# From Opacity to Clarity: Leveraging XAI for Robust Network Traffic Classification

Chamara Sandeepa[1], Thulitha Senevirathna[1], Bartlomiej Siniarski[1], Manh-Dung Nguyen[2], Vinh-Hoa LA[2], Shen Wang[1], and Madhusanka Liyanage[1]

[1] School of Computer Science, University College Dublin, Ireland
{abeysinghe.sandeepa, thulitha.senevirathna}@ucdconnect.ie,
{bartlomiej.siniarski, shen.wang, madhusanka}@ucd.ie
[2] Montimage EURL, France
{manhdung.nguyen,vinh_hoa.la}@montimage.com

**Abstract.** A wide adoption of Artificial Intelligence (AI) can be observed in recent years over networking to provide zero-touch, full autonomy of services towards the next generation Beyond 5G (B5G)/6G. However, AI-driven attacks on these services are a major concern in reaching the full potential of this future vision. Identifying how resilient the AI models are against attacks is an important aspect that should be carefully evaluated before adopting these services that could impact the privacy and security of billions of people. Therefore, we intend to evaluate resilience on Machine Learning (ML)-based use case of network traffic classification and attacks on it during model training and testing stages. For this, we use multiple resilience metrics. Furthermore, we investigate a novel approach using Explainable AI (XAI) to detect network classification-related attacks. Our experiments indicate that attacks can clearly affect the model integrity, which is measurable with the metrics and detectable with XAI.

**Keywords:** AI Resilience · Explainable AI · AI Attacks · Activity Classification · Communication Networks

## 1 Introduction

Real-world use of AI can have various caveats. Once deployed in the cloud, the models will constantly undergo different types of intrusion attempts. Although pragmatically, these attempts are unlikely to pass through the API gateway; it is not a certainty that it will be the case always. On the off chance that one of these attacks becomes successful, the model can be either stolen, taken under the control of the attacker, or completely/partially disabled. Therefore, achieving model robustness requires constant analysis of the model for various attacks.

However, despite the efforts to implement security and privacy, ML can still undergo numerous attacks that can compromise trust in the final outcome of the model. These attacks include evasion, poisoning, backdoor, and inference attacks [6]. This can create an issue in applying AI to real-world applications such as network automation. If an AI model performing critical network operations is compromised, the user's data that is handled via the AI is at risk, and the end users can lose trust in the network [13]. Yet, AI-driven automation is expected to be the future, which requires fully automated solutions to handle complex networks with trillions of devices in beyond 5G.

Hence, analysing a model's resilience against attacks is essential in AI before integrating it into a real-world application in future networks. A major component in resilience analysis is the metrics used to analyse the models. These metrics can uncover information regarding the model and its effectiveness against attacks. However, these metrics need to be customized during the implementation process for each attack type. By identifying the vulnerabilities and the model's effectiveness against attacks, stakeholders can make important decisions regarding the model and the system's health. To further verify and obtain more detailed information, the stakeholders can be provided with explanations generated with XAI methods. XAI is a set of mechanisms that can reveal the information on the black-box AI models, explaining how the AI decision-making process is done [15]. In this paper, we investigate on the possibility of using XAI as a detection tool for resilience attacks against AI by observing changes in post-hoc explanations with the attacks.

As a practical scenario of using AI in networking, we use a target use case of network activity classification, which can be regarded as an essential component in future network automation processes. It can differentiate various types of data transmitted over the network, which can be used for tasks such as scaling and prioritising highly demanding traffic, identifying malicious traffic, and personalisation services. To analyse the resilience of the network activity classifier, we launch two attacks: poisoning and evasion attacks. Then, we analyse the metrics and use XAI to determine if it is possible to identify attack scenarios.

### 1.1 Our Contributions

Despite many techniques to implement ML-based systems on networks, we observe a lack of metrics for resilience evaluation in the related literature. Moreover, the analysis of the impact of attacks on AI models via XAI is a novel area that can be considered as a potential path for attack detection. We bring the discussion of these topics, which are essential aspects to be considered for future network-based AI. Therefore, we summarise **our contributions** as follows:

- Presenting a use case of network activity classification model and launching network attacks against the resilience of the model.
- Assessing two attacks, evasion, and poisoning of ML models via multiple metrics, including measuring the impact on the models and complexity of the attacks.
- Performing a comprehensive XAI Shapley Additive Explanations (SHAP)-based attack analysis on the two attacks under different conditions.

We also provide an open-source repository[3] that consists of the implementation of this work, along with the dataset sources.

The rest of the sections are arranged as follows. Section 2 discusses associated works in the AI-based network classification use case and possible attacks. Details of our system model are presented in Section 3. Section 4 provides methods for implementing AI attacks on the system model. Section 5 provides experiments and their results. We discuss possible research directions in Section 6. The paper is summarised and concluded in Section 7.

## 2 Related Works

AI-based network traffic analysis is an important aspect that network systems should maintain for maintaining a secure and efficient network environment. Several works are available in the literature that has implemented AI-based systems to analyse network traffic. Authors in [7] present a detection technique for slowDoS attacks on encrypted traffic using clustering-based AI. Work in [1] uses Neural Networks (NN) and Principal Component Analysis (PCA) to analyse and classify network traffic for potential identification of malware traffic. In [2], the authors develop a Recurrent Neural Network (RNN)-based heterogeneous traffic detection and classification system for 5G networks.

However, detecting vulnerabilities in such AI applications that classify the traffic is a significant concern since the models trained on AI also need to train on accurate data in a secure and private environment. For example, in decentralised systems such as distributed AI or Federated Learning (FL) [12], AI models are trained by aggregation of models trained by many clients. In such a case, providing guaranteed privacy and security for each client would not be feasible. Even centralised systems can be vulnerable to attacks, where attackers can exploit a vulnerability and alter the operations in the training of AI models.

Attacks such as data poisoning [3] can cause degradation in the performance of ML models. Here, an attacker attempts to cause damage to the model predictions by manipulating the input data used to train the model. Several types of poisoning attacks exist, which include: 1) untargeted poisoning, which hinders convergence of the target model, leading to denial-of-service, 2) targeted poisoning, which causes abnormal predictions for some inputs; and 3) backdoor attacks, that uses targeted poisoning attacks go unnoticed via techniques such as hidden triggers embedded in the training data [14]. In network traffic classification scenarios, an attacker may use this type of attack to evade attacker-specific malicious traffic by poisoning the model to be insensitive to malicious traffic.

Evasion attacks are another type of attack where the attacker aims to evade the decisions made by the learned model during the test time; however, unlike poisoning attacks, they do not interfere with the training data [4]. Instead, they add perturbations to the inference data, which can cause erroneous output from the trained model. If an evasion attack is made for tasks such as network traffic classification scenarios, an attacker can successfully evade any malicious traffic

[3] https://github.com/Montimage/activity-classification

by having a small perturbation, misleading the classification of the AI model, even if the model is trained accurately and safely in the training phase.

Therefore, by considering these issues, assessing the impact of these attacks on AI is a significant requirement to identify the level of resilience of the overall intelligent system against the attacks. Furthermore, having XAI-based inspection on the model outputs can provide more explainability on model behaviour. Especially with attacks, if the model behaviour is changed with attacks, XAI itself can be used as a potential detection mechanism against the attacks, even if the changes are subtle and not visible, merely observing model predictions. Thus, we have two benefits of using XAI in the use case: 1) enhancing model transparency via explanations and 2) its potential use to evaluate resilience against poisoning and evasion attacks. However, we observe a lack of related literature considering both these aspects of XAI. Hence, in our work, we highlight the novel concept of using XAI as a potential tool for network classification scenarios for the detection of adversarial attacks.

## 3 System Model

The network activity classification is based on users interacting with a user's end personal devices. An overview of the system model is presented in Fig. 1.

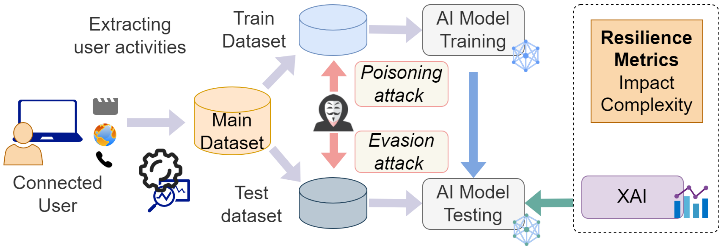


Fig. 1: Overview of the system with XAI and metrics implementation. The attacker is performing attacks; meanwhile, the resilience metrics and XAI analyse the impact, complexity, and possible detection of the attacks.

Here, network packets are analysed by a packet-capturing service. The metadata in the network packets includes the type of data that the payload of the packet consists of. This information is preprocessed and collected into a main dataset. Next, the data is split to train and test sets, where an AI model is first trained with the train data. The objective of this model is to classify the type of data that the payload consists of for a given metadata. Next, the trained model is tested with the test set to identify the performance of the model.

However, an attacker can interfere with the model training or the testing phase. They can launch poisoning attacks on the training data or target the test data to launch evasion attacks. Therefore, the model is vulnerable in both phases of the ML process.

Therefore, resilience metrics can be obtained to evaluate how the model is resilient against the attacks on AI. For this, we use two types of metrics: impact and complexity [11]. The impact metric measures how critical the attack is by analysing the changes in the accuracy metrics. The complexity is the level of difficulty of launching it from the attacker's side. Since we simulate the attacks, we can analyse the trade-off between the impact and complexity to get overall attack performance. Furthermore, we get the XAI metrics to identify internal changes in the ML model with the attacks. Such changes can lead to a potential attack detection technique, where we investigate from experiments in Section 5.

# 4 Attacks on AI

This section presents details of the two attacks we implement on the network traffic classification system. We use two attacks: 1) poisoning attacks during the training phase and 2) evasion attacks in the testing or inference phase. We highlight the procedure of launching the attacks, their variations, and constraints in the implementation.

## 4.1 Attacks at Training Time: Poisoning Attacks

The resilience of an explanatory platform can be evaluated by analysing how the platform would behave during an attack event. The observations would be the deviations in the model parameters when compared with the models before the attack and the output explanations provided by the trained models after the attack. The level of impact that an attack can be quantified with suitable metrics for the specific attack type. To simulate such a scenario, a poisoning attack was implemented on a network activity classification system, where an attacker is expected to perform different types of poisoning to the original dataset used to train the classification model. The overview of the attack is shown in Fig. 2.

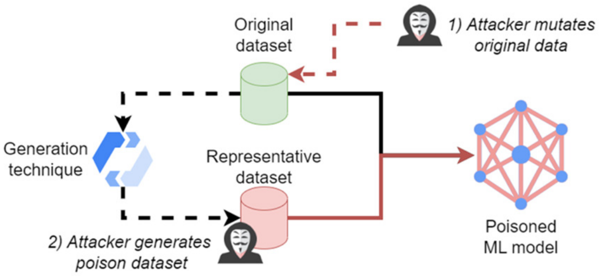


Fig. 2: Poisoning attacks on the activity classification model.

Here, the attacker can perform the data poisoning attack in two approaches: 1) modifying/mutating the original dataset or 2) adding new adversarial dataset examples to the existing dataset. For the second type, the attacker can use a generation technique like Generative Adversarial Networks (GAN) [16] for the attack. For the use case, we perform three types of poisoning attacks:

- **Random label flipping attack -** attacker randomly changes the labels in the original dataset. The objective of this attack is to reduce the utility of the model.
- **Targeted label poisoning attack -** a selected label type is favoured in flipping the labels by the attacker. The attacker intends to create an artificially biased dataset with the selected label to change the decision boundary of the model.
- **GAN-based poisoning attack -** here, the attacker uses a GAN to generate new data instances. Depending on the GAN-generated data's similarity level, this attack will be more challenging to detect. The main intention of the attacker here is to identify how the model's internals behave without disrupting its performance. Such poisoning can be used for privacy attacks like membership inference [17].

**Metrics to Assess Poisoning Attack** For assessing how the model has changed with the attack, we use two metrics: impact and complexity.

The impact of the poisoning on the ML model can be defined as:

$$Impact = \frac{Error(F_p) - Error(F)}{Error(F)} \tag{1}$$

where it measures the original accuracy of a benign ML model $F$ compared to compromised model $F_p$. The impact is high when the difference between the benign and the compromised model is high.

The complexity metric is defined as:

$$Complexity = \frac{|D_p|}{|D + D_p|} \tag{2}$$

which computes the ratio of poisoned data $D_p$ and benign data $D$. It is regarded in poisoning that higher poisoning ratio results in higher complexity to perform the attack.

### 4.2 Attacks at Test Phase: Evasion Attacks

One of the popular attack types in adversarial AI is the evasion attack. Here the attacker attempts to misclassify the model output for maleficent advantages. In our implementation, we adopt FGSM (Fast Gradient Sign Method) [8] white box attack technique to generate adversarial data points used to test the web activity classification model. Here the intuition is to add a non-random noise with the direction same as the gradient of the cost function with respect to the data. i.e., according to the authors of this attack, the direction of the perturbation only needs to have a positive dot product with the gradient of the cost function while the absolute magnitude of the perturbation is just enough to skip over the decision boundary. This evasion method can generate sequences of adversarial data points that can be generalized for ML models trained with different sub sets of training data. For the purpose of demonstrating the capability of metrics and XAI, FGSM attack is an ideal candidate because of the said extensibility with other ML models which we intend to expand the results.

**Metrics to Assess Evasion Attacks** We use two resilience metrics, impact and complexity, that would manifest the attack's effectiveness against the models. The impact of evasion attacks is the attack success rate for a successfully evaded percentage of adversarial samples.

$$Impact = \frac{1}{|A|} \sum_{X_a \in A} \begin{cases} 1, & \text{if } F(X_a) \neq F(X) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The complexity of the model can be defined as the CPU usage for the generation of evasion adversarial samples.

$$Complexity = t_{cpu} \times N_{adv} \tag{4}$$

where $t_{cpu}$ denotes the total CPU time and $N_{adv}$ is the number of adversarial samples generated.

# 5 Experiments

This section is organised under each attack type we ran our experiments on. In addition to that, we also generate explanations through the XAI technique SHAP to provide a granular insight into the model features and their importance distribution.

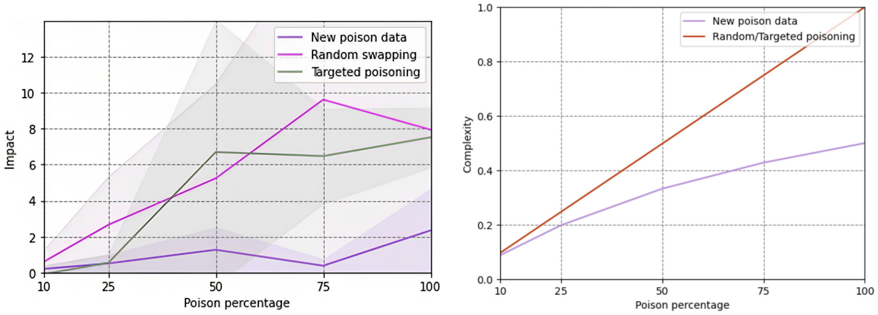
## 5.1 Evaluating Poisoning Attacks

To perform the experiments, first, we trained a NN model for classifying the network testbed dataset. The data was collected from 279 training and 103 test instances of converted network PCAP files, with 22 features for each. We use a sequential model for the NN with 3 layers with 12, 8, and 3 dense layers, respectively. The configurations of the NN include Adam optimiser, batch size of 10, and binary cross-entropy for loss calculation. After training the model for 150 iterations, NN model test accuracy without attacks is obtained as 96.12%.

**Impact and Complexity Analysis** Next, we perform poisoning attacks and assess the impact and complexity of the ML model over different poisoning types. As discussed, we have three types of poisoning: random swapping and poisoning by targeted flipping of labels were performed on the existing training data. The models are trained over varying poison percentages of 10%, 25%, 50%, and 100% of the original data being poisoned. We also generated a GAN-dataset using a small representative dataset from the original dataset, using a library called CTGAN[4]. Then, we run GAN-based poisoning by mixing GAN data together with the original data. Next, the impact and complexity were calculated for each poisoning percentage. Graphs in Fig. 3 show how impact and complexity metrics vary over different poisoning levels.

Here, the impact of a new GAN-generated poisoning attack is relatively lower than the random and targeted poisoning since the new examples added by the

[4] https://github.com/sdv-dev/CTGAN



(a) Impact vs. poison percentage. (b) Complexity vs. poison percentage.

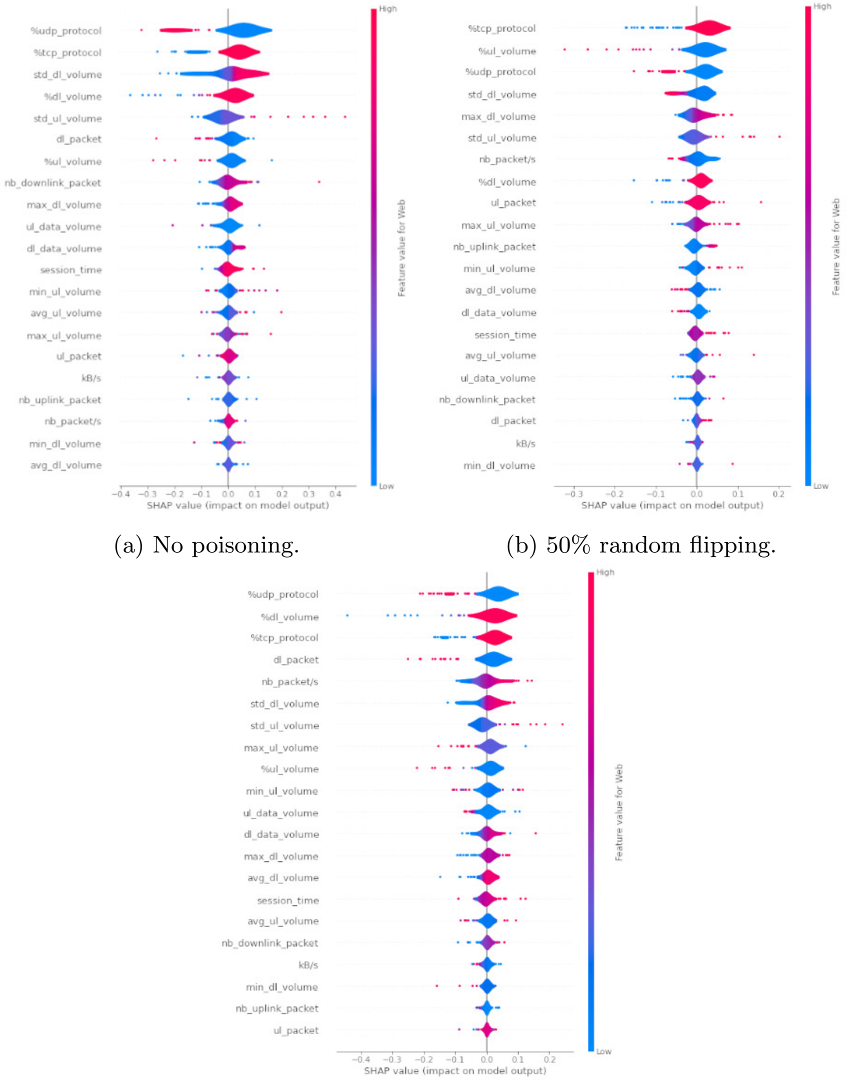
Fig. 3: Impact and complexity metrics over different poison percentages.

GAN-generated dataset are very similar to the original data examples. In GAN-based poisoning, the attacker's main intention is not to reduce the model utility, since most GAN-based attacks are aimed for revealing privacy leakages without disrupting the model utility [17]. Therefore, the differences in original accuracy vs. the accuracy of the poisoned model here are lower. This results in lower impact. GAN-based attack is, therefore, more difficult to detect. However, for the other two attacks, the poisoned mode has significantly deviated from the original model, causing a higher impact. Similarly, the complexity metrics are higher for the random and targeted poisoning types than for the new GAN-generated poisoning.

**XAI Feature Changes with Poisoning** To compare how key explanations from XAI are different with the poisoning attacks, SHAP [9] values are taken for the model for three scenarios of a) no poisoning, b) random label flipping, and c) new GAN-based poisoning as follows:

The SHAP values of the model without poisoning are similar to the GAN-based poisoning, which can be expected since they have a lesser impact on the model. However, in random flipping, the top feature values of the model have significant differences when compared with the model without poisoning, resulting from the high impact of the random flipping attack.

Therefore, in summary, the resilience of the model has been impacted by the poisoning attacks. They can be measured via suitable metrics for the attack scenario. For example, metrics such as distances among clustering of XAI feature values can be used to detect how impactful the attack is. Such XAI metrics can also provide a possibility of identifying a comparable deviation from the normal model when an attack event occurs. When considering the novelty of our approach by using XAI, poisoning attacks can initially be detected by observing the deviation of the most significant features compared with a non-poisoned model. This can further be analysed with the use of metrics that can quantify the impact of the attacks via XAI, which is considered a future work in our proposal.



(a) No poisoning.

(b) 50% random flipping.

(c) 50% GAN poisoning.

Fig. 4: SHAP values variation over no poisoning vs. different poison types.

### 5.2 Evaluating Evasion Attacks

The same dataset dimensions as in the poison attack (with 279 training data samples and 103 test samples) were also used in the evasion attack evaluation. A sequential keras model with 12 input nodes, 1 hidden layer with 8 nodes, and, a final output layer with 3 nodes is fitted with the training dataset. The rest of the hyper-parameters of the model were kept the same as the poisoning attack scenario. This neural network was then used to generate adversarial attack samples using the FGSM implementation in [10]. We generated 103 adversarial samples from the 103 test data samples we initially obtained. Exploiting the transferability of adversarial attacks [5], we expand our work on two more model types that are popularly used in network function predictions, namely, LightGBM(LGBM) and XGBoost. Here we used the same adversarial dataset generated using the neural network to launch attacks on the boosted tree models.

### 5.3 Observations and analysis

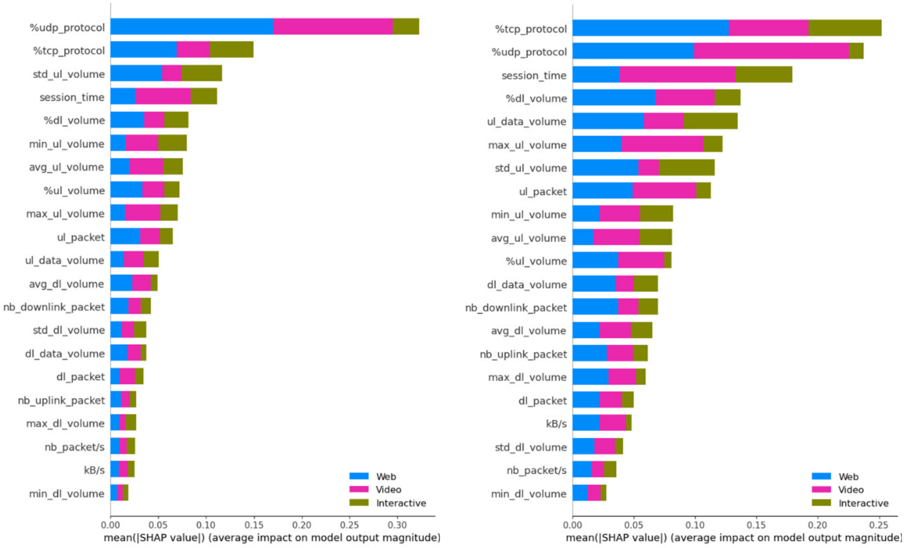
We observed the following results in Table 1 for each metric before and after each model was exposed to the evasion attack.

Table 1: Attack performance for different attack types in multiple models.

| Model | Attack type | Accuracy w/o attack | Accuracy with attack | Impact | Complexity |
|---|---|---|---|---|---|
| Neural Network | FSGM | 96% | 71% | 29% | 37.86$\mu s$<br>(1000 itr) |
| LGBM | Transferred from NN | 94% | 72% | 28% | 37.86$\mu s$<br>(1000 itr) |
| XGBoost | Transferred from NN | 94% | 54% | 45% | 37.86$\mu s$<br>(1000 itr) |

Since the generation of adversarial samples was carried out with the neural networks, the model complexity is consistent with all the models. With the complexity metric, a stakeholder can realize the convenience for an adversary to generate these samples in various devices and compare different attack types on their expected frequency. This will allow system operators to efficiently manage defense mechanisms adhering to resource constraints.

We observed a significant drop in accuracy when evasion samples were involved. It is further manifested in the impact metric presenting how vulnerable the model is to FGSM attacks. For instance, the XGBoost model is 17% more likely to be affected by an adversarial model than a LightGBM model when compared to a LightGBM model with the same level of accuracy. This is valuable information for a stakeholder when deploying a model for real-world usage. To further strengthen our analysis, we also generate explanations that would be at the disposal of stakeholders, as shown in Fig. 5. It is apparent that the model's priorities, in terms of features, change in the face of adversarial attacks. The Shapley values for web activities have decreased (around 16%) for the *udp_protocol* causing the feature to drop to the second place in ranking while the importance of the *tcp_protocol* has almost doubled. Such a significant feature change would provide more meticulous information for stakeholders to take immediate action. Towards identifying evasion attacks on models, shapley value changes are also suitable as a validator alongside resilience metrics. However, with these revelations, in the future, it is possible to develop automated solutions with increased accuracy without losing accountability for detecting evasion attacks with high confidence.



(a) Benign model (NN).

(b) Adversarial attack (NN).

Fig. 5: SHAP values for benign model vs. evasion attack event.

## 6 Discussion

The analysis of XAI in our work identifies that we can observe significant differences in a NN model undergoing an adversarial attack compared to a benign model. In evasion attacks, if the attacker starts flooding the model with evasion data points, the effects of the output will be reflected in the feature importance scores. This also depends on the frequency of the attackers' intrusions. In short, during the pre-deployment stage of a model, system developers can use the impact and complexity metrics to identify the vulnerability of the model against known evasion attacks, while XAI is more of a detection and reconciliation tool. In the case of poisoning attacks, the attack may go unrecognised in case if the attack is GAN-based. This also depends on the similarity of the produced GAN data with the original data of users. Obtaining such a representative sample from the original may also be difficult for an attacker unless they gain access to the original data or be an insider. Therefore, a GAN-produced poisoning attack may also be possible to detect and defend if we provide more security and privacy for the original data. Furthermore, the diversity and distribution of the original data can also affect the original reference model XAI outputs. For example, in the case of distributed ML like FL, many clients can have local models based on their non-Independent and Identically Distributed (non-IID) data that is specialised on a particular client. In such a case, XAI's outputs can vary from client to client. Detecting benign models from adversarial models via variations in XAI is a possibility to expand in the future.

Furthermore, when comparing the key outcomes of our work with the related works, we observe that none of the works related to network-based classifications discuss the potential of using XAI as a detection mechanism against attacks. A comparison of model resilience metrics of impact and complexity during an attack incident and their trade-offs is another key finding we delivered in our work that is not found in other works. Table 2 provides a summary comparison of our work with key highlights.

Table 2: Summary contribution of our work.

| Characteristics | Ref [1] | Ref [3] | Ref [7] | Ref [16] | Ref [17] | **Our work** |
|---|---|---|---|---|---|---|
| Network related use case for B5G/6G AI and attack scenarios | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Launching resilience attacks against trained ML models | ✓ | ✓ | - | ✓ | - | ✓ |
| Proposal of metrics to quantify resilience on the model during attacks | ✓ | ✓ | - | ✓ | - | ✓ |
| XAI-based visualisation and detection strategy | - | - | - | - | - | ✓ |
| Possible directions on enhancing resilience and limitations | - | - | ✓ | ✓ | ✓ | ✓ |

## 7 Conclusion

In this research, we analysed resilience metrics of impact and complexity of two adversarial attacks, poisoning and evasion attacks, against network activity classification scenarios. We observe that the impact and complexity can vary in different attack categories, which makes attacks such as GAN-based poisoning relatively difficult to detect, meanwhile having less complexity on the attacker's side. Thus, attacks that cause a lesser impact on the model can be more threatening to the security and privacy of the model due to its difficulty in tracing the attack. The XAI analysis with SHAP also shows that the higher the attack's impact on the model, the easier it is to determine by observing the variation of the SHAP-based feature contributions. We observe higher variation over the adversarial attacks, which cause a higher impact on the model. Therefore, XAI can be applied as a potential detection mechanism for attack detection. It can easily be adapted for assessing the model's possible resilience compromises before deploying the AI models in real-world network applications.

## Acknowledgment

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

## References

1. Arivudainambi, D., KA, V.K., Visu, P., et al.: Malware Traffic Classification using Principal Component Analysis and Artificial Neural Network for Extreme Surveillance. Computer Communications **147**, 50–57 (2019)
2. Artem, V., Ateya, A.A., Muthanna, A., Koucheryavy, A.: Novel AI-based Scheme for Traffic Detection and Recognition in 5G based Networks. In: Internet of Things, Smart Spaces, and Next Generation Networks and Systems: 19th International Conference, NEW2AN 2019, and 12th Conference, ruSMART 2019, St. Petersburg, Russia, August 26–28, 2019, Proceedings 19. pp. 243–255. Springer (2019)
3. Aryal, K., Gupta, M., Abdelsalam, M.: Analysis of Label-Flip Poisoning Attack on Machine Learning Based Malware Detector. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 4236–4245. IEEE (2022)
4. Deldjoo, Y., Noia, T.D., Merra, F.A.: A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial Networks. ACM Computing Surveys (CSUR) **54**(2), 1–38 (2021)
5. Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., Roli, F.: Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In: 28th USENIX security symposium (USENIX security 19). pp. 321–338 (2019)
6. Eigner, O., Eresheim, S., Kieseberg, P., Klausner, L.D., Pirker, M., Priebe, T., Tjoa, S., Marulli, F., Mercaldo, F.: Towards Resilient Artificial Intelligence: Survey and Research Issues. In: 2021 IEEE International Conference on Cyber Security and Resilience (CSR). pp. 536–542. IEEE (2021)
7. Garcia, N., Alcaniz, T., González-Vidal, A., Bernabe, J.B., Rivera, D., Skarmeta, A.: Distributed Real-time SlowDoS Attacks Detection over Encrypted Traffic using Artificial Intelligence. Journal of Network and Computer Applications (2021)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
9. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. Advances in neural information processing systems (2017)
10. Nicolae, M.I., Sinn, M., Tran, M.N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I., Edwards, B.: Adversarial robustness toolbox v1.2.0. CoRR **1807.01069** (2018), https://arxiv.org/pdf/1807.01069
11. Park, S., et al.: Deliverable 2.2 Define Parameters and Elements to Construct Accountability, Resilience, and Privacy Metrics. European Union, Horizon 2020 SPATIAL (2023)
12. Pei, J., Zhong, K., Jan, M.A., Li, J.: Personalized Federated Learning Framework for Network Traffic Anomaly Detection. Computer Networks **209**, 108906 (2022)
13. Sandeepa, C., Siniarski, B., Kourtellis, N., Wang, S., Liyanage, M.: A Survey on Privacy for B5G/6G: New Privacy Challenges, and Research Directions. Journal of Industrial Information Integration p. 100405 (2022)

14. Tian, Z., Cui, L., Liang, J., Yu, S.: A Comprehensive Survey on Poisoning Attacks and Countermeasures in Machine Learning. ACM Computing Surveys (2022)
15. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In: Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8. pp. 563–574. Springer (2019)
16. Zhang, J., Chen, J., Wu, D., Chen, B., Yu, S.: Poisoning Attack in Federated Learning using Generative Adversarial Nets. In: 2019 18th IEEE International Conference on Trust, Security and Privacy In Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE). pp. 374–380. IEEE (2019)
17. Zhang, J., Zhang, J., Chen, J., Yu, S.: GAN Enhanced Membership Inference: A Passive Local Attack in Federated Learning. In: ICC 2020-2020 IEEE International Conference on Communications (ICC). pp. 1–6. IEEE (2020)