Explainability-based Metrics to Help Cyber Operators Find and Correct Misclassified Cyberattacks

Robin Duraz, David Espes, Julien Francq, Sandrine Vaton

Explainability-based Metrics to Help Cyber Operators Find and Correct Misclassified Cyberattacks

Robin Duraz
robin.duraz@ecole-navale.fr
Chaire of Naval Cyberdefense, Lab-STICC
Brest, France

David Espes
david.espes@univ-brest.fr
UBO, Lab-STICC
Brest, France

Julien Francq
julien.francq@naval-group.com
Naval Group (Naval Cyber Laboratory, NCL)
Ollioules, France

Sandrine Vaton
sandrine.vaton@imt-atlantique.fr
IMT Atlantique, Lab-STICC
Brest, France

ABSTRACT

Machine Learning (ML)-based Intrusion Detection Systems (IDS) have shown promising performance. However, in a human-centered context where they are used alongside human operators, there is often a need to understand the reasons of a particular decision. EXplainable AI (XAI) has partially solved this issue, but evaluation of such methods is still difficult and often lacking. This paper revisits two quantitative metrics, Completeness and Correctness, to measure the quality of explanations, i.e., if they properly reflect the actual behaviour of the IDS. Because human operators generally have to handle a huge amount of information in limited time, it is important to ensure that explanations do not miss important causes, and that the important features are indeed causes of an event. However, to be more usable, it is better if explanations are compact. For XAI methods based on feature importance, Completeness shows on some public datasets that explanations tend to point out all important causes only with a high number of features, whereas Correctness seem to be highly correlated with prediction results of the IDS. Finally, besides evaluating the quality of XAI methods, Completeness and Correctness seem to enable identification of IDS failures and can be used to point the operator towards suspicious activity missed or misclassified by the IDS, suggesting manual investigation for correction.

CCS CONCEPTS

• **Security and privacy → Intrusion detection systems**; • **Human-centered computing → HCI design and evaluation methods**; • **Computing methodologies → Machine learning**.

KEYWORDS

Metrics; Explainable AI (XAI); Intrusion Detection Systems (IDSs)

SAFE '23, December 8, 2023, Paris, France

ACM ISBN 979-8-4007-0449-9/23/12…$15.00
https://doi.org/10.1145/3630050.3630079

ACM Reference Format:
Robin Duraz, David Espes, Julien Francq, and Sandrine Vaton. 2023. Explainability-based Metrics to Help Cyber Operators Find and Correct Misclassified Cyberattacks. In *Explainable and Safety Bounded, Fidelitous, Machine Learning for Networking (SAFE '23), December 8, 2023, Paris, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3630050.3630079

1 INTRODUCTION

XAI is an important tool in making ML-based IDSs more trustable and usable in real world scenarios. It is especially the case when IDSs are supporting and supported by human decision-making. Among the various XAI methods developed, feature importance methods are more commonly used for tabular data. LIME [24] and SHAP [17] are two methods often encountered when using XAI with IDSs [12, 18, 32]. In other cases, as suggested in [25], methods relying on inherently interpretable ML algorithms are employed [1, 6, 30].

However, XAI is still an emerging field and suffers from various difficulties. Most notably, despite research trying to solve this issue [9, 11, 21], XAI methods are still lacking evaluation metrics used as a standard [10], especially in the context of IDSs [22, 33]. Furthermore, XAI methods can sometimes be mistaken [14] or manipulated [8]. In such a context, it is actually unclear which benefits XAI methods can provide, and research suggests that AI might be enough by itself [5, 26]. Moreover, when explanations are ostensibly wrong, it is actually unclear if the fault lies with the XAI methods or with the IDS. Possible solutions are relying on specific tests as in [2] or defining new metrics.

This paper revisits two metrics introduced as properties in [21]: *Completeness* and *Correctness*, and explore their relation with the number of important features returned by the XAI method. *Completeness* reflects the fact that explanations are sufficient to explain the cause of a prediction, i.e., all important causes are visible in the explanation. *Correctness* reflects the fact that the most important features returned by the XAI methods indeed have the biggest influence on the prediction given by the IDS. Besides serving as evaluation metrics for XAI methods, Completeness and Correctness can also be used as debugging tools to identify when the IDS is potentially mistaken. In a context where IDSs are used alongside human operators, XAI methods along with the two metrics can explain the decision process of an IDS and point out cases where it behaves abnormally, resulting in higher chances of making wrong

predictions, to instead help human operators by suggesting manual investigations to correct IDS misclassifications.

The rest of the paper is organized as follows: Section 2 presents related works. Section 3 describes the two metrics along with the proposed approach. Section 4 presents and analyzes the results. Finally, Section 5 discusses results and their limitations, and Section 6 concludes the paper and discusses future avenues of research.

2 RELATED WORK

XAI methods are generally used in a user-centric context, and as such, need to cater to users. Therefore, much work has been done to identify and define criteria of "good" explanations [9, 13, 19]. However, these criteria are often qualitative and focus on how well-received explanations are, and do not measure how much these explanations reflect the actual decision process of the explained ML method.

Quantitative metrics have also been used in an attempt to evaluate various properties of XAI methods such as Faithfulness, Robustness or Complexity [29]. Faithfulness of an explanation is a desirable property that describes the ability to capture the features used by a predictor [7], but can be quite complex to compute. However, other methods can also be used to compute metrics related to Faithfulness [16, 31]. Robustness measures the stability and consistency of a given XAI method [15, 28], while Complexity [15] ensures that explanation would be easily understandable by users. Finally, it can be interesting to measure how explanations coincide with ground truth [28, 32]. However, IDSs often rely on weak signals to detect less obvious cyberattacks, relying on many features. Feature importance methods can specify a number of features present in the explanation. While this does not matter for some applications, e.g., images, it is important in the context of IDSs and often forgotten when evaluating the quality of explanations.

Finally, incorrect explanations according to a given metric does not necessarily mean that the XAI method is to blame. Instead, it might be caused by the IDS being incorrect and behaving abnormally, because of the lack of data or simply from spurious patterns in the training data. In these cases, XAI methods might be used to correct either the dataset used in training [3, 23] or the IDS itself.

3 METHODOLOGY

In the context of IDSs, whether as a tool to help in human decision-making or to enable auditing, it is important for the cause of predictions to be understood. The XAI methods that will be used in this paper for explanations is LIME [24], a method that is based on a surrogate linear model to output feature importance. Code used to realize experiments, as well as instructions to reproduce experiments, are available on GitLab[1].

3.1 Metrics

Because verification by human operators is generally a time-consuming task, it is important for XAI tools to explain using the most important causes of a prediction, while hopefully ensuring that there are only a few causes. When viewing explanations, it might confuse the operator if causes that are more aligned with his knowledge are absent. As such, the first metric to consider is Completeness and relates to the fact that the explanation is self-sufficient, i.e., all features in the explanation are enough to explain a prediction and no other features are needed.

Another important property of explanations is their faithfulness to represent the behavior of the IDS. If features present in the explanation are not actually considered causes of the event by the IDS, it might instead mislead the user. As such, the second metric to consider is Correctness and relates to the fact that features present in the explanation are indeed important, i.e., their influence is higher for the given prediction than for other classes. Both metrics will be computed with different numbers of features to research the impact of this parameter on the quality of explanations.

3.1.1 Completeness. In order to test the Completeness of an explanation, non-important features are "deleted" (as in replaced by median values computed over the whole dataset). The algorithm to compute Completeness is described in Algorithm 1. To do so, an *IDS* that outputs prediction probabilities is needed, as well as the median value (*median*) for all features. An instance x that is explained is also given along with important features ($feat\_importance$) returned by the explanation of x's prediction. An explanation is thus deemed complete if the prediction using only important features is the same as the original prediction.

Algorithm 1: Completeness computation

```
Data: IDS, x, median, feat_importance
pred ← IDS.probas(x);
incomplete_x ← median;
incomplete_x[feat_importance] ← x[feat_importance];
if IDS.probas(incomplete_x) is pred then
    /* Explanation was complete */
    return True;
else
    return False;
end
```

3.1.2 Correctness. In order to test the Correctness of an explanation, features are "deleted" incrementally and the impact on output probabilities is measured. The explanation is correct if the relative change in probability is the highest for the predicted class. The algorithm to compute Correctness is described in Algorithm 2.

3.2 Datasets, ML and XAI algorithms

In order to test the influence of the number of features as well as the performance of the IDS on the quality of the explanations, three datasets were used: WADI [4], which is an Industrial Control System (ICS) dataset of a water plant, and CIC-IDS2017 [27] and UNSW-NB15 [20], two network traffic datasets. All three datasets were split using a stratified scheme into 70% train (60% and 10% validation) and 30% test sets.

For the WADI dataset, features such as Row, Date, Time and four other features that are missing all values were removed. The resulting dataset has 124 features. Attacks are named `Attack_i` (i

[1] https://gitlab.com/RobinKD/completeness-and-correctness-to-evaluate-xai-and-improve-ids

Algorithm 2: Correctness computation

Data: $model, x, median, feat\_importance$
$p \leftarrow model.probas(x)$;
$pred \leftarrow argmax(p)$;
$previous\_p \leftarrow p$;
/* max_di is used to retain max values of
deletion_impact ($|max\_di| = |labels|$) */
$max\_di \leftarrow [0, 0, ..., 0]$;
for $feature$ *in* $feat\_importance$ **do**
  $incomplete\_x[feature] \leftarrow median[feature]$;
  $incomplete\_p \leftarrow model.probas(incomplete\_x)$;
  /* Unit-wise division */
  $deletion\_impact = \frac{p}{incomplete\_p}$;
  for $i \leftarrow 0$ *to* $size(deletion\_impact)$ **do**
    $max\_di[i] = max(deletion\_impact[i], max\_di[i])$;
  end
  $previous\_p \leftarrow incomplete\_p$;
end
if $argmax(max\_di)$ *is* $pred$ **then**
  /* Explanation was correct */
  return True;
else
  return False;
end

from 1 to 15, e.g., `Attack_1`) and have different targets and objectives. They can either affect physical equipment, starting pumps, opening or closing valves, or manipulate sensor readings. For the UNSW-NB15 dataset, features such as IP addresses, timestamps, *attack_cat* were removed, while categorical features or features having a small number of unique values, were one-hot encoded. The resulting dataset has 229 features. For the CIC-IDS2017 dataset, two features and 5792 instances were removed because of problematic or missing values. A further eight features were removed because they only had one value. The resulting dataset has 70 features.

For the ML algorithm used as an IDS, the Neural Network (NN) algorithm is retained as a first experiment for multiple reasons. First, it is often one of the best performing algorithms. Secondly, NNs are also among the less inherently interpretable ML algorithms, thus the interest in explaining their predictions. The NN architectures used are those obtaining the highest accuracy on the three datasets. They are fully connected with six hidden layers of size 256, 512, 1024, 512, 256, 128, with a ReLU activation function.

To explain NNs, as well as compute Completeness and Correctness, LIME is used. This particular method has been chosen over other methods because this is the most extensively used compared to other similar feature importance methods, along with SHAP, but is more computationally efficient. Raw values of feature importances returned by LIME are used to compute XAI metrics, as shown in Algorithm 1 and Algorithm 2, where feature importances are given with the parameter *feat_importance*.

4 RESULTS

First, because performance of the IDS might impact the performance of XAI methods, it is important to evaluate the IDS with metrics such as Accuracy that represents the proportion of correctly classified instances. Therefore, Accuracy on the different datasets is reported in Table 1. More detailed Accuracy results are available in the Appendix.

Table 1: NN Accuracy on the three datasets

Dataset	WADI	UNSW-NB15	CIC-IDS2017
Accuracy (%)	99.96	97.92	99.62

4.1 XAI Method Evaluation with Completeness and Correctness

Completeness and Correctness were evaluated for each data point of each class for all three datasets, depending on the number of features asked of the explanation method. Results obtained are shown in Figure 1 with values for Completeness and Correctness, both representing the percentage of instances in each class having, respectively, a complete and correct explanation. Features are ordered by importance, i.e., retaining three features only keeps the three most important features. Because predicting as `Normal`, i.e., not an attack, seems to be the default behavior of the IDS, explanations would be complete regardless of the number of features, thus results for the `Normal` class have not been represented.

Figure 1a shows Completeness results on the WADI dataset. First, Completeness seems to be correlated with the number of features in most cases, with a higher number of features meaning a more complete explanation. Secondly, there are two exceptions to the previous observation. `Attack_8` and `Attack_9` have a very high Completeness value even with the lowest number of features returned. `Attack_9` (turns on a pump) is relatively simple to detect considering the feature `1_P_006`'s (showing the status of the pump) value that is normally 0, is 2 when that attack occurs. However, in some cases during `Attack_9`, the feature `1_P_006`'s value remains 0, causing both a prediction error, and the explanation to not be complete. In the case of `Attack_8` (opens the Motor Control Valve, MCV, 007), a high enough value ($value \geq 30$) for `2_MCV_007_CO` (showing opening percentage) also seems to be the only important feature. Finally, `Attack_13` (reduces a booster set point pressure) explanations are often not complete, because this attack is weakly impacting the whole water plant, thus needing many features to be detected. In Figure 1d showing Correctness on WADI, results seem to not be correlated with the number of features, and interestingly, Correctness results for each class tend to be relatively close to each class' Accuracy.

In the case of the UNSW-NB15 dataset, as shown in Figure 1b, explanations generally need less features to be complete than for other datasets. This probably means that some features in this dataset offer more discriminating power with regard to detecting some attacks. However, Completeness is often above Accuracy, which means these important features might be given more importance

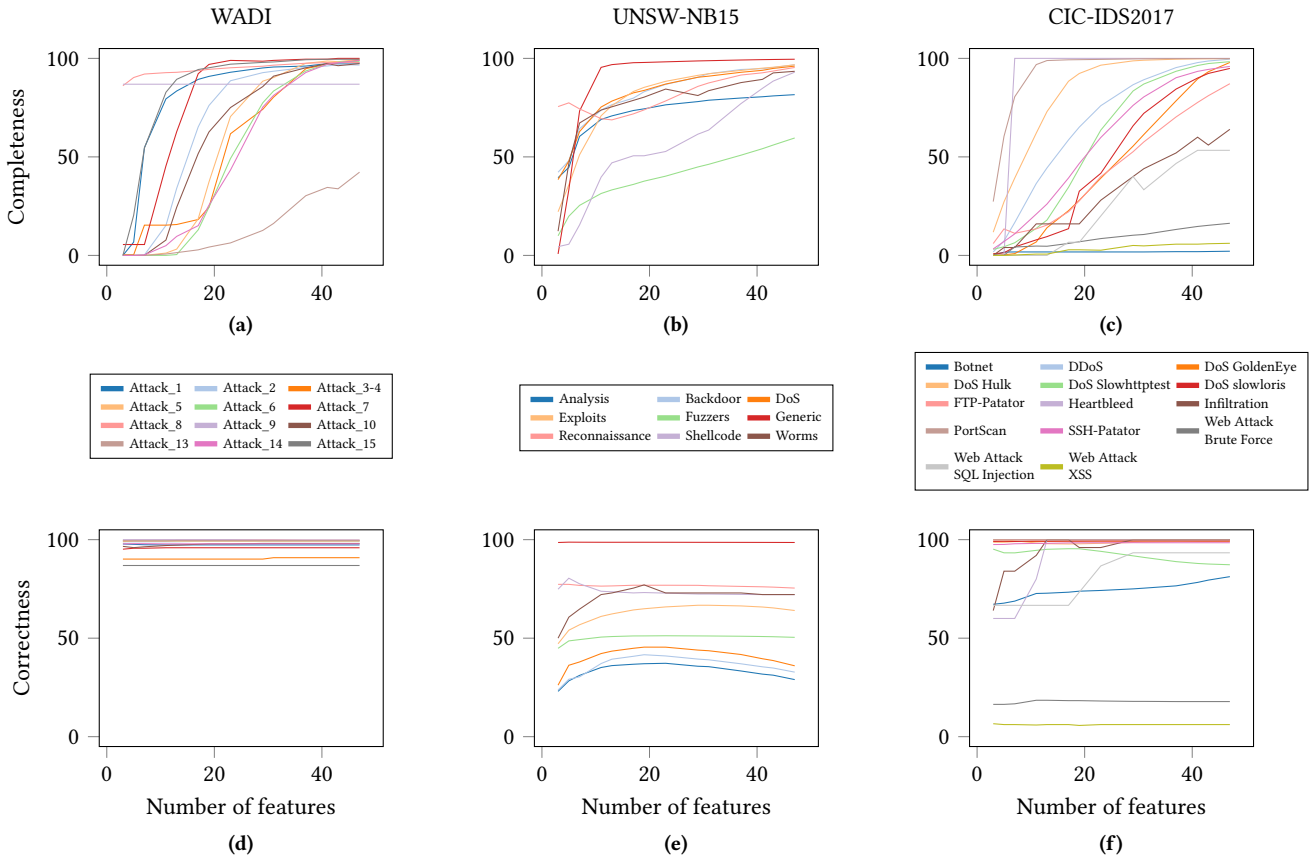

Figure 1: Completeness and Correctness results for each class on WADI (a) and (d), UNSW-NB15 (b) and (e), and CIC-IDS2017 (c) and (f).

than they actually have, thus sometimes pushing the IDS towards the wrong prediction. For Correctness, increasing the number of features seems to have a negative impact for some attacks, which seems a peculiar behavior. The most likely possibility is that important features in these cases tend to be more important for another class. Last but not least, similarly as Completeness, Correctness values are above class Accuracy for many classes, which means the IDS is often confidently wrong, e.g., Accuracy on Analysis and Backdoor is less than 10% whereas Correctness is higher than 25%, which is concerning.

For CIC-IDS2017, increase rate in Completeness seems disparate for the different attack classes. This means that attack complexities differ a lot for this dataset. Completeness remains close to 0% for `Web Attack XSS`, `Web Attack Brute Force`, and `Botnet`. For `Botnet` where Accuracy is around 60%, it probably means that a combination of many features is generally required to predict correctly these classes. For `Web Attack XSS` (and `Web Attack Brute Force`), it is probably a result of poor performance (Accuracy is 2% for `XSS`, 13% for `Brute Force`). Interestingly, the IDS also seems often confidently wrong in the case of `Web Attack SQL Injection` where Completeness reaches around 50% whereas Accuracy is 0%. The same behavior seems to be validated by Correctness results for Web Attacks.

Correctness does not seem to be impacted much by the number of features. Overall, it means that features in the explanation indeed have the biggest impact on the class predicted by the IDS, so explanations indeed properly reflect the IDS's decision process. Moreover, Correctness also seems to be highly correlated with performance on the different classes. This is important because it possibly means that incorrect explanations might be due to incorrect predictions, thus allowing to find IDS errors. Completeness, however, is very dependent on the number of features, except in cases where one or a few features make the prediction too obvious, e.g., `Attack_9` in WADI. Performance also seem to have an impact, e.g., `Web Attack XSS`, `Web Attack Brute Force` in CIC-IDS2017, although the impact is lower than for Correctness. This metric also shows that many attacks are generally complex and require many features to be properly explained. Therefore, relying on explanations to explain and validate predictions might be more time-consuming than expected for a human operator.

Furthermore, the IDS performance seem to heavily influence the ability to obtain complete and correct explanations. Attack classes that are poorly detected generally leads to their explanations being

incorrect and often also incomplete, e.g., `Web Attack XSS` and `Web Attack Brute Force` explanations are almost all incorrect and incomplete, regardless of the number of features.

4.2 Completeness and Correctness to identify IDS errors

Because Correctness seems to be highly correlated with class Accuracy, it is interesting to explore correlation between both metrics and prediction results of the IDS. In case of a high correlation, both metrics might be useful in detecting errors in prediction. Correlation with Completeness might provide additional information, especially at a high number of features, or when some features are by themselves the determining factor, e.g., for `Attack_8` and `Attack_9` in WADI. Results on the train set are shown in Figure 2.

For WADI, in two cases, correlation between Correctness and prediction results is equal to 0 because performance on the class is 100%, thus lowering artificially the correlation to 0 as long as one instance's explanation is not correct, which is the case here.

Overall, for Completeness, correlation with prediction results seems very dependant on model performance. The lower the performance, the lower the correlation. However, for `Attack_8` and `Attack_9` in WADI, correlation is (or reaches) 100% which means it could be used to correct all errors. For Correctness, the impact of model performance seems lower, but nonetheless still present, and correlation seems positive, or even highly positive, for all three datasets. Both metrics could thus be used to point out errors in prediction and would be effective in different cases.

To test the potential of both metrics to find errors on the test set, Completeness and Correctness are computed for uncertain predictions (uncertain means the probability of the second most likely class is superior to a threshold, to reduce unnecessary computation). If either Correctness or Completeness invalidates prediction, the instance is given to a human operator (considered as an oracle) to investigate. The number of features returned by explanations needs to be fixed for both metrics. It has been fixed at 40 for Completeness (because correlation prediction results seem higher with more features) and 30 for Correctness (the number does not matter much). Accuracy at 0.5 in threshold value represents the original Accuracy. As the threshold values decreases, Figure 3a shows the increase in Accuracy, while Figure 3b shows the required manual investigations. Manual investigations are mainly required for the `Normal` class. This is expected because many attacks are missed and classified as `Normal`, thus the need to investigate this class. For `DoS Hulk`, the class often possesses signal of other DoS attacks, but the

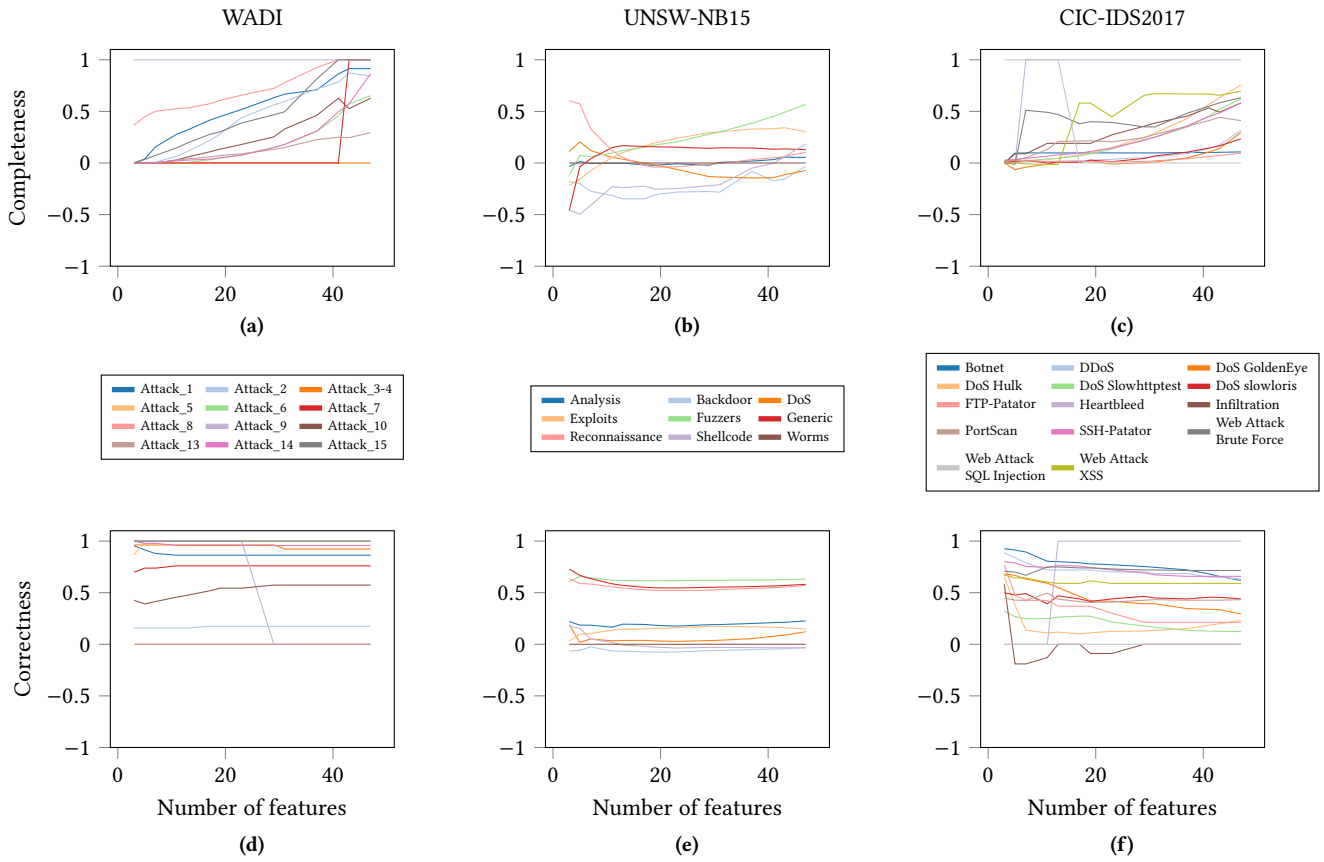

Figure 2: Completeness and Correctness correlation with errors in prediction for each class on WADI (a) and (d), UNSW-NB15 (b) and (e), and CIC-IDS2017 (c) and (f).

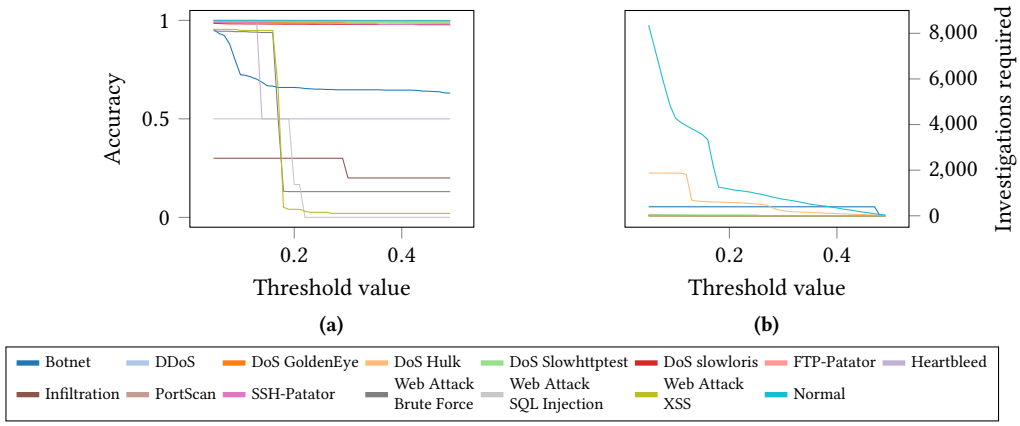

Figure 3: Possible gains in Accuracy (a) and Manual investigations required (b) for each class, with XAI pointing errors.

IDS is nevertheless correct, thus resulting in wasteful investigations. Interestingly, by pointing out potential errors with XAI for the CIC-IDS2017 dataset, there is a definite improvement for many classes compared to only using an IDS. For attacks where performance was originally low (below the 75% mark), the gain in Accuracy ranges between 5-10% to even 90% for `Web Attack XSS`. However, this creates a bigger load for a human operator, which can grow quickly as traffic increases and the chosen threshold decreases. Therefore, there is a need to find a correct trade-off between performance increase and human workload.

5 DISCUSSION AND LIMITATIONS

Completeness tends to show that a low number of features (because there is a limit to the amount of information a human user is able to handle) seems to generally not be enough to identify a specific class (explanations are not complete). This means that features potentially important or more understandable for the operator might not be present in the explanation. However, experiments were performed using NNs as IDSs and explanations would possibly be more sparse and thus more easily complete with other ML algorithms such as Decision Trees or NNs using dropout, possibly at the cost of lower performance.

Correctness, however, seems to be more correlated with prediction results (the prediction being correct or not) than it is correlated with the number of features. Correctness, and to a lesser extent Completeness, seem to be able to point out errors in prediction, but it requires a human to investigate. It would be interesting to see if these two metrics could be used to automate correction of the predictions, to reduce human workload.

Overall, the IDS's performance also impacts the ability of XAI methods to deliver sound and useful explanations. There are at least three possible causes of this behavior. First, the IDS's inferred definition of a class presents flaws or is too broad, which could be improved if the IDS can perform better. Secondly, important features might be shared between multiple classes, but remain more important for one of them. This might cause the explanations to be incorrect for the other classes for which the feature is also important. Finally, some instances might not correspond to what is expected of a specific class, possibly because of errors in data collection or labelling inconsistencies, e.g., the misclassified `Attack_9` instances in WADI.

6 CONCLUSION AND FUTURE WORK

Completeness has shown that usability of explanations might be heavily impacted by the number of important features provided to the user, because important causes of a prediction might not be present in an explanation. Correctness is much less impacted by the number of features, which means that features (with their values) present in an explanation are indeed more important for the given prediction than for other classes. Furthermore, Correctness tends to be more highly correlated with the results of an IDS's predictions, which can help in pointing out errors. When considering the prediction and its explanation with Completeness and Correctness results, an explanation of a prediction both complete and correct will be more trustable for a user. On the other hand, if it is either incomplete or incorrect, it might instead point to the prediction being wrong.

In future works, the usefulness of both metrics will be tested with other ML algorithms and potentially other XAI methods such as SHAP. Furthermore, both metrics will be researched to explore the potential of automating the correction of IDS errors.

ACKNOWLEDGMENTS

This work is supported by the Chair of Naval Cyber Defence and its partners Ecole Navale, ENSTA-Bretagne, IMT-Atlantique, Naval Group and Thales.

REFERENCES

[1] Jesse Ables, Thomas Kirby, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. 2022. Creating an Explainable Intrusion Detection System Using Self Organizing Maps. *CoRR* (2022).
[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. *CoRR* (2018).
[3] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging Tests for Model Explanations. *CoRR* (2020).
[4] Chuadhry Mujeeb Ahmed, Venkata Reddy Palleti, and Aditya P. Mathur. 2017. WADI. In *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks.*

[5] Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does Explainable Artificial Intelligence Improve Human Decision-Making? *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 8 (2021), 6618–6626.

[6] Giuseppina Andresini, Annalisa Appice, Francesco Paolo Caforio, Donato Malerba, and Gennaro Vessio. 2022. Roulette: a Neural Attention Multi-Output Model for Explainable Network Intrusion Detection. *Expert Systems with Applications* 201 (2022), 117144.

[7] Umang Bhatt, Adrian Weller, and José M. F. Moura. 2020. Evaluating and Aggregating Feature-Based Model Explanations. *CoRR* (2020).

[8] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations Can Be Manipulated and Geometry Is To Blame. *CoRR* (2019).

[9] Finale Doshi-Velez and Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. *CoRR* (2017).

[10] Anna Hedström, Leander Weber, Dilyara Bareeva, Daniel Krakowczyk, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M. C. Höhne. 2022. Quantus: an Explainable Ai Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *CoRR* (2022).

[11] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable Ai: Challenges and Prospects. *CoRR* (2018).

[12] Zakaria Abou El Houda, Bouziane Brik, and Lyes Khoukhi. 2022. "Why Should I Trust Your Ids?": an Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks. *IEEE Open Journal of the Communications Society* 3 (2022), 1164–1176.

[13] Janet Hui-wen Hsiao, Hilary Hei Ting Ngai, Luyu Qiu, Yi Yang, and Caleb Chen Cao. 2021. Roadmap of Designing Cognitive Metrics for Explainable Artificial Intelligence (XAI). *CoRR* (2021).

[14] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2017. The (Un)reliability of Saliency Methods. *CoRR* (2017).

[15] Ding Li, Yan Liu, Jun Huang, and Zerui Wang. 2022. A Trustworthy View on XAI Method Evaluation.

[16] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St. Jules, Xiao Yu Wang, and Alexander Wong. 2019. Do Explanations Reflect Decisions? a Machine-Centric Strategy To Quantify the Performance of Explainability Algorithms. *CoRR* (2019).

[17] Scott Lundberg and Su-In Lee. 2017. A Unified Approach To Interpreting Model Predictions. *CoRR* (2017).

[18] Shraddha Mane and Dattaraj Rao. 2021. Explaining Network Intrusion Detection System Using Explainable Ai Framework. *CoRR* (2021).

[19] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights From the Social Sciences. *CoRR* (2017).

[20] Nour Moustafa and Jill Slay. 2015. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 Military Communications and Information Systems Conference (MilCIS)*. 1–6.

[21] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2022. From Anecdotal Evidence To Quantitative Evaluation Methods: a Systematic Review on Evaluating Explainable Ai. *CoRR* (2022).

[22] Subash Neupane, Jesse Ables, William Anderson, Sudip Mittal, Shahram Rahimi, Ioana Banicescu, and Maria Seale. 2022. Explainable Intrusion Detection Systems (X-IDS): a Survey of Current Methods, Challenges, and Opportunities. *CoRR* (2022).

[23] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. 2021. Finding and Fixing Spurious Patterns With Explanations. *CoRR* (2021).

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* (2016).

[25] Cynthia Rudin. 2018. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *CoRR* (2018).

[26] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*.

[27] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy*. 108–116.

[28] Kashif Siddiqui and Thomas E. Doyle. 2022. Trust Metrics for Medical Deep Learning Using Explainable-AI Ensemble for Time Series Classification. In *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. 370–377.

[29] Sédrick Stassin, Alexandre Englebert, Géraldin Nanfack, Julien Albert, Nassim Versbraegen, Gilles Peiffer, Miriam Doh, Nicolas Riche, Benoît Frenay, and Christophe De Vleeschouwer. 2023. An Experimental Investigation Into the Evaluation of Explainability Methods. *CoRR* (2023).

[30] Mateusz Szczepanski, Michal Choras, Marek Pawlicki, and Rafal Kozik. 2020. Achieving Explainability of Intrusion Detection System by Hybrid Oracle-Explainer Approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*. 1–8.

[31] Syed Wali and Irfan Khan. 2021. Explainable AI and Random Forest Based Reliable Intrusion Detection system.

[32] Maonan Wang, Kangfeng Zheng, Yanqing Yang, and Xiujuan Wang. 2020. An Explainable Machine Learning Framework for Intrusion Detection Systems. *IEEE Access* 8 (2020), 73127–73141.

[33] Zhibo Zhang, Hussam Al Hamadi, Ernesto Damiani, Chan Yeob Yeun, and Fatma Taher. 2022. Explainable Artificial Intelligence Applications in Cyber Security: State-Of-The-Art in Research. *IEEE Access* 10 (2022), 93104–93139.

A DETAILED ACCURACY RESULTS

A.1 WADI

Table 2: NN Accuracy for each class on WADI (Part 1)

Class	Attack_1	Attack_10	Attack_13	Attack_14	Attack_15	Attack_2	Attack_3-4	Attack_5	Attack_6
Accuracy (%)	0.975	0.995	0.852	0.994	0.989	0.971	1.0	1.0	0.970

Values were truncated to the third decimal.

Table 3: NN Accuracy for each class on WADI (Part 2)

Class	Attack_7	Attack_8	Attack_9	Normal
Accuracy (%)	1.0	0.975	0.888	0.999

Values were truncated to the third decimal.

A.2 UNSW-NB15

Table 4: NN Accuracy for each class on UNSW-NB15

Class	Analysis	Backdoor	DoS	Exploits	Fuzzers	Generic	Reconn-aissance	Shellcode	Worms	Normal
Accuracy (%)	0.028	0.034	0.176	0.884	0.456	0.982	0.740	0.821	0.0	0.996

Values were truncated to the third decimal.

A.3 CIC-IDS2017

Table 5: NN Accuracy for each class on CIC-IDS2017 (Part 1)

Class	Botnet	DDoS	DoS GoldenEye	DoS Hulk	DoS Slowhttp-test	DoS slowloris	FTP-Patator	Heartbleed	Infiltration
Accuracy (%)	0.630	0.999	0.985	0.998	0.988	0.978	0.996	0.5	0.2

Values were truncated to the third decimal.

Table 6: NN Accuracy for each class on CIC-IDS2017 (Part 2)

Class	PortScan	SSH-Patator	Web Attack Brute Force	Web Attack SQL Injection	Web Attack XSS	Normal
Accuracy (%)	0.999	0.977	0.130	0.0	0.020	0.996

Values were truncated to the third decimal.