Metrics for Evaluating Interface Explainability Models for Cyberattack Detection in IoT Data

Amani Abou Rida[1,3], Rabih Amhaz[1,2], and Pierre Parrend[1,3]

[1] ICube - Laboratoire, des sciences de l'ingénieur, de l'informatique et de l'imagerie, UMR 7357, Université de Strasbourg, CNRS, 67000, Strasbourg, France
[2] ICAM, site de Strasbourg - Europe, 67300 Schiltigheim, France
[3] Laboratoire de Recherche, de L'EPITA (LRE), 14-16 rue Voltaire, 94270 Le Kremlin Bicêtre, France

Abstract. The importance of machine learning (ML) in detecting cyberattacks lies in its ability to efficiently process and analyze large volumes of IoT data, which is critical in ensuring the security and privacy of sensitive information transmitted between connected devices. However, the lack of explainability of ML algorithms has become a significant concern in the cybersecurity community. Therefore, explainable techniques are developed to make ML algorithms more transparent, thereby improving trust in attack detection systems by its ability to allow cybersecurity analysts to understand the reasons for model predictions and to identify any limitation or error in the model. One of the key artifacts of explainability is interface explainability models such as impurity and permutation feature importance analysis, Local Interpretable Model-agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP). However, these models are not able to provide enough quantitative information (metrics) to build complete trust and confidence in the explanations they generate. In this paper, we propose and evaluate metrics such as reliability and latency to quantify the trustworthiness of the explanations and to establish confidence in the model's decisions to accurately detect and explain cyberattacks in IoT data during the ML process.

Keywords: Explainability · Trust · ML · Cybersecurity · Cyber-attacks

1 Introduction

Explainability is a crucial factor for establishing trust in ML models [Gui+18]. According to the High-Level Expert Group on Artificial Intelligence (AI) established by the European Union, trust is a critical factor in promoting the development and adoption of AI technologies [AI19a]. Trust is defined as the belief that an AI system will perform as expected and compliant with ethical and legal norms. To achieve trust in AI, the group [AI19a] recommends considering several properties, including transparency, accountability, reliability, safety, and privacy. However, without the ability to understand and interpret the decisions made by the model, trust cannot be established. This is where explainability comes in. It enables users to understand how the model arrived at its decisions, and therefore plays a crucial role in building trust.

In addition, trust and explainability are critical for the effectiveness of ML models in cybersecurity. As cyber threats continue to evolve in IoT data, it is crucial to prioritize explainability since it allows security analysts to comprehend the complex interactions between devices [Mol20], which can be difficult due to the distributed nature of IoT systems. With a better understanding of the reasons behind detection, analysts gain valuable insights which help them improve the detection process. However, achieving explainability using ML models in IoT data can be a complex task. To address this challenge, we extend on the framework for explainability in ML developed by Arrieta [Arr+20]. The framework identifies five key artifacts of explainability that are crucial for ensuring transparency throughout the data analysis process: data traceability [Mor+21], model understandability [Mur+19], output comprehensibility [Fer+19], interface explainability [Fer+19], and human interpretability [Arr+20]. Among these artifacts, interface explainability is the key for building trust and for increasing the adoption of ML systems. While models such as impurity feature importance, permutation feature importance analysis, LIME, and SHAP provide intuitive explanations, they have limitations in offering complete quantitative information and facilitating direct comparisons between explanations.

We therefore propose a novel approach for evaluating the quality of interface explainability models, with a specific focus on reliability of features and latency for providing explanations. We first evaluate existing metrics for interface explainability models, then develop new ones that can improve decision-maker's trust and confidence in the models they select. We aim at providing decision-makers with a solid foundation for selecting and evaluating interface explainability models that are transparent, trustworthy, and effective in making critical decisions. To accomplish this objective, we address the following question: How to evaluate quantitatively the explanations provided by interface explainability models to ensure trust in the data analysis process?

The paper is organized as follows. Section 2 introduces the state of the art, and Section 3 shows the benchmark for evaluating the quality of interface explainability models. Section 4 presents the datasets used and the implementation. Section 5 provides the evaluations and Section 6 discusses the significance of obtained results. Section 7 concludes this work.

2 State of the art

2.1 Trusted AI

Definition of Trusted AI According to the High-Level Expert Group on AI [AI19a], appointed by the European Commission, Trusted AI refers to artificial intelligence systems that meet specific properties and are developed and utilized in a transparent, ethical, and responsible manner.

Properties of Trusted AI In April 2019, the High-Level Expert Group on AI [AI19a] released Ethics Guidelines for Trustworthy Artificial Intelligence, which outlines seven key properties that AI systems should meet to be considered trustworthy. These properties include respecting human autonomy and fundamental rights, ensuring system robustness and safety, protecting the privacy and data governance, promoting transparency, avoiding bias and discrimination, considering societal and environmental well-being, and holding AI systems accountable for their outcomes. Moreover, the guidelines offer a specific assessment list to verify compliance with each property. This list includes measures such as fundamental rights impact assessments, cybersecurity testing, privacy protection, transparency and explainability, non-discrimination and fairness, and impact assessment tools.

Explainability for Trusted AI The High-Level Expert Group on AI [AI19a] emphasizes that incorporating transparency and explainability into AI systems is essential for ensuring their trustworthiness and promoting wider societal acceptance. By promoting a greater understanding of how AI systems make decisions and producing more trustworthy outcomes, transparency and explainability increase the social acceptance of AI and mitigate concerns related to issues such as bias, discrimination, and privacy.

2.2 Explainable AI

Definition of Explainable AI Arrieta [Arr+20] has defined Explainable Artificial Intelligence (xAI) in the context of ML models as follows:" Given an audience, an explainable Artificial Intelligence model produces details or reasons to make its functioning clear or easy to understand."

Artefacts for Explainability in the Data Analysis Process The data analysis process in ML involves five key components, as identified by Arrieta [Arr+20]: input data, model, output, user interface, and the human element. Input data refers to the information provided to the ML model to produce results. The user interface includes explanations or visualizations that allow humans to interact with and understand how the model works and how the results are derived. The human element involves the people who use the model and interpret the results by making decisions based on the insights generated from the data analysis. To achieve transparency in the data analysis process, Arrieta highlights the importance of incorporating five key artifacts of explainability: traceability, understandability, comprehensibility, explainability, and interpretability [Arr+20]. Traceability ensures the ability to track the origin and lineage of data used to train a model [Mor+21]. Understandability refers to the characteristic of a model to make its function understandable to humans without any need to explain its internal structure [Mur+19]. Comprehensibility is the ability of a learning algorithm to represent its learned knowledge in a way that is understandable to humans [Fer+19]. Explainability is the property of an AI system that enables it to provide an interface between humans and a decision-maker that accurately represents the decision-making process while being understandable to humans [Fer+19]. Interpretability is the ability to explain or provide meaning in understandable terms to a human [DK17].

Interface Explainability Models Interface explainability models in ML enhance the interpretability and transparency of complex models by providing insights into their decision-making processes. To achieve this, researchers have developed various techniques such as impurity and permutation-based feature importance analysis [AI19b; HMZ21], as well as model-specific methods like LIME [RSG16] and SHAP [LL17]. In this section, we highlight the strengths of these models and explore how they work. We also compare these techniques according to the data they use, the ML models they work on, the type and level of explanation they provide, and their limitations as shown in Table 1. Impurity Feature Importance (FI) and Permutation Feature Importance (PFI) are metrics used to measure feature importance in ML models. Their objective is to identify the most important features in the model's predictions by evaluating their impact on model performance. Impurity FI calculates the reduction in impurity that occurs when a feature is used to split the data and selects the feature with the largest decrease in impurity as the most important [AI19b]. PFI shuffles a feature's values to break its relationship with the target variable and measures the decrease in the model's score, with the feature having the largest drop in score considered the most important [HMZ21]. LIME is a method used to explain black-box ML model predictions locally, by showing the contributions of each feature to the prediction for a specific instance or a subset of the data. It generates new instances by sampling a neighborhood around the instance being explained and applies the model to it, weighting the generated instances based on their distances to the instance being explained. This results in a linear model that provides an understandable explanation of the black-box model's behavior for the instance being explained. The SHAP technique explains ML model output by attributing the prediction to input features, providing a consistent way of computing feature importance globally and locally. It uses Shapley values from game theory to distribute the effect of a feature fairly among all input features, resulting in a feature importance score. By repeating this process for all input features, SHAP provides a final score indicating the relative importance of each feature in the model's prediction.

Metrics for Explainable AI Metrics are critical for assessing the effectiveness of XAI systems [AI19a]. They allow for a meaningful comparison of how well a model fits the definition of explainability. Arrieta [AI19a] highlights the significance of metrics in evaluating the impact of explanations on the trust and reliance of the audience, as well as the need for concrete tools to compare the explainability between different models. Hoffman [Hof+18] emphasize the importance of developing clear and effective measurement concepts for evaluating the effectiveness of explainable systems. Although their framework identifies key questions about measuring effectiveness, it does not include specific quantitative measures or evaluation methods. Sovrano [Sov+22] provide a qualitative analysis

Comparison	**Impurity Feature Importance (FI)**	**Permutation Feature Importance (PFI)**	**LIME**	**SHAP**
Data used	Trained model	Trained model	Any datapoint	Any datapoint
Model types	Only trees - Model specific	Only trees - Model agnostic	Any - Model agnostic	Different explainer's types - Model agnostic
Explainability Level	Global	Global	Local	Global and local
Explainability Type	feature importance, visualization	feature importance, visualization	Simplification, visualization	feature relevance, visualization
Limitations	Computed on training set statistics and therefore do not reflect the ability of feature to be useful to make predictions that generalize to the test set (data), Unsuitable for linear models and for continuous features (only suitable for tree models) (ML model)	Unsuitable to explain time series models or when there are strongly correlated features (data), Do not provide information on how the feature affects the model's predictions for specific instances or subsets of the data (global explanation) (data and output), Does not account for any interactions between features (feature relevance)	Does not provide a complete picture of the model's behavior for the entire dataset (local explanation) (data and output), Depending on the number of fake instances you generate and the kernel width you select this introduces inaccuracies and leads to a loss of information (data and output)	Cannot be used to make statements about changes in prediction for changes in the input (data and output), Difficult to compare explanations from different ml models (each model requires a different type of explainer) (ML model), Large computational time because the training time grows exponentially with the number of features (output)

Table 1: Compare Interface Explainability models in ML

of metrics for evaluating the quality of explainability in AI systems. However, they do not provide specific quantitative measures or thresholds for each metric.

2.3 Explainability for Cybersecurity in IoT data

Detecting cyber attacks in IoT systems is crucial for maintaining the confidentiality, availability, and integrity of information transmitted over the Internet. ML algorithms are useful in handling the complexity of data generated from various sources and adapting to changing attack patterns [SS20]. However, the lack of explainability of ML algorithms has become a significant concern in the cybersecurity community. As highlighted in [Sri+22], explainability helps cybersecurity experts understand the reasons behind cyber attacks and detect any biases or vulnerabilities in their security systems. However, no research has compared the effectiveness of different explainability models in achieving these goals. Therefore, it is essential to identify and evaluate the most effective explainability model to achieve intuitive explanations of ML models' behavior in cybersecurity.

3 Metrics for evaluating the quality of Interface Explainability Models

Reliability between different Interface Explainability Models Reliability is a crucial aspect of interface explainability models, which refers to the consistency and similarity of feature importance scores across different models. If the output of various interface explainability models is consistent and similar, the user can have more confidence in the explanations provided by these models. Conversely, if they are inconsistent or differ widely between models, users may have difficulty interpreting the results and identifying the important features. We propose two metrics to measure the reliability of interface explainability models. The first metric is Top5Ratio, measures the similarity between the top 5 most important feature scores generated by different interface explainability models and is represented according to this formula:

$$\text{Top5Ratio} = \frac{|\text{Top}_5(score(x, m, iexp)) \cap \text{Top}_5(score(x, m, iexp'))|}{5} \tag{1}$$

This metric indicates the similarity between important features across different interface explainability models. The Top5Ratio metric is calculated by identifying the top 5 important features that consistently appear in the top 5 across all interface explainability models based on their feature scores, where x is the data, m is the ML model, *iexp* and *iexp'* are different interface explainability models. The top5Ratio metric produces results between 0 and 1, where 1 represents complete agreement between interface explainability models in terms of the top 5 important features and their scores, and 0 represents no agreement. To further assess the consistency of the feature scores across different models, we propose the average reliability metric according to this formula:

$$\text{Average Reliability} = \frac{1}{N}\sum_{i=1}^{N} |score_i(x, m, iexp) - score_i(x, m, iexp')| \tag{2}$$

This metric gives a measure of how much the feature importance scores differ between the two models. Then, computing the absolute distance between the average reliability scores, gives us a measure of how consistent the two models are in terms of their overall reliability. A lower absolute distance between the average reliability of interface explainability models indicates greater consistency, meaning that the two interface explainability models are more similar in terms of their reliability scores. A higher absolute distance indicates that the two models are less reliable.

Latency between different Interface Explainability Models In the context of providing explainability for cyber-attack models, latency refers to the amount of time it takes for the interface explainability model to generate explanations. The objective of the latency metric is to evaluate the efficiency of the explanation model in generating explanations for cyber-attacks. The metric aims to balance the need for quick and accessible explanations with the need for accurate and comprehensive explanations. It measures the time required for the explanation model to generate an explanation, with the time measured in units such as seconds. Local explanations, such as LIME [RSG16], provide insights into the model's decision-making process for each individual data entry in the dataset by analyzing the features that are most important for the model's decision for that specific data entry. The latency metric measures the time taken to generate a local explanation for each data entry in the dataset. The formula below computes the average time taken per instance over the dataset.

$$\text{Latency} = \frac{1}{N}\sum_{i=1}^{N}(T_{end} - T_{start}) \tag{3}$$

where N is the number of instances in the dataset, and T_{end} and T_{start} are the end and start times for generating the explanation for instance i.

For global explanations, we calculate the time taken to generate the explanation for the entire dataset, given by the formula: Latency $= T_{end} - T_{start}$

This metric measures the average time taken by the interface explainability model to generate explanations for cyber-attacks. A lower value of Latency indicates that the explanation model generates explanations more quickly, while a higher value of Latency indicates slower explanations.

4 Datasets and Implementation

Datasets To evaluate the explainability of cyber-attacks in ML, we used three datasets created by the Cyber Security Research Group at the University of New South Wales [Boo+21; Kor+19; MS16]. These datasets include Ton-IoT [Boo+21], BoT-IoT [Kor+19], and UNSW-NB15 [MS16], and contain both normal traffic and various types of cyber-attacks with varying sizes and features.

Implementation To implement explainability for cyber-attacks in ML, we utilize Jupyter Notebook running on the Larry server of HPE DL385 generation 10+. The server is equipped with two AMD EPYC 7552 48-Core Processors and 3 TB of RAM, providing sufficient computing power to perform the required analyses.

5 Evaluation

Comprehensibility of Output Performance Comprehensibility of output performance is essential in building trust in ML models, as end-users are more likely to trust and accept the results when they can understand and interpret the output of the model. To ensure the performance output is easily comprehensible, we utilized a range of metrics to evaluate the performance of various multi-classification models, including CART, Random Forrest, XGBoost, and MLP, on the datasets mentioned in Section 4. These metrics comprised unbalanced accuracy, MCC, accuracy, precision, recall, F1 score, True Negative Rate (TNR), loss, and multi-classification fit and pred time. Our evaluation revealed that XGBoost achieved the highest accuracy in detecting cyber-attacks in the Ton-IoT and UNSW-NB15 datasets, while CART is the most accurate model for the BoT-IoT dataset. Additionally, CART exhibit the shortest prediction time among all the models evaluated, for all the datasets as shown in Table 2.

Interface Explainability Models To increase trust in the model's predictions of XGBoost, we used four different interface explainability models: FI, PFI, LIME, and SHAP as shown in Fig 1. The FI and PFI models provided feature importance scores that contributed to XGBoost's predictions, while the LIME model simplified individual predictions to highlight the most important factors in predicting "MITM" attacks and reveal insights into the model's decision-making process. In contrast, the SHAP model presented a summary plot of the

Dataset	**Ton-IoT (45 features) num of classes = 10**				**UNSW (45 features) num of classes = 10**				**BoT-IoT Nb15 (46 features) num of classes = 4**			
Learning Models	Random Forest	XGBoost	CART	MLP	Random Forest	XGBoost	CART	MLP	Random Forest	XGBoost	CART	MLP
Precision	0.61369	**0.99991**	0.99981	0.65621	0.72632	**0.87941**	0.85395	0.50275	0.99989	0.99999	**1**	0.92654
Recall	0.77512	**0.99991**	0.99981	0.0484	0.73545	**0.87858**	0.84999	0.52941	0.99989	0.99999	**1**	0.84948
TNR	0.16053	**0.99798**	0.99581	0	0	**0.02264**	0.01893	0	0.99987	0.99999	**1**	0.82715
Accuracy	0.77512	**0.99991**	0.99981	0.0484	0.73545	**0.87858**	0.84999	0.52941	0.99989	0.99999	**1**	0.84948
F1 score	0.68192	**0.99991**	0.99981	0.01446	0.66093	**0.87161**	0.85185	0.46165	0.99986	0.99999	**1**	0.86858
Balanced Accuracy	0.38676	**0.99984**	0.99973	0.10115	0.29817	**0.62515**	0.59959	0.23236	0.79865	0.99999	**1**	0.48904
Mcc	0.55888	**0.99984**	0.99967	0.02514	0.69429	**0.84705**	0.81028	0.41077	0.99953	0.99997	**1**	0.62287
loss	0.55888	**0.00008**	0.00018	0.95159	0.26454	**0.12141**	0.15001	0.41077	0.00011	0.000005	**0**	0.15051
Fit time sec	12.4332	483.81	**0.00008**	84.0137	0.25997	1081.45	**1.6835**	53.4782	26.661	104.101	**0.98176**	134.22
Pred time sec	0.93504	0.06234	**0.01611**	0.23403	0.41471	0.056119	**0.01842**	0.10782	0.73737	0.047371	**0.02597**	0.25997

Table 2: Comprehensibility of output performance for CART, Random Forrest, XGBoost, and MLP models for different datasets

most important features ranked in descending order based on their impact on the prediction for each attack class. These models provided different perspectives on XGBoost's behavior.

Metrics for evaluating Interface Explainability Models

Reliability To evaluate the reliability of different interface explainability models, we first calculate the top5 ratio to determine the similarity of the five most important features in the Ton-IoT dataset, as shown in Fig. 2a. The results indicate that SHAP has a higher ratio compared to FI and PFI, indicating greater feature intersection among the most important features. We further analyze the reliability of these features by calculating the average reliability for each feature across all models, as illustrated in Fig. 2b. When comparing the reliability of different interface explainability models, we calculated the average difference between each model and the rest. The results show that SHAP has the smallest average difference of 0.02 compared to the other interface explainability models, indicating higher reliability. On the other hand, FI has an average difference of 0.116, PFI has an average difference of 0.134, and LIME has an average difference of 0.142 when compared to the other interface explainability models. These findings are depicted in Figure 3.

Latency To compare the latency of different interface explainability models, we present a 3D plot that shows the tradeoff latency, accuracy, and time performance. Fig.4 illustrates this tradeoff by showing the latency of the four interface explainability models, with respect to the accuracy and fit time for the four multi-classification models.

6 Discussion

Comprehensibility of Output Performance The evaluation of the comprehensibility of output performance revealed that XGBoost can accurately detect a large proportion of actual attacks while minimizing false positives, which refer to non-attacks being wrongly classified as attacks. While XGBoost showed superior accuracy, CART was faster and achieved comparable accuracy to XGBoost.

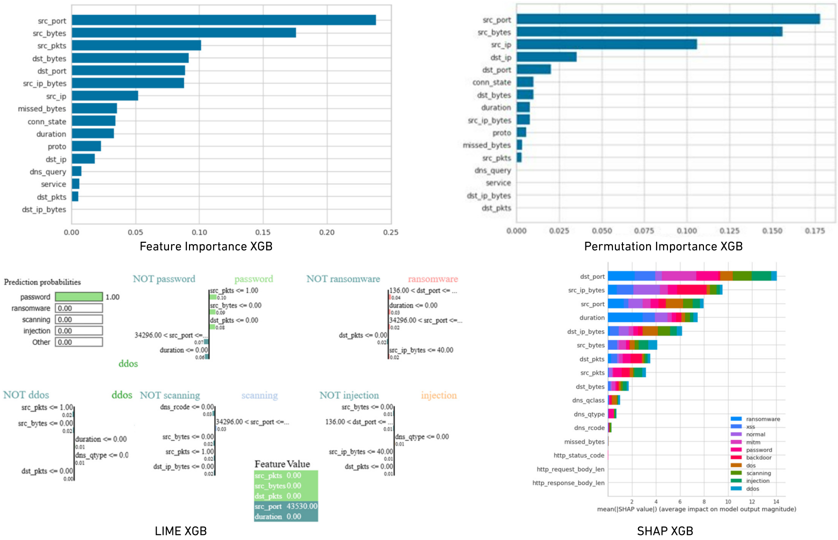


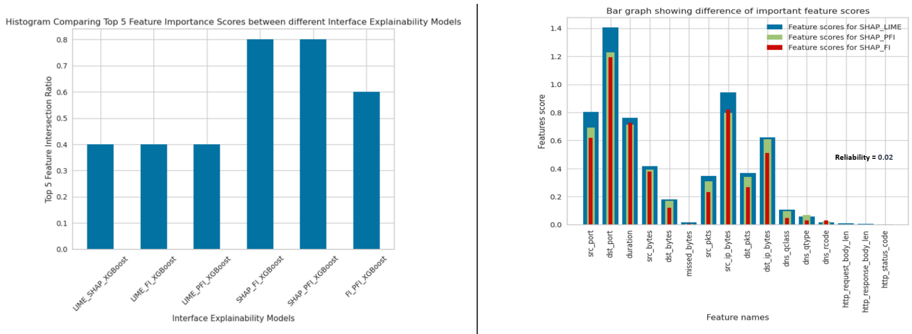
Fig. 1: Interface Explainability Models using XGBoost multi-classification model on Ton-IoT dataset

This highlights the significance of striking a balance between accuracy and time against cyber-attacks.

Interface Explainability Models The examination of various interface explainability models has provided distinct perspectives on how XGBoost predicts results. While FI and PFI models offer a global explanation of the overall model performance, they do not provide any insights into the classification of specific classes. On the other hand, LIME provides a local explanation of how the "MITM" class is classified for a particular instance, but not for the entire dataset. In contrast, SHAP provides an XGBoost summary plot that highlights the feature relevance scores. The visualization indicates that the feature "dst port", which is situated at a high level in the plot, resulted in the classification of "Ransomware" more frequently than the other class types. This demonstrates that the feature relevance scores and visualization provided by SHAP offer valuable insights into the model's predictions.

Metrics for evaluating Interface Explainability Models

Reliability Despite the variations in methodology and algorithms used by different interface explainability models, we can obtain similar important features when comparing them. This should provide users with confidence in the explainability model they choose. The top5Ratio metric shows that SHAP produces similar most common important features, even though it works differently to



(a) Exploring the Top5Ratio for Common Features in Interface Explainability Models for XGBoost Multi-Classification Model

(b) Average Reliability of SHAP's Feature Scores in Comparison with LIME, FI, and PFI

Fig. 2: Interface Explainability Model Comparison for XGBoost Multi-Classification on Ton-IoT dataset: Top 5 Ratio and Average Reliability Analysis

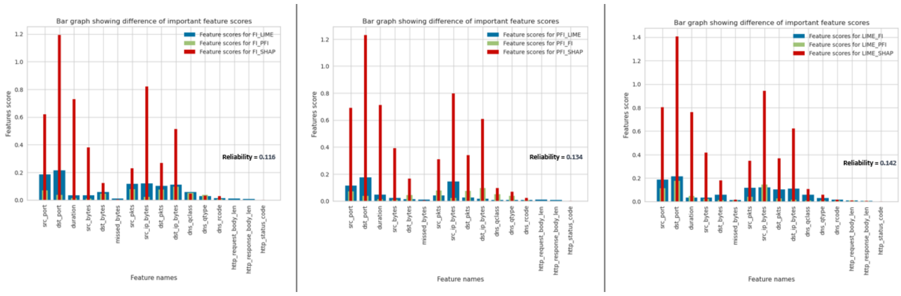


Fig. 3: Average Reliability Comparison of Feature Scores for different Interface Explainability Models

compute them. We further analyzed the difference in features between the different models and found that SHAP consistently performed better, indicating that we can rely on it for cybersecurity explanations.

Latency Given the time-sensitive nature of cyberattacks, the tradeoff between latency, accuracy, and time performance is crucial for classifying and explaining attacks. While XGBoost has the best accuracy, its latency for detecting an attack is slower compared to other approaches. Similarly, SHAP has a slower latency for providing explanations compared to FI and LIME. It is essential to balance these factors to quickly and accurately classify and explain an attack, allowing incident responders to take appropriate actions in a timely manner.

7 Conclusions and Perspectives

The reliability and latency metrics proposed in this paper serve as a strong foundation for evaluating interface explainability models in detecting cyber-attacks.

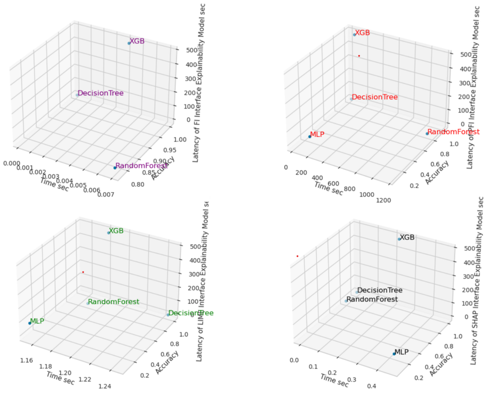


Fig. 4: Latency of different Interface Explainability Models using different multi-classification models on Ton dataset

These metrics ensure transparency and trust by demonstrating consistency and similarity in feature scores across various models, and highlighting the tradeoff between accuracy, latency, and time performance. However, to comprehensively evaluate interface explainability models, additional metrics need to be developed in the data analysis process, such as stability when data changes, similarity across ML models with different parameters, and clarity for human interpretability. Moreover, we plan to enhance the data visualization and dashboard for our metrics and provide a more detailed methodology for the training and testing process to facilitate a better understanding of the results and make our research transparent and easily reproducible for other researchers.

References

[AI19a] HLEG AI. *High-level expert group on artificial intelligence*. 2019.

[AI19b] Susan Athey and Guido W Imbens. “Machine learning methods that economists should know about”. In: *Annual Review of Economics* 11 (2019), pp. 685–725.

[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.

[DK17] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[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.

[Gui+18] Riccardo Guidotti et al. "A survey of methods for explaining black box models". In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[HMZ21] Giles Hooker, Lucas Mentch, and Siyu Zhou. "Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance". In: *Statistics and Computing* 31 (2021), pp. 1–16.

[Hof+18] Robert R Hoffman et al. "Metrics for explainable AI: Challenges and prospects". In: *arXiv preprint arXiv:1812.04608* (2018).

[Kor+19] Nickolaos Koroniotis et al. "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset". In: *Future Generation Computer Systems* 100 (2019), pp. 779–796.

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

[Mol20] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[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.

[MS16] Nour Moustafa and Jill Slay. "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set". In: *Information Security Journal: A Global Perspective* 25.1-3 (2016), pp. 18–31.

[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.

[Sov+22] Francesco Sovrano et al. "Metrics, explainability and the European AI act proposal". In: *J* 5.1 (2022), pp. 126–138.

[Sri+22] Gautam Srivastava et al. "XAI for cybersecurity: state of the art, challenges, open issues and future directions". In: *arXiv preprint arXiv:2206.03585* (2022).

[SS20] Yash Shah and Shamik Sengupta. "A survey on Classification of Cyber-attacks on IoT and IIoT devices". In: *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE. 2020, pp. 0406–0413.