A Benchmark for Explainability of Machine Learning for Cyberattacks Detection

1st Amani Abou Rida
ICube - Laboratoire
des sciences de l'ingénieur,
de l'informatique et de l'imagerie
UMR 7357, Université de Strasbourg
CNRS, 67000, Strasbourg, France
Laboratoire de Recherche
de L'EPITA (LRE), 14-16 rue Voltaire
94270 Le Kremlin-Bicêtre, France.
abou-rida.amani@etu.unistra.fr

2nd Rabih Amhaz
ICube - Laboratoire
des sciences de l'ingénieur,
de l'informatique et de l'imagerie
UMR 7357, Université de Strasbourg
CNRS, 67000, Strasbourg, France
Icam, site de Strasbourg-Europe
67300 Schiltigheim, France
amhaz@unistra.fr

3rd Pierre Parrend
ICube - Laboratoire
des sciences de l'ingénieur,
de l'informatique et de l'imagerie
UMR 7357, Université de Strasbourg
CNRS, 67000, Strasbourg, France
Laboratoire de Recherche
de L'EPITA (LRE), 14-16 rue Voltaire
94270 Le Kremlin-Bicêtre, France.
parrend@unistra.fr

***Abstract*—The importance of explainability in machine learning models for cyber security is growing, as it provides a better understanding of the decision-making process and improves the ability to defend against attacks. However, the application of explainability in the context of cyber security is becoming a challenge, as there is currently a lack of standard methodologies for evaluating and comparing the performance and explanations of different models. This paper presents a proposal for a benchmark in the field of explainability of cyberattacks, aimed at enhancing organizations' analysis and response capabilities.**

***Index Terms*—explainability, classification of cyberattacks, cybersecurity**

I. Introduction

Classification models are used to identify specific cyberattacks by identifying patterns and anomalies in large amounts of data with a high accuracy rate. However, it is impossible for the security analyst to trust a Machine Learning (ML) model, and thus to take decisions based on its insights, without understanding how and why it makes its decisions. This shows the importance of using explainable AI (xAI) that can help improve the transparency, interpretability and trustworthiness [Hle19] of AI-based cybersecurity systems. The objective is to extract the dataset features that lead to the classification and the data points triggering an alert, to provide suitable information about how to respond to a specific attack. In this paper, we present a benchmark for explaining cyberattacks in ML to evaluate and compare the performance and explainability of different Explainable AI (xAI) models. The purpose is to provide specifications, a clear and standardized methodology for evaluating and comparing xAI models on the problem of classification of cyberattacks. To accomplish this objective, the research addresses the following 2 questions:

- How to apply xAI on ML process for the classification of cyberattacks?
- How to evaluate and compare xAI models for the classification of cyberattacks?

The paper is organised as follows. Section II introduces the state of the art, Section III identifies the benchmark for explainability of cyberattacks. Section IV presents its implementation and used dataset. Section V provides the evaluations and discussion. Section VI concludes this work.

II. State of the Art

A. Explainable AI (xAI) in ML

In the context of ML models, the definition of the term Explainable Artificial Intelligence (xAI) is given by [Arr+20]:" Given an audience, an explainable Artificial Intelligence model is one that produces details or reasons to make its functioning clear or easy to understand."

1) Explainability Properties in ML: The main properties for achieving explainability refers to: **Traceability** the ability to trace and understand the decision-making processes [Mor+21], **Understandability** the characteristic of a model to make a human understand its function [Mur+19], **Comprehensibility** the ability of an algorithm to represent its learned knowledge in a human understandable way [Fer+19], **Explainability of user interface** a clear and understandable explanations for decisions to its users [Arr+20], and finally **Interpretability** the ability of providing a meaning for a model in terms understandable and shareable by human [Arr+20].

2) Explainability of ML Models: The most commonly used explainability approaches focusing on the ML models themselves are impurity-based feature importance (IP), permutation feature importance (PFI), LIME, and SHAP. **IP** [Alt+10] is a metric used in tree-based models to evaluate the contribution of each feature to the prediction performance. It is calculated by measuring the average impact that each feature has on the quality of the model's predictions, based on the reduction in impurity or improvement in accuracy. **PFI** [Alt+10] measures the decrease in model performance after shuffling feature values, but is computationally expensive and sensitive to noise in the data. **LIME** [RSG16] provides explanations that are easy for non-experts to understand and can be used

[0] RESSI Thesis contribution - 1/10/2020-30/9/2023

with any model, but only provides local explanations into how a particular input or feature influenced the prediction outcome. **SHAP** [LL17] breaks down the contribution of each feature and is model-agnostic and can handle continuous and categorical features, but is computationally expensive and may be affected by correlated features.

III. Benchmark

According to the explainability properties defined by [Arr+20], we have extended their framework to apply and evaluate these properties in machine learning by providing specifications for each step as shown in section III-A and metrics as shown in section III-B.

A. How to apply Explainability on ML process for detecting cyberattacks?

In this section we provide some specifications to follow when applying explainability properties. For **data's traceability**, compare the labeling of cyberattacks according to their definition to the labeling in the data input, calculate the correlation between values and class membership, check the Indicators of Compromise (IoC) in the dataset for each feature, and analyse the important features in the dataset such as ports and packets size. For **model's understandability**, understand the architecture of the ML model to identify any potential issues with the model's design and ensure that it is suitable for the task of identifying cyberattacks. In addition, understand which features the model is using to make decisions, and how important they are in terms of identifying cyberattacks. For **output's comprehensibility**, observe the output of the classifiers by evaluating its performance on a test set of data. For **user's interface explainability**, use interactive visualizations such as feature importance plots that can help users understand the decision-making process of the model and how it is identifying cyberattacks. Also use interactive dashboards to display the model's performance. Finally, for **human's interpretability**, provide a human-centered evaluation where human users are asked to interact with the model and provide feedback on its interpretability and usefulness. For example: Can I explain the significance of data analysis so that my colleague understand it? Is the measurement of feature scores understandable for detecting cyberattacks?

B. How to evaluate Explainability properties for cyberattacks?

In this section we propose three phases to evaluate explainability for cyberattacks: **Explainability of Performance** Refers to the ability to understand the classification of cyberattacks. The main objective is to help organizations better understand the strengths and limitations of their AI systems to increase trust and improve overall performance. This includes providing information on the model's accuracy, precision, recall, F1 score, TNR, and loss. These values are used to quantify the performance of the prediction where users can understand how well the model is performing and identify any potential issues with its predictions, and thus help us achieve **understandability** of the model and **comprehensibility** of the output by providing a detailed understanding of the model's behavior and the reasoning behind its predictions. **Explainability of feature relevance** Refers to the impact of features on the classification model and to the visualization that provides a clear representation of how different features of the model's input influence its predictions and aids **user interface explainability**. This can be achieved through explainability models that are mentioned in II-A2. Moreover, feature relevance explanations can help make the model's decision-making process more transparent and identify the key inputs that drive a model's decisions, which can further aid in **traceability** and **model understandability**. Additionally, interactive visualizations can be used to allow users to explore the model's behavior in different regions of the feature space, which can further enhance the **interpretability** of the model. **Tradeoff explainability-accuracy-time performance** Refers to the relationship between the accuracy of a model and the amount of time it takes to make predictions. This trade-off can solve the issues of under-fitting that result in a lower accuracy and a shorter training time and over-fitting of the data that result longer training time and a lower accuracy on the test set. On the other hand, the explainability-accuracy trade-off refers to the balance between a model's ability to accurately predict cyberattacks and its ability to provide clear and understandable explanations for its predictions.

IV. Dataset and Implementation

The dataset used is TON-IoT [Boo+21] which contains 461043 rows, 45 features, and nine attack categories. For the implementation, we use a Jupyter notebook version 1.0.0 that runs on a Larry server with model Gen10+, cpu 48-Core processor, and ram 3 TB.

V. Evaluation and Discussion

The evaluation of the benchmark for explainability of cyberattacks consists of three phases.

A. Explainability of Performance

We apply the explainability of performance metrics that we mentioned in Section III-B to compare the performance of 4 classifiers: CART, Random Forrest, XGBoost, and MLP. Table.I shows the comparison of the performance metrics. Although XGBoost shows better accuracy, CART was faster and had similar accuracy as XGBoost. This shows the importance of having a balance between accuracy and time in order to effectively protect against cyberattacks.

B. Explainability of feature relevance

We apply 4 different explainability models that are mentioned in II-A2 to evaluate the explainability of feature relevance. Fig.1 shows the summary plots for each attack type when applying SHAP tree explainer model that is applied only on tree models such as XGBoost, Random Forest, and Decision Tree classifiers. These plots show the importance of features and ranked them in descending order based on the

TABLE I
Compare Explainability Performance

Performance Metrics	**learning algorithm**			
	CART	***Random Forest***	***XGBoost***	***MLP***
Precision	0.992388	0.5374	**0.994333**	0.792293
Recall	0.992358	0.712189	**0.99423**	0.80414
TNR	0.848892	0.07061	**0.99423**	0.102611
Accuracy	0.848892	0.712189	**0.99423**	0.80414
Balanced Accuracy	0.992372	0.242396	**0.99423**	0.518739
F1 score	0.992372	0.603236	**0.99423**	0.787025
MCC	0.992372	0.385834	**0.99423**	0.634509
zero one loss	0.008	0.288	**0.006**	0.267
Fit time sec	**1.231235**	11.873	445.818	341.461
Pred time sec	**0.028048**	1.056989	0.146705	0.151560

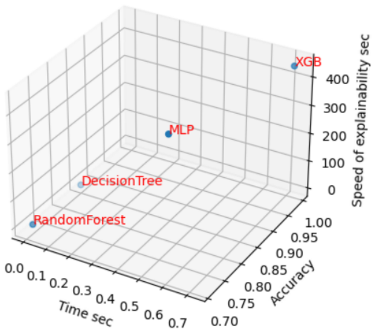

Fig. 2. Tradd-off SHAP explainability-accuracy-time performance on different classifiers

effect the prediction has on each class. For example, taking XGBoost summary plot, we see that the feature called "dst port" for destination port, which is located in high level of this plot, resulted in classes "Password" more than the remaining class types. This shows how the feature relevance scores and visualization provide valuable insights into the model.

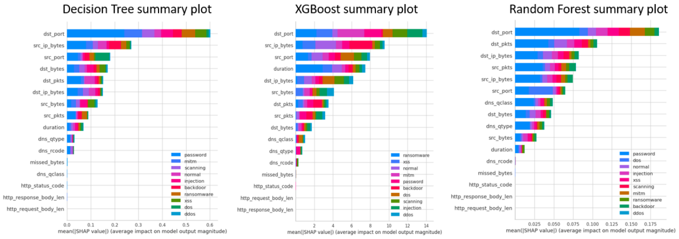

Fig. 1. SHAP summary plots for different classifiers

C. *Tradeoff explainability-accuracy-time performance*

We apply a 3D plot to show the tradeoff explainability-accuracy-time performance. Fig.2 shows the tradeoff using SHAP model as speed of explainability, accuracy, and Fit time. This is important due to the time-sensitive nature of cyberattacks, to quickly and accurately classify and explain an attack so that incident responders can take appropriate actions in a timely manner. For example, the XGBoost classifier has the best accuracy, but the speed of attack explaination is significantly slower than competing approaches.

VI. Conclusions and Perspectives

Our proposed benchmark for explainability of Machine Learning models for detecting cyberattacks helps organizations improve their own analysis and response. This benchmark highlights how to apply explainability properties on ML processes and how to evaluate and compare the explainability for detecting cyberattacks. However, ML requires high computational power to detect and explain evolving attacks. This opens a great challenge for explaining cyberattacks using graph learning models that allows for the visualization and analysis of complex relationships and for detecting advanced cyberattacks.

References

[Alt+10] André Altmann et al. "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10 (2010), pp. 1340–1347.

[Arr+20] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information fusion* 58 (2020), pp. 82–115.

[Boo+21] Tim M Booij et al. "ToN_IoT: The role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion data sets". In: *IEEE Internet of Things Journal* 9.1 (2021), pp. 485–496.

[Fer+19] Alberto Fernandez et al. "Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?" In: *IEEE Computational intelligence magazine* 14.1 (2019), pp. 69–81.

[Hle19] AI Hleg. "Ethics guidelines for trustworthy AI". In: *B-1049 Brussels* (2019).

[LL17] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[Mor+21] Marçal Mora-Cantallops et al. "Traceability for trustworthy ai: A review of models and tools". In: *Big Data and Cognitive Computing* 5.2 (2021), p. 20.

[Mur+19] W James Murdoch et al. "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" 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*. 2016, pp. 1135–1144.