

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/379413078>


Towards Accountable and Resilient AI-Assisted Networks: Case Studies and Future Challenges

Conference Paper · June 2024

CITATIONS
0

READS
234


7 authors, including:



Chamara Sandeepa
University College Dublin

19 PUBLICATIONS 69 CITATIONS


SEE PROFILE



Thulitha Theekshana Senevirathna
University of Ruhuna

10 PUBLICATIONS 13 CITATIONS


SEE PROFILE



Bartlomiej Siniarski
University College Dublin

20 PUBLICATIONS 45 CITATIONS

SEE PROFILE



Manh-Dung Nguyen

11 PUBLICATIONS 690 CITATIONS

SEE PROFILE

Towards Accountable and Resilient AI-Assisted Networks: Case Studies and Future Challenges

Shen Wang*, Chamara Sandeepa*, Thulitha Senevirathna*, Bartlomiej Siniarski*,
Manh-Dung Nguyen†, Samuel Marchal‡ and Madhusanka Liyanage*

*School of Computer Science, University College Dublin, Dublin, Ireland.

Email: {shen.wang, bartlomiej.siniarski, madhusanka}@ucd.ie, {abeysinghe.sandeepa, thulitha.senevirathna}@ucdconnect.ie

†Montimage, France, email: manhdung.nguyen@montimage.com

‡ VTT Technical Research Center of Finland and Aalto University, Espoo, Finland, email: samuel.marchal@vtt.fi

Abstract—Artificial Intelligence (AI) will play a critical role in future networks, exploiting real-time data collection for optimized utilization of network resources. However, current AI solutions predominantly emphasize model performance enhancement, engendering substantial risk when AI encounters irregularities such as adversarial attacks or unknown misbehaves due to its “black-box” decision process. Consequently, AI-driven network solutions necessitate enhanced accountability to stakeholders and robust resilience against known AI threats. This paper introduces a high-level process, integrating Explainable AI (XAI) techniques and illustrating their application across three typical use cases: encrypted network traffic classification, malware detection, and federated learning. Unlike existing task-specific qualitative approaches, the proposed process incorporates a new set of metrics, measuring model performance, explainability, security, and privacy, thus enabling users to iteratively refine their AI network solutions. The paper also elucidates future research challenges we deem critical to the actualization of trustworthy, AI-empowered networks.

Index Terms—AI, Security, Privacy, Explainability, Malware, Traffic Classification, Federated Learning

I. INTRODUCTION

AI has emerged as one of critical enabling technologies for the future networks, providing groundbreaking capabilities in network automation, management, and security [1]. Due to its inherent capacity for learning, adapting, and predicting, AI can address the complex challenges associated with future networks including high-frequency millimeter wave propagation, dynamic spectrum management, and efficient energy utilization. However, with the increasing ubiquity and sophistication of AI, the demand for transparency and interpretability has escalated beyond the improvement of the AI/ML model performance only. The intrinsic “black-box” decision process of AI models will lead to uncontrolled negative network automation and unawareness of AI attacks. Explainable AI (XAI) arises recently to ensure trustworthiness and reliability in future networks [2], [3] fostering an understanding of AI decision-making processes in network management.

However, most of the existing work using XAI for trustworthy networks focus on specific task [4], [5], which is difficult to generalise to wider network applications in a stakeholders-involved and ever-evolving environment. This paper attempts to extract a common high-level process from initial experi-

ments on three typical AI applications for future networks: encrypted network traffic classification, malware detection, and federated learning. Specifically, we make the following contributions:

- Propose a high-level process with active engagement of internal stakeholders and a full set of metrics for assessing the model’s accountability and resilience feature.
- Demonstrate the initial results when applied to three aforementioned AI use cases. Those results show that a better AI trade-off objective can be achieved iteratively by more transparent indicators of the system’s accountability and resilience.
- Present research challenges for XAI deployment in future networks. Those research challenges are summarised from these three initial use cases and likely play a vital role in the related research community.

II. PROPOSED PROCESS FOR ACCOUNTABLE AND RESILIENT AI-ASSISTED NETWORK

A. Stakeholders

The proposed process actively engages the internal stakeholders of the system (e.g., developers, testers, system operators, etc.) who have domain knowledge about the targeted AI system. We focus on the internal stakeholders only as it has been reported [6] that most existing successful XAI are deployed for technical experts debugging their AI systems, rather than enhancing the trust of the end-users. As shown in Figure 1, use case stakeholders trigger the process by setting up the trade-off objectives, while terminating the process when the pre-set objectives are achieved, using the proposed set of resilience metrics together with the existing model utility metrics and posthoc XAI methods.

B. Four-step iterative process

As shown in Figure 1, our process firstly collects trade-off objectives from the user case stakeholders in terms of how they expect the importance among multiple targets: model utility, accountability, resilience, and privacy. It is important to note that there is no flawless AI solution that can simultaneously achieve optimal model performance, resource efficiency, accountability, and resilience. For instance, if a high

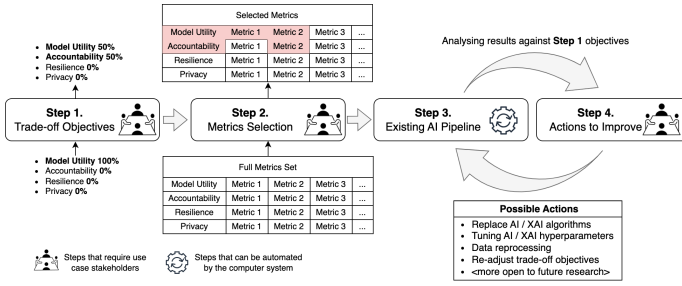


Fig. 1. Proposed Four-Step Process for Accountable and Resilient AI-Assisted Networks

explainability is required, it often necessitates compromising the model's performance to some extent. Secondly, a list of key metrics will be chosen according to the pre-set trade-off targets and the implementation feasibility of the use case. Thirdly, these metrics will be used to comprehensively test how well the existing AI system performs against the trade-off targets that do not only consider model utility as before. Finally, possible actions (e.g., tuning AI/XAI methods, and feature engineering) will be taken to improve the existing AI system to achieve a better trade-off in the next iteration. We reflect on each step of the proposed process in all three use cases in Table IV at the end of our paper.

C. Metrics

1) *Model Utility*: Model utility refers to the traditionally used machine learning metrics that assess how well the trained model performs at the specific classification or regression task. For example, model accuracy, precision, recall, F1 score, and so on.

2) *Accountability*: The accountability of the AI model refers to the stakeholders' expectation that the lifecycle of AI's behaviour should comply with the relevant regulations. XAI is one of the common ways to implement accountable AI by making its decision-making process more transparent. We introduce *currentness* to describe the ratio of the time for executing different XAI methods to the time of executing AI models. The shorter the better. The second metric is called *feature importance* which refers to how important each training data feature contributes to AI model results. We also propose to use *stability*, *compactness*, and *consistency* metrics with the support of Shapash library to evaluate different XAI methods. The *stability* metrics describe how stable a certain XAI method can explain similar data instances. The *compactness* attempts to show if most of a certain XAI / AI results can be explained by the minimum possible features. Lastly, the *consistency* measures how consistently different XAI methods can explain the results given by the same AI model. Further detail can be found in Shapash library.

3) *Resilience*: We propose the following metrics to measure the difficulty faced by an attacker to succeed in an adversarial ML attack. *Impact* quantifying the effect of the attack on the system from an integrity, availability or confidentiality perspective. A high impact decreases resilience. *Complexity*

quantifying the effort required by an attacker to achieve a successful attack. A high complexity increases resilience. *Detectability* (of the attack exploiting the vulnerability) quantifying the effort required by a defender to detect and mitigate an attack exploiting the vulnerability. A high detectability increases resilience. *Capability (privileges)* required, quantifying the capabilities required by an attacker to achieve the attack. Higher required capabilities increase resilience.

We further define these metrics as follows for evasion attacks, which are mainly studied in our initial work due to their high popularity:

- **Impact**: the ratio of adversarial examples being successful at evading the target ML model to all generated adversarial examples.
- **Complexity**: the average number of queries required by the ML model to generate a successful adversarial example.
- **Detectability**: the modifications required to generate a successful adversarial sample. The distance between original X and adversarial examples X_a , is computed as an L_0 norm, which counts the number of non-zero elements in each vector. For example if $X = (0, 0)$, $X_a = (0, 5)$, the L_0 distance should be 1.

4) *Privacy*: We introduced two metrics that relate to the privacy-preserving. The first one is the ϵ of differential privacy (DP) which basically defines how difficult the attacker can derive private information. The higher the easier. The second metric introduced is called *user diversity*. The higher the user diversity, the more difficult it is for each individual from being identified.

III. USE CASE 1: ENCRYPTED NETWORK TRAFFIC CLASSIFICATION

A. Use Case Introduction

The growing prevalence of HTTPS and Virtual Private Networks (VPN) has resulted in a significant rise in encrypted Internet traffic. By 2022, approximately 95% of all Internet traffic is encrypted, with more than 85% of attacks occurring within encrypted traffic. While encryption is essential for user privacy, it poses challenges for security tools responsible for analysing and classifying traffic. This encrypted network traffic classification will become more challenging in the future networks with exponentially increased volume and diversity of traffic. AI models fit perfectly to this big data challenge for traffic classification, however, the black box AI could misbehave for unknown reasons, which decreases the trustworthiness of using AI for wider areas.

The dataset is generated by capturing local network traffic at Montimage. Specifically, we utilize Wireshark to create pcap files with a size of 2.15 GB when a user engages in normal activities on a single host. The main dataset comprises multiple network traffic traces, each associated with a specific user activity. Here, the network traffic traces contain essential information such as the source and destination IP addresses, protocols, port numbers, packet timestamps, packet

size, etc, depending on the specific features we choose. After applying filtering processes, the final dataset consists of 382 labelled traces across three traffic classes: Web, Interactive, and Video activities, with 304, 34, and 44 traces respectively. The processed CSV files derived from this dataset are used for the analysis and evaluation of our AI-based classification system. Feature extraction reveals 21 features categorized into five main categories: duration, protocol, uplink, downlink, and speed. We employ various machine learning classification algorithms, including Neural Networks (NN), LightGBM (LGBM), and XGBoost. We generated 103 adversarial samples from the 103 test data samples we initially obtained to launch a common white-box evasion attack: Fast Gradient Sign Method (FGSM) [7].

B. Trade-off Objectives

The objective of this use case is to study how resilient the existing AI solution is to a common AI attack. Moreover, we will also explore how XAI technologies can help in making this AI solution more robust. Most importantly, the increased accountability by introducing XAI can not largely compromise the model's utility.

C. Metrics Selection

We choose accuracy as the model utility metric as the dataset we use does not have an imbalance issue. We use all proposed accountability metrics as it is the core of our study in this use case. We choose impact and complexity as the resilience metric. There are no privacy-related metrics selected for this study as it is not included in our trade-off objectives.

D. Initial Iterative Results

We firstly run our experiments using NN before and after the FGSM attack, then calculate the pre-selected metrics for trade-off analysis. However, the existing Shapash library does not support stability, compacity and consistency analysis for NN. We then changed the NN model to the tree-based models LightGBM and XGBoost which are also widely used in industry yet still considered as black-box models due to their high complexity. Except for the resilience metric, complexity ($37\mu s$, averaged over 1000 iterations), which is calculated by iterating the generation of adversarial samples over 1000 times and getting the average value to generate per one sample, we summarised all other key results in Table I and showed an accountability case in Figure 2.

As shown in Table I, NN achieves the best model utility as expected. But it does not support advanced accountability analysis as mentioned before. Although XGBoost can have a relatively good accountability and model utility, it is the most vulnerable model to the evasion attack compared with NN and LGBM. To achieve a good trade-off between accountability, resilience and model utility, LGBM is recommended. We also suggest SHAP [8] with LGBM as it is not so slow compared with LIME [9].

Additionally, as shown in Figure 2, the feature importance ranking has changed significantly before and after the FGSM

attack. We have also observed a similar property when LGBM and XGBoost are used. This finding highlights a great potential to use XAI methods for detecting the possible evasion attacks that lead to the misbehaviour of the targeted AI systems.

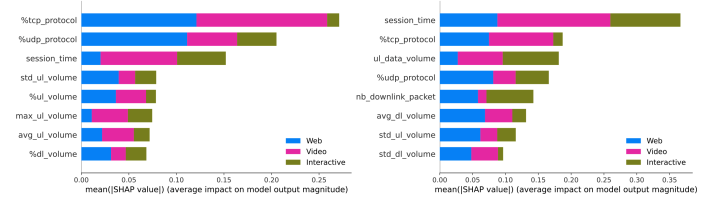


Fig. 2. List of features that are sorted by the average SHAP value (feature importance score using SHAP) for each traffic type classification (Left: before FGSM; Right: after FGSM).

IV. USE CASE 2: MALDOC DETECTION

A. Use Case Introduction

Remote collaborative working style becomes prevalent across various computing platforms (e.g., desktop, mobile, cloud, etc.), especially after COVID-19. However, this increases the risk of sharing files in a secure and private-preserving manner. Therefore, we conduct our initial experiment in a use case called *MalDoc*. It is an ML-based malicious document detection system meant to classify Microsoft Office files as one of two classes: “benign” or “malware”. It takes as input a Microsoft Office document and it outputs a binary decision together with a probability score depicting the likelihood of a certain document being malicious. The MalDoc system uses an XGBoost model for the classification of documents. This XGBoost classifier is trained using a training dataset composed of 30,600 benign and 10,900 malicious Microsoft Office files. These files are parsed and 143 features are extracted from each of them before being used for training. Most of the features are summing up counts of keywords contained in macros, which depict the activity found in the document. A testing dataset containing 7,600 benign and 2,700 malicious Microsoft Office files is used for testing, validation, and computation of performance metrics such as accuracy, precision and recall. An additional attack dataset composed of 100 malicious Microsoft Office files is used for empirical security testing and computation of resilience metrics. The evasion attacks launched in this study are: Simple blackbox Attack (SimBA) [10], Zeroth Order Optimization based black-box attacks (ZOO) [11], and Natural Evolution Strategy (NES) [12].

B. Trade-off Objectives

The MalDoc system must meet two main requirements: high accuracy and resilience against evasion attacks. Thus, the resilience/accuracy trade-off is paramount for this use-case. We have also explored how XAI method SHAP can help to achieve a better trade-off objective.

TABLE I
A SUMMARY OF INITIAL ACCOUNTABILITY AND RESILIENCE ANALYSIS FOR ENCRYPTED NETWORK TRAFFIC CLASSIFICATION.

AI Model	Accuracy w/o attack	Accuracy with attack	Impact	Currentness (SHAP/LIME)	Stability	Compacity	Consistency
Neural Network	96%	71%	29%	25976 / 8535	unknown ¹	unknown	unknown
LightGBM	94%	72%	28%	15992 / 9738	medium ²	medium	medium
XGBoost	94%	54%	45%	14678 / 6726	high	high	high

¹ The existing software library that supports stability, compacity, and consistency analysis, does not work for NN-based models.

² For simplicity and considering the page limit, we ignore the figures for the comparative accountability studies and only summarise the results. Specifically, this means XGBoost is better than LightGBM in terms of stability, compacity, and consistency.

C. Metrics Selection

In terms of the model utility, we select the following three metrics:

- Accuracy: the ratio of samples getting correct predictions from the ML model over all tested samples.
- Recall: the ratio of malicious samples getting correct predictions from the ML model over all malicious samples.
- False positive rate (FPR): the ratio of benign samples getting incorrect predictions from the ML model overall benign samples.

We use the full set of proposed resilience metrics for analysis. In particular, we conduct evasion attacks under the following three attacker's capability levels:

- High: No constraint on modifications. Feature values are optimized and kept as floating numbers. This is unrealistic since feature values represent keyword counts and are supposed to be integers.
- Medium: Modified features must be integers to respect the keyword counts they represent.
- Low: Modified features must be integers and their values can only be increased. Any modification would correspond to the insertion of a new keyword in the document. It means that it won't compromise its integrity or break its functionality. This is the most realistic capability.

D. Initial Iterative Results

First, we test the utility of the MalDoc system in normal working condition (no attack) and when subject to the evasion attacks under three different capability levels. After carefully examining the features most manipulated by the evasion attacks and the most important features for XGBoost decision-making using SHAP, we remove 14 features that are not important for the ML models but highly ranked in the list of most manipulated features by the tested evasion attack. We then retrain the XGBoost model using 129 features and run the same test again before and after launching the evasion attacks. We summarise the resilience analysis results in Table II and the corresponding model utility changes in Table III.

We can see that under high capability, all attacks have a high impact (equal or close to 1), while depicting very different complexity and detectability. SimBA is the least complex and detectable attack modifying 9 features on average, while NES is the most complex and detectable, modifying 115 features on average. As the adversarial capability decreases (from high to low), the impact of the attacks decreases down to 0.05

TABLE II
A SUMMARY OF RESILIENCE ANALYSIS FOR MALDOC DETECTION. (O: ORIGINAL 143 FEATURES; R: RETRAINED RESILIENT 129 FEATURES.)

Capability(att.)	Impact		Complexity		Detectability	
	O	R	O	R	O	R
High(ZOO)	1.00	0.99	206,044	206,868	22.5	26.7
High(SimBA)	1.00	0.87	340	498	9.0	11.8
High(NES)	0.96	0.88	567,524	550,827	115.8	104.7
Med(ZOO)	0.80	0.84	206,044	207,501	13.0	21.2
Med(SimBA)	1.00	0.89	348	509	9.2	10.8
Med(NES)	0.98	0.85	558,743	576,300	116.1	104.0
Low(ZOO)	0.05	0.02	206,044	206,044	7.4	6
Low(SimBA)	0.05	0.02	202	52	3.6	2
Low(NES)	0.05	0.01	649,097	532,444	96.6	80.0

with the lowest capability (biggest constraints). Yet, a 5% impact remains a significant vulnerability for this use case where malicious documents need to be identified with high recall. The complexity is not affected much by the decrease in capability while the detectability goes down overall. This can be explained by the decrease in impact, the most difficult adversarial example to generate will be unsuccessful (explaining the low impact), while the ones that are successful require smaller modifications and are easy to turn into adversarial examples to start with.

As for the final trade-off shown in TableIII, this resilience improvement comes with a very little cost of less than 0.7% reduction in any of the used three model utility metrics.

TABLE III
MODEL UTILITY RESULTS BY THE ORIGINAL AND RETRAINED RESILIENT MALDOC ML MODEL

Model	Accuracy	Recall	FP Rate
Original (143 features)	99.82%	99.31%	0%
Resilient (129 features)	99.71%	98.99%	0.04%

V. USE CASE 3: FEDERATED LEARNING

A. Use Case Introduction

Federated Learning (FL) is a typical technology for privacy-preserving AI-based future networks. It enables decentralized learning from distributed data across devices while maintaining data privacy. FL's distributed nature also complements edge computing capabilities, enhancing performance and reducing latency. However, although the private user data is always kept locally for the FL training, the model updates every iteration can still be used to reconstruct the user's private data. Privacy-preserving technologies such as differential

privacy (DP) can be used for reducing the probability of such reconstruction, but it could also lead to the convergence problem of the FL algorithm due to the high noise added to the model updates.

To explore the best trade-off strategy between model utility and privacy-preserving, we conduct our initial experiments using an FL framework named Flower [13] with one aggregation server and ten virtual clients and non-IID MNIST dataset (i.e., 10-class classification problem by recognising hand-written digit number from 0 to 9). To create non-IID data distribution, each client was assigned 1000 data samples where 90% of them is a class (i.e., a certain digit number) it is allocated. The remaining 10% dataset could contain any of 10 classes. As shown in Figure 3, the user's private data always stays at the client side and only send the centralised aggregator the model updates when the one training epoch is completed locally.

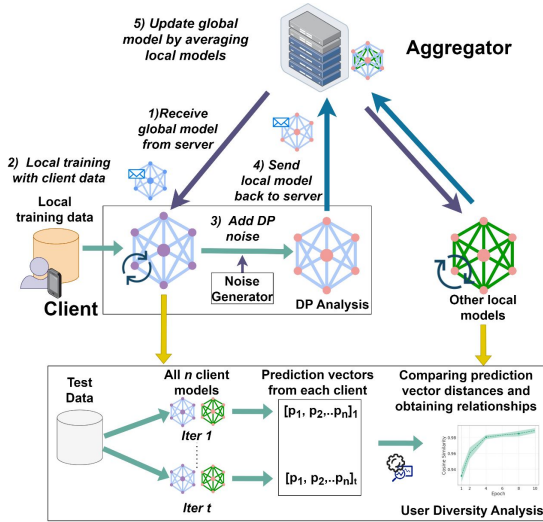


Fig. 3. This FL workflow shows how the DP noise is added for privacy-preserving at Step 3), and how user diversity metrics are obtained at each training iteration by averaging the cosine similarities among prediction vectors shown in the bottom box.

B. Trade-off Objectives

The objective of this use case is to study how much noise can be added using DP to keep a satisfactory model utility. Additionally, we study if user diversity matters for privacy-preserving FL training.

C. Metrics Selection

We use accuracy as a model utility metric. We also select the DP ϵ as one of the privacy-preserving metrics. To obtain another privacy metric, the user diversity at each training iteration, the prediction vectors from the local model on each client are kept by testing it with a sample dataset with 100 records consisting of randomly distributed 10 classes. We firstly calculate the cosine similarity of the given client prediction vector to the one from all other 9 clients. Then, the user diversity is the averaged cosine similarities of all 9 prediction vector pairs.

D. Initial Iterative Results

As shown in the following Figure 4, when increasing the level of privacy protection, it can be observed that the accuracy of the models gets reduced. This demonstrates that although it is getting increasingly difficult to reconstruct the private data using the local model updates with the stronger noise added, the training process per-se has also been severely impacted due to the faded useful information for effective learning. The utility degradation can be severe where the accuracy levels drop significantly for over 70% on the MNIST data.

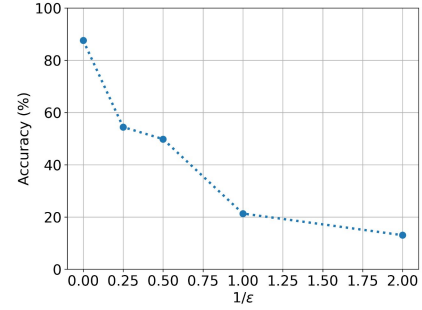


Fig. 4. The evolution of FL model accuracy with the increasing level of DP privacy protection (the increasing value of $\frac{1}{\epsilon}$ means from “no privacy” to increased privacy-protection level).

Another interesting observation can be found in Figure 5 which shows that the user diversity is reducing sharply (i.e., cosine similarity is increasing) at the very first 4 training iterations while plateaus at a low level afterwards. This observation reveals that the concerns about privacy leakage may only exist at the beginning of the training iterations. After this “warm-up” phase, the user diversity can be well preserved by the central aggregator’s model integration one iteration after another.

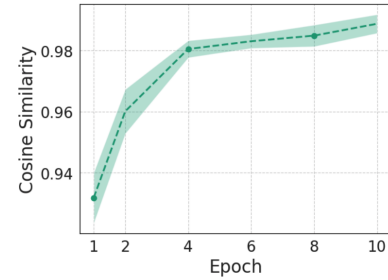


Fig. 5. The trend of averaged cosine similarity among prediction vectors from all clients, over the increasing FL training epoch.

VI. DISCUSSION AND FUTURE RESEARCH CHALLENGES

As summarised in Table IV, beyond model utility, we have done our initial trade-off analysis on three AI network applications towards model accountability, resilience, and privacy. Specifically, we choose a subset of metrics that are feasible and most relevant to the particular use case. Then we trigger the existing AI pipeline to obtain the first-round results, making our trade-off objectives quantitatively (resilience) and

TABLE IV
SUMMARY OF THREE STUDIED AI APPLICATIONS WITH THE REFLECTION TO THE PROPOSED FOUR-STEP PROCESS

	Network Traffic Classification	MalDoc Detection	Federated Learning
Dataset	Montimage 5G networks traffic	Proprietary dataset at WithSecure	MNIST
Algorithms	NN, XGBoost, LGBM; LIME, SHAP	XGBoost; SHAP	FedAvg
1. Trade-off Objectives	Accountability, Resilience (FGSM)	Resilience (ZOO, SIMBA, NES)	Privacy Preserving
2. Metrics Selection	Accuracy, Impact, Complexity, Currentness, Stability, Consistency, Compacity	Accuracy, Recall, FPR, Capability, Impact, Complexity, Detectability	Accuracy, ϵ of DP, User Diversity
3. Existing AI Pipeline	NN achieves the highest model utility but does not have good accountability; XGBoost has good accountability but is vulnerable to FGSM attack.	There exist some features that are mostly manipulated by the evasion attacks but not important for model inference	Model utility can drop significantly with gradually decreased ϵ .
4. Actions to Improve	Better trade-off can be achieved by using LightGBM with SHAP.	Better trade-off can be achieved by removing those features to retrain the XGBoost model	Better trade-off can be achieved by 1. Fine-tuning ϵ until the lowest acceptable model utility; 2. the impact of user diversity on privacy is reduced when FL training round is increasing.

qualitatively (accountability) visible. Finally, we provide our suggested improvement actions with initial results. Having demonstrated promising results in using XAI for more robust AI cybersecurity solutions, we have also listed the following three research challenges.

- **XAI for time-series scenarios.** In the future networks, there will be various devices and sensors constantly collecting data for the downstream ML tasks in real-time. Devising better XAI approaches for multi-variant time-series analysis can be critical to expanding the applications of XAI to future full-stack AI networks. Recent research on explainable natural language processing using sequence models may also inspire this direction.
- **Metrics calculation for scalability.** Next generation networks will provide much higher bandwidth and reliability with much lower latency. Therefore, it is essential to test if the calculation of proposed metrics can be scalable. Similar to TreeSHAP which accelerates the computation time for tree-based models, some time-consuming metrics may also need to be approximated in the distributed large-scale user request environments.
- **Engaging external stakeholders.** External stakeholders such as end-users and legal auditors are important for an accountable AI-assist networks. More interdisciplinary research is required to engage external stakeholders in the existing closed-loop process. Methodologies like the System Usability Scale (SUS) based questionnaire might be a good start to collect their feedback.

VII. CONCLUSION

This paper proposes an iterative process for internal network stakeholders to help their “black-box” AI solutions achieving a better trade-off among model performance, accountability, and resilience. This process are validated on three AI-based network applications: encrypted network traffic classification, malware detection, and federated learning. This paper also presents research challenges for deploying XAI for future intelligent networks.

ACKNOWLEDGMENT

This work is partly supported by the European Union under the SPATIAL project (Grant ID. 101021808) and by Science Foundation Ireland under CONNECT phase 2 (Grant no. 13/RC/2077_P2) projects.

REFERENCES

- [1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, “The roadmap to 6g: Ai empowered wireless networks,” *IEEE communications magazine*, vol. 57, no. 8, pp. 84–90, 2019.
- [2] S. Wang, M. Atif Qureshi, L. Miralles-Pechuan, T. Reddy Gadekallu, M. Liyanage *et al.*, “Explainable ai for b5g/6g: technical aspects, use cases, and research challenges,” *arXiv e-prints*, pp. arXiv:2112, 2021.
- [3] T. Senevirathna, V. H. La, S. Marchal, B. Siniarski, M. Liyanage, and S. Wang, “A survey on xai for beyond 5g security: technical aspects, use cases, challenges and research directions,” *arXiv preprint arXiv:2204.12822*, 2022.
- [4] P. Barnard, N. Marchetti, and L. A. DaSilva, “Robust network intrusion detection through explainable artificial intelligence (xai),” *IEEE Networking Letters*, vol. 4, no. 3, pp. 167–171, 2022.
- [5] F. Rezazadeh, H. Chergui, and J. Mangues-Bafalluy, “Explanation-guided deep reinforcement learning for trustworthy 6g ran slicing,” *arXiv preprint arXiv:2303.15000*, 2023.
- [6] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. Moura, and P. Eckersley, “Explainable machine learning in deployment,” in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 648–657.
- [7] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- [8] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.
- [9] M. T. Ribeiro, S. Singh, and C. 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.
- [10] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, “Simple black-box adversarial attacks,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 2484–2493.
- [11] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, “Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models,” in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.
- [12] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, “Black-box adversarial attacks with limited queries and information,” in *International conference on machine learning*. PMLR, 2018, pp. 2137–2146.
- [13] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão *et al.*, “Flower: A friendly federated learning research framework,” *arXiv preprint arXiv:2007.14390*, 2020.