Mitigating opaque trigger alerts of black-box ML models with Explainable Artificial Intelligence (XAI) in Zero Trust Architectures (ZTA)

Madhuri Rao[1], Ganesh Khekare[2]

[1] Lincoln University College, Malaysia; [2] Vellore Institute of Technology, Vellore, India

pdf.madhurirao@lincoln.edu.my

Abstract: Zero Trust Architecture (ZTA) is based on the concept of never trusting and always verifying. Even with continuous verification and strong authentication with access control mechanisms, data breaches are yet possible. Hence Intrusion Detection Systems (IDS) are essential in ZTA. Modern IDS incorporated with Machine Learning (ML) models can detect anomalies and can also identify novel attack patterns. They are critical for cloud native environments and or application with microservice architecture. ML based IDS systems can trigger alerts and can adapt to evolving threats. Complex ML models such as XGBoost and Voting ensemble methods are even more challenging to comprehend and cause operator distrust. Such black box models could also trigger alerts without any understandable cause. Here, opaque and false trigger alerts are addressed with the help of Explainable Artificial Intelligence (XAI) SHapley Additive exPlanations (SHAP) model.

Keywords: Zero Trust Architecture; Intrusion Detection System; Explainable Artificial Intelligence; SHAP; XGBoost, Voting Ensemble, Random Forest

Introduction

Zero trust is preferred over implicit trust as enterprises adapt to cloud systems for scaling and cost cutting. This requires continuous check on credibility for grating access or supporting continued access. Zero Trust Architecture (ZTA) requires evaluating entities such as users, devices and applications continuously for disparities related to behavioral, contextual and security attributes [1]. In [2] ZTA is devised for Tactical Edge Networks (TEN) which are possible because of cloud architectures, however understanding the system is a challenge, Today, we cannot imagine how things can work without AI. AI is needed in decision making primarily in health, e-commerce and disaster management systems. These systems cannot afford to be compromised as trust and faith in the system adversely affects its use [3]. Here in this aspects, Explainable AI can be very instrumental. **Figure 1** depicts the components of Zero Trust Architecture, where Trust evaluation is the core component. The other supporting components are dynamic policy engine and a point for policy enforcement. Data As A Service (DAAS) Management ensures secure data delivery, access logging and continuous monitoring.

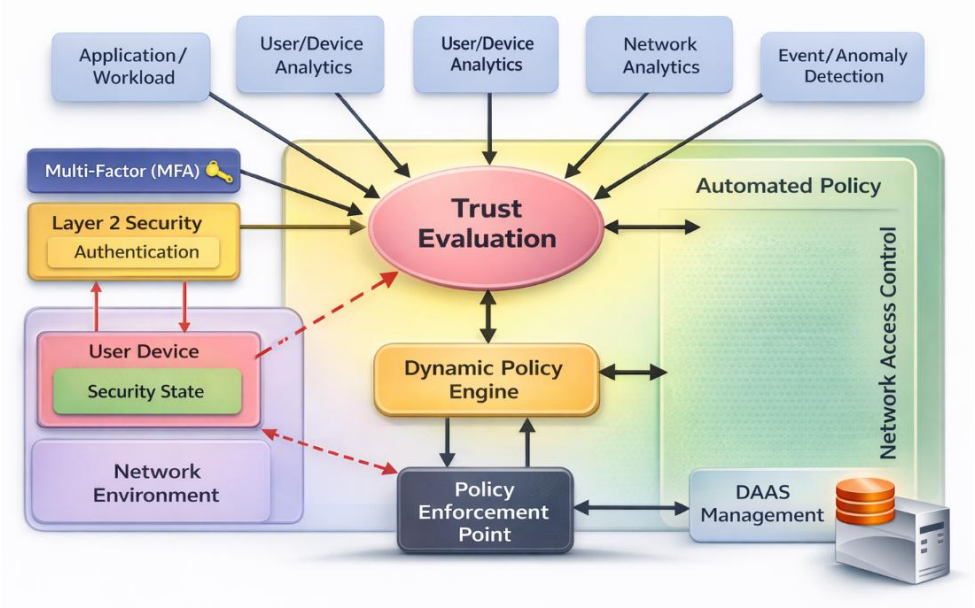

Fig1: Components of Zero Trust Architecture [1]

Figure 2 depicts the overview of XAI Models. Human in the Loop is very crucial aspect of a XAI based system that could help interpret ML Models with local and global explanations.

Figure 2: Overview of XAI

Feature Attribution Engine quantifies impact magnitude and supports trust scoring. SHAP is a strong feature attribution method that has been explored in this study.

Related work

Table 1 compares the work presented in this article with previous work done by researchers in the field of ZTA and XAI. While ZTA is crucial in Cyber Forensics, Legal Systems, Tactical Edge Networks and Cloud Native Environments, its comprehension is tedious and cumbersome. XAI based interpretations of ZTA mechanisms can build user trust in the systems.

Table 1. Compares this work with the related work or previous research by other researchers

	Applications	Models	Findings
[3]	Cyber Forensics (CF)	LIME, Anchors	Survey of various XAI models in CF
[4]	Software Defined Networks in tactical Edge Networks	Preferred Reporting Items for Systematic Reviews and Meta- analyses (PRISMA)	ZTA Systems are difficult to comprehend while essential with evolving threat landscape.
[1]	Legacy Systems	NIST Special Publication SP800-207	Discusses challenges, steps and points to consider when moving from legacy Systems to ZTA systems.
This work	Cloud Native Environments	SHAP	Understanding Alarms triggered in Intrusion Detection Systems with Ensemble Voting, Random Forest and XGBoost

Key Contribution

In this work SHAP XAI model is explored to interpret the cause of alerts that are triggered in a ML Based Intrusion Detection System. Three models are explored here including XGBoost, Random Forest and Voting Ensemble method. Four major attacks were studied in this work namely, DoS, Probe, U2R and R2L. All models performed exceptionally well on the test set with RF method having 99.97% accuracy, XGBoost having 99.99% accuracy and Voting Ensemble having achieved 99.98% accuracy and thus intrusion was successfully detected. SHAP based XAI explanations revealed and explained the top features namely count, protocol type, src_bytes, srv_count and dst_bytes contributing towards the intrusion that was successfully detected.

Method, Experiments and Results

KDD Cup dataset, widely used benchmark datasets in cybersecurity intrusion detection research was explored in this study. It has about 5 million records, with 41 input features and 1 target label. **Figure 3** shows the proposed model that comprises of Xai engine based on SHAP. After preprocessing, three ML

techniques were explored namely Random Forest (RF), XGBoost and Voting Ensemble method. If intrusion alert is triggered then the value of prediction is 1. SHAP provided global feature analysis, local explanations about why packets are flagged and feature contribution ranking as well.

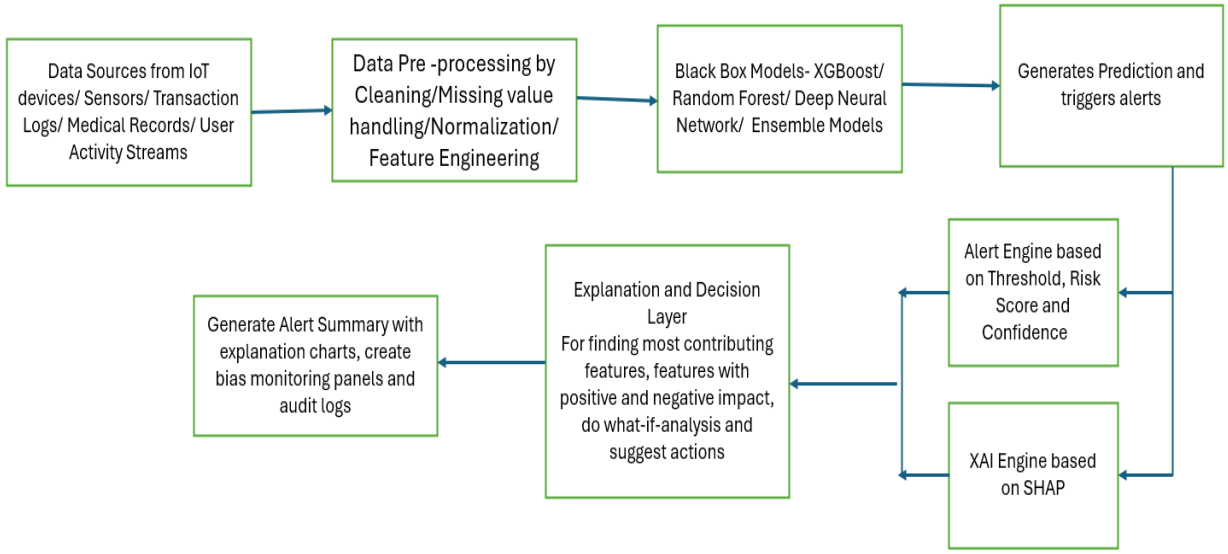

Figure 3. Proposed XAI model for Intrusion Detection with SHAP Engine

In this work four types of network intrusion attacks are detected namely Denial of Service (DoS), Probe, User to Root (U2R) and Remote to Local (R2L) attacks. **Table 2** explains these attacks briefly.

Table2: attack comparison table

Attack Type	Objective	Traffic Pattern	Severity	Key SHAP Drivers
DoS	Disrupt service	Very high volume	High	count, srv_count, dst_bytes
Probe	Reconnaissance	Scanning behavior	Medium	srv_count, service, diff_host_rate
U2R	Privilege Escalation	Low traffic, high system missuses	Very high	hot, privilege indicators
R2L	Unauthorized access	Repeated login attempts	high	logged_in, error rates

Discussions

Figure 4 depicts the Bee swarm plot which helps to interpret feature importance of ML models used for intrusion detection. It clearly depicts the features contributing for model's prediction as intrusion vs normal. The Y axes depict the features ordered from top to bottom based on overall impact on the model.

Table 3 depicts the SHAP value for the top features with count feature having utmost significance. The X axes in Figure 4 depicts the SHAP value calculated for the features based on **Eq. (1)** [5].

$$(\emptyset_i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad \text{Eq. (1)}$$

Here, $\emptyset_i$ is SHAP value of Feature i, N is set of all 41 features. S is considered as the subset of features not containing feature i, |S| is the number of features in subset S, while |N| is the total number of features. f(S) is the model's prediction using only features in subset S and $f(S \cup \{i\})$ is the prediction attained after adding feature i.

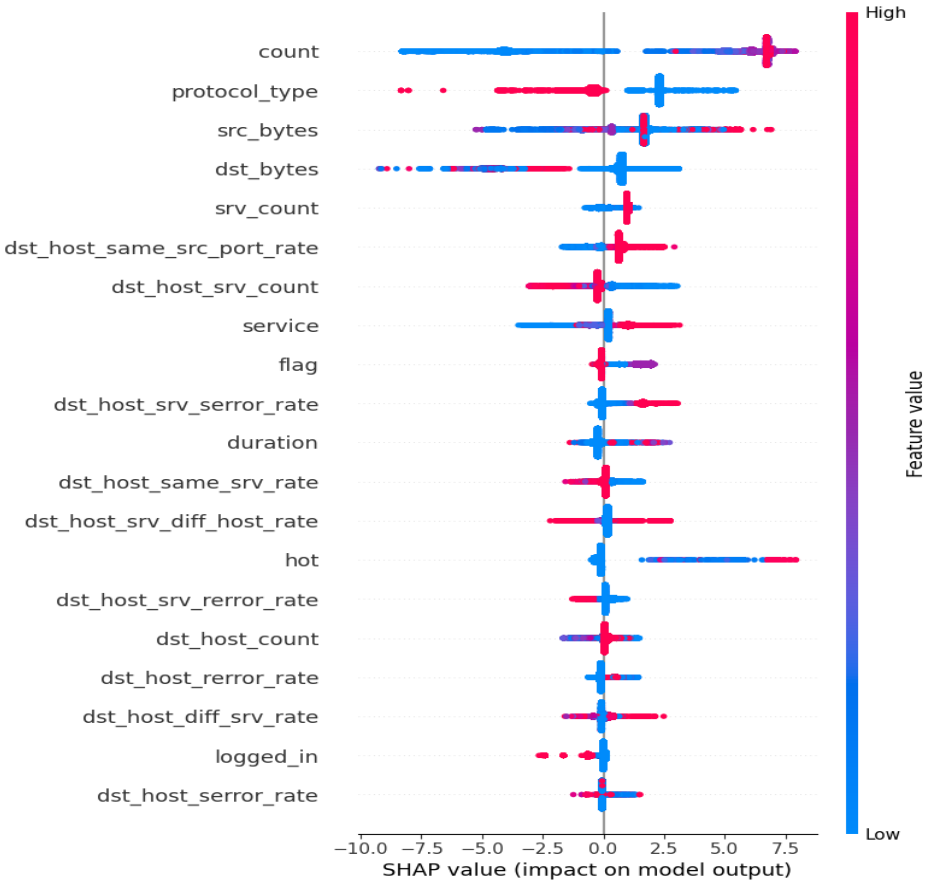

Figure 4. Bee swarm plot with SHAP explanations for XGBoost

Hence, from Figure 4 we can understand the features that have strong influence in intrusion detection systems as count, protocol_type, src_bytes. Here it may be noted that high values in red push the prediction to the right. For an intrusion detection system, high connection count in a short time increases attack probability, likely indicating DoS attack or a scanning behavior of an intruder which is essential aspect of a Zero Trust architecture showing that the system is continuously monitoring. The second most important feature noted as per Figure 4 is protocl_type as certain protocols strongly influence attack classification like ICMP or abnormal TCP.

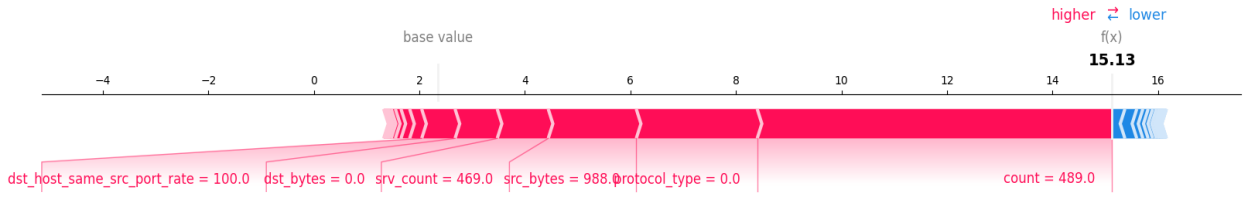

Figure 5. Local explanations for selected Sample

Figure 5 depicts SHAP Force Plot that explains the single prediction made by the Intrusion Detection model [6]. It indicates why the model predicted the given instance as an attack. Here f(x) is 15.13 that therefore strongly predicts an attack. The features that enabled the system to consider this as an attack are depicted in red with count feature having a value as 489, src_bytes as 988, srv_count as 469, dst_host_same_src_port_rate = 100.0.

Table 3: Feature contribution Table

ColumnID	Feature	SHAP Value
22	count	6.724935
1	protocol_type	2.292205
4	src_bytes	1.668737
23	srv_count	0.966768
5	dst_bytes	0.799057
35	dst_host_same_src_port_rate	0.632767
2	service	0.227493
36	dst_host_srv_diff_host_rate	0.164594
33	dst_host_same_srv_rate	0.106007
40	dst_host_srv_rerror_rate	0.078598

Table 2 represents the global SHAP feature importance values of the intrusion detection model. Our SHAP model relies heavily on connection frequency matrix, traffic volume features and protocol behavior,

helping our model primarily detect high-volume network anomalies, behavioral deviations and service based attack patterns.

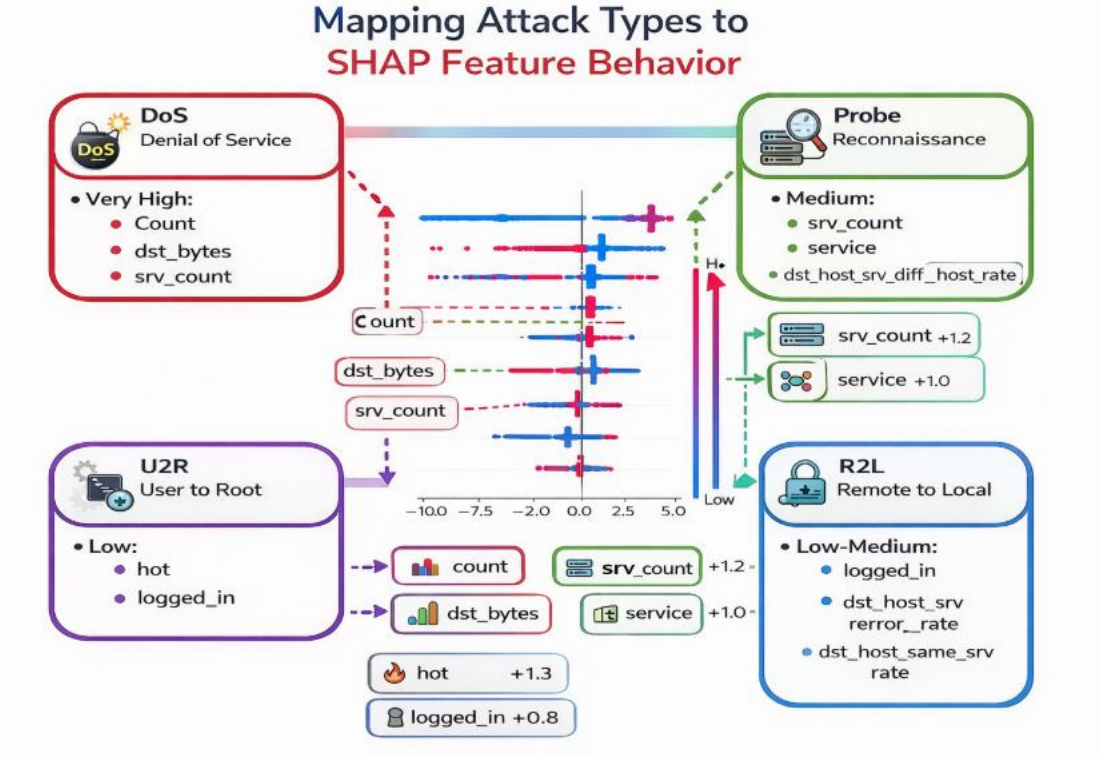

Figure 6 SHAP feature behavior for detecting attacks

Figure 6 depicts the four types of attacks as studied and detected in this research work using SHAP XAI engine. It also depicts the features that are insightful in detecting a given attack. SHAP distributions reveal that feature count dominance bias empowers U2R detection. The model detected DoS because of high connection count (+4.2 SHAP) and high service count (+1.8 SHAP) strongly increased attack probability.

Conclusions

1. Problem statement Addressed/ Motivation
   In this research work SHAP-based XAI explanations are attained for ML-based IDS models in order to enhance transparency, trustworthiness, and operational effectiveness within Zero Trust Architecture. Existing IDS research prioritizes accuracy over explainability. There is a clear need for an interpretable, XAI-driven intrusion detection framework that maintains high predictive performance while providing feature-level explanations to enhance trust, compliance, and adaptive security policy enforcement, which we address in this work. With XAI in ZTA trigger automated containment is aimed to achieve.

2. Method used

In this work three Machine Learning models namely Random Forest, XGBoost and Voting Ensemble methods were explored for Intrusion Detection and categorize DoS, Probe, U2R and R2L attacks that were further interpreted using SHAP Xai Engine.

3. Key findings

   The KDD Cup 1999 dataset, a benchmark intrusion detection dataset containing 41 network traffic features and labeled normal or attack connections was studied here. Four major attack categories—DoS, Probe, U2R, and R2L, widely used to evaluate machine learning-based intrusion detection systems were detected here, despite limitations such as class imbalance and outdated attack patterns. The SHAP summary plot reveals that connection count, protocol type, and traffic volume features significantly influence intrusion classification. High values of 'count' and 'hot' strongly push predictions toward attack class, indicating behavioral anomaly patterns consistent with scanning and DoS attacks. This demonstrates the suitability of SHAP for interpretable intrusion detection in Zero Trust environments.

4. Limitations of the work and future work that should be carried

   The study would be extended to realistic datasets like: NSL-KDD, CICIDS2017, UNSW-NB15. Detection of modern attack types like encrypted traffic, realistic network behavior will also be explored in further work. Currently traditional ML models have been explored but in future we intend to explore CNN-based methods for traffic classification, LSTM/GRU for sequential attack detection, Graph Neural Networks (GNNs) for network topology attacks. Future systems that detect zero-day attacks, model temporal attack patterns, handle large-scale enterprise traffic and reduce alert fatigue and support regulatory compliance are other possible directions of future work.

References

1. S. Teerakanok, T. Uehara, A. Inomata, "Migrating to Zero Trust Architectures: Reviews and Challenges", Security and Communication Networks, 2021. **https://doi.org/10.1155/2021/9947347**
2. K. Ramezanpour, J. Jagannath, "Intelligent zero trust architecture for 5G/6G networks: Principles, challenges, and the role of machine learning in the context of O-RAN", Computer Networks, vol. 217, part no. 109358, 2022. https://doi.org/10.1016/j.comnet.2022.109358
3. S. Alam, Z. Altiparmak, XAI-CF-Examining the role of explainable artificial intelligence in cyber forensics", Engineering Applications of Artificial Intelligence, vol. 167, part 3, 113892, March 2026. https://doi.org/10.1016/j.engappai.2026.113892
4. F. A. Ruambo, D. Zou, I. O. Lopes, B. Yuan, A. Muthanna, M. Zakarya, M. S. A. Muthanna, "Trust scoring algorithms for zero trust-based software-defined perimeter architectures: A systematic literature review of advancements, challenges, and future directions", Computers and Electrical Engineering, vol. 132, part no. 111002, 2026. https://doi.org/10.1016/j.compeleceng.2026.111002
5. F. Doshi-Velez, B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning", arXiv:1702.08608, 2017. https://doi.org/10.48550/arXiv.1702.08608

6. S. Lundberg, S. Lee, “A Unified Approach to Interpreting Model Predictions”, arXiv:1705.07874, 2017. https://doi.org/10.48550/arXiv.1705.07874