/DL model. The files should have features from different analysis methods. VirusSign possesses a large dataset of malware samples for free and subscription use. [19]

### 5.2. Model Training

Choose the desired model based on identified needs, capabilities, and feasibility. Begin training the model on a collection of malware and benign samples.

### 5.3. XAI Integration

Begin applying the XAI technique/tool that is identified as the most appropriate integration for the model and end-user that will analyze the explanations.

### 5.4. AdvXAI Defenses

Identify the defense measures to implement. The ones from the previous section are preferred. However, if constraints exist, choose appropriate measures for the desired use case. Multiple defense measures may be put in place.

### 5.5. Evaluation

Monitor the performance of the model by defined goals and metrics to include:

- Performance
- Explainability

- Trade-Off Analysis

### 5.6. Deployment

When the model meets the desired metrics, deploy for use in malware detection and analysis workflow.

### 5.7. Continued Evaluation and AdvXAI Defense Evaluation

Following deployment, continually monitor the system's ability to meet desired goals in malware detection and explainability for malware analysts. Continually test the model against implemented defense features. Assess for AdvXAI to ensure explainability does not create a vector for manipulation or evasion within the model. Perform security assessments, such as penetration tests, to ensure adversaries do not have direct access to the model.

### 5.8. Feedback

Allow for analyst comments on the model's predictions that are able to be looped back into allow the system to improve over time.

## 6. TRADE-OFF ANALYSIS

The framework is designed for balancing explainability with security. Therefore, it is necessary to identify the level of explainability desired without revealing significant sensitive information. For this framework, a least amount of explainability necessary approach is preferred.

Those who manage AI systems for malware analysis and malware analysts who use these systems should consider how much explainability they need to understand and trust the decisions the model makes. They must analyze the explanations and use expert knowledge to determine if sensitive information is in the output that an attacker can use. This is in addition to the defense mechanisms of the framework.

To ensure that security is not compromised, there needs a to be policy for a trade-off analysis. Trade-off analysis in XAI is usually a term for evaluating explainability and performance or privacy, in this paper, it is applied to a trade-off of security. A trade-off analysis team will review explanations and determine if all descriptions are necessary to achieve goals. When explanations surpass necessary value, the review team will determine if it creates a risk to AdvXAI attacks and put mitigations in place.

## 7. LIMITATIONS

This framework is theoretical and needs thorough testing to ensure its effectiveness in practical usage. At this stage, the individual components are proven effective measures through previous studies and applications. However, using the framework in seamless integration with an entity performing malware analysis and testing entity-specific metrics will assist in ensuring its effectiveness. Additionally, the framework is broad allowing for scaling and tool selection. This provides flexibility but does not specify specific requirements in the framework. Entities that use this framework will need to create a policy that follows the objective, components, and workflow.

Another limitation is that there are some ML/DL models, XAI techniques, and AdvXAI defensive measures listed in this paper; researchers need to investigate further into the vast set of tools available that appropriately meet their objectives. Expected challenges may occur through change management processes as organizations decide on which models, techniques, and measures they will implement and that those components integrate effectively with each other.

## 8. OPPORTUNITIES FOR FUTURE WORK

There are various opportunities in AdvXAI and XAI. Researchers may evaluate and update this framework for more efficiency. Researchers may also create metrics for balancing security and explainability. Furthermore, there is still a need for a graphical user interface (GUI) dashboard to assist in clear understanding, visual representation, and reporting of XAI.

## 9. CONCLUSION

This paper created a working framework that serves several purposes focusing on AdvXAI defensive response measures to increase resiliency against adversarial attacks when AI models are used for malware analysis.

The paper analyzed the evolution of existing models used for malware analysis and highlighted evolving techniques and challenges. It then created a framework allowing for robust, transparent, and secure models. Each step allows an analyst or organization the ability to choose which components work best for them while ensuring they meet the steps for security against AdvXAI attacks.
By implementing this framework, analysts will be empowered to trust AI predictions and decisions while ensuring reduced risks against adversarial attacks.

## ACKNOWLEDGMENTS

The author would like to thank all of my professors at Marymount University. Especially Dr. Alex Mbaziira for his support in seeking publication and Dr. Donna Schaeffer for her expertise and advice. I would also like to thank my family who allow me to study and pursue my passions.

### REFERENCES

[1] S. Ali *et al.*, “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence,” *Inf. Fusion*, vol. 99, p. 101805, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.

[2] DARPA, “Explainable Artificial Intelligence (XAI).” DARPA, Aug. 10, 2016. [Online]. Available: https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

[3] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, “A historical perspective of explainable Artificial Intelligence,” *WIREs Data Min. Knowl. Discov.*, vol. 11, no. 1, p. e1391, Jan. 2021, doi: 10.1002/widm.1391.

[4] R. Goebel *et al.*, “Explainable AI: The New 42?,” in *Machine Learning and Knowledge Extraction*, vol. 11015, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds., in Lecture Notes in Computer Science, vol. 11015. , Cham: Springer International Publishing, 2018, pp. 295–303. doi: 10.1007/978-3-319-99740-7_21.

[5] A. Kuppa and N.-A. Le-Khac, “Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, United Kingdom: IEEE, Jul. 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9206780.

[6] H. Baniecki and P. Biecek, "Adversarial attacks and defenses in explainable artificial intelligence: A survey," 2023, doi: 10.48550/ARXIV.2306.06123.

[7] H. Manthena, S. Shajarian, J. Kimmell, M. Abdelsalam, S. Khorsandroo, and M. Gupta, "Explainable Malware Analysis: Concepts, Approaches and Challenges," Sep. 09, 2024, *arXiv*: arXiv:2409.13723. doi: 10.48550/arXiv.2409.13723.

[8] M. Saqib, S. Mahdavifar, B. C. M. Fung, and P. Charland, "A Comprehensive Analysis of Explainable AI for Malware Hunting," *ACM Comput. Surv.*, vol. 56, no. 12, pp. 1–40, Dec. 2024, doi: 10.1145/3677374.

[9] G. Mikriukov, G. Schwalbe, F. Motzkus, and K. Bade, "The Anatomy of Adversarial Attacks: Concept-based XAI Dissection," 2024, *arXiv*. doi: 10.48550/ARXIV.2403.16782.

[10] Y. Gan *et al.*, "'Is your explanation stable?': A Robustness Evaluation Framework for Feature Attribution," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Los Angeles CA USA: ACM, Nov. 2022, pp. 1157–1171. doi: 10.1145/3548606.3559392.

[11] J. Gafur, S. Goddard, and W. Lai, "Adversarial Robustness and Explainability of Machine Learning Models," in *Practice and Experience in Advanced Research Computing 2024: Human Powered Computing*, Providence RI USA: ACM, Jul. 2024, pp. 1–7. doi: 10.1145/3626203.3670522.

[12] B. M. Khammas, "Ransomware Detection using Random Forest Technique," *ICT Express*, vol. 6, no. 4, pp. 325–331, Dec. 2020, doi: 10.1016/j.icte.2020.11.001.

[13] A. Vaidya, M. Pande, S. Shankrod, T. Dorkar, and S. Aundhakar, "DETECTION OF MALWARE USING SVM," *Int. Res. J. Mod. Eng. Technol. Sci.*, vol. 5, no. 3, pp. 2927–2931, Mar. 2023, doi: 10.56726/IRJMETS34910.

[14] R. Ch, J. Manoranjini, S. Pallavi, U. Naresh, S. Telang, and S. Kiran, "Advancing Malware Detection Using Memory Analysis and Explainable AI Approach," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Coimbatore, India: IEEE, Aug. 2024, pp. 518–523. doi: 10.1109/ICoICI62503.2024.10696406.

[15] Z. Keita, "Explainable AI - Understanding and Trusting Machine Learning Models," DataCamp. [Online]. Available: https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models

[16] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent Advances in Adversarial Training for Adversarial Robustness," Apr. 21, 2021, *arXiv*: arXiv:2102.01356. doi: 10.48550/arXiv.2102.01356.

[17] I. Goodfellow, "Gradient Masking Causes CLEVER to Overestimate Adversarial Perturbation Size," Apr. 21, 2018, *arXiv*: arXiv:1804.07870. doi: 10.48550/arXiv.1804.07870.

[18] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," in *Proceedings 2018 Network and Distributed System Security Symposium*, San Diego, CA: Internet Society, 2018. doi: 10.14722/ndss.2018.23198.

[19] "VirusSign - Giant Malware Sample Repository." https://github.com/VirusSign/malware-samples. [Online]. Available: https://github.com/VirusSign/malware-samples