Interpreting Machine Learning-Based Intrusion Detection in IoT using Explainable AI: A Case Study with Explainable Boosting Machine*

Yizhou Xu and Parteek Kumar
Department of Computer Science
Whitman College
Walla Walla, WA 99362
{xuy2, kumarp}@whitman.edu

Abstract

In this article, we utilize an explainable AI approach, the Explainable Boosting Machine (EBM), to perform feature analysis on an extensive Internet of Things (IoT) dataset collected from real-world devices. After feature selection and data processing, our training dataset includes 2.88 million traffic data instances, categorized into six classes (DDoS, DoS, Mirai, Recon, Spoofing, and Benign). The EBM trained on this dataset achieved a impressive accuracy rate of 99.4% and an F1 score of 92.8%. Using the resultant model, we interpreted its predictions based on feature importance. The identified feature importance aligned well with established cybersecurity principles, indicating the model's potential. However, our analysis revealed that the machine learning model's predictions were strongly tied to the specific characteristics of the training IoT dataset, thereby raising concerns about the model's reliability when applied to real-world attack detection. Future research could explore the use of more diverse and balanced datasets or the applicability of the machine learning model in different IoT contexts, aiming to enhance the model's generalizability and practical relevance.

1 Introduction

Machine learning is progressively becoming a fundamental component of intrusion detection systems for the Internet of Things (IoT). Although numerous studies have shown promising results using machine learning, a common limitation is the lack of depth in explaining their methodologies, posing concerns about these methods' real-world applicability. In this article, we harness the power of explainable AI, specifically the Explainable Boosting Machine (EBM), to demystify the predictions made by our machine learning model. The EBM was selected for its robust performance, superior interpretability, and effective visualization capabilities. EBM, which combines the simplicity of linear or decision tree models with the precision of complex models like gradient boosting or random forests, functions as a "glass box" model, allowing both its predictions and its decision-making processes to be interpretable by humans.

We utilized a comprehensive IoT dataset as our training data[7]. Upon training the EBM on this dataset, we achieved an impressive accuracy rate of 99.4% and an F1 score of 92.8%. The feature importance produced by the EBM facilitated the interpretation of the model's predictions.

The major contributions of our work are:

- The decision-making of the machine learning model, trained on the IoT dataset, was interpreted using Explainable AI techniques (EBM).
- We cross-referenced the decisions made by our machine learning model to verify their alignment with fundamental cybersecurity principles.
- After thorough validation, we underscored the limitations of the machine learning model, particularly when trained on a specific dataset.

2 Related Work

The detection of malicious activity through internet traffic data remains a compelling avenue in cybersecurity research, with numerous explorations spanning several years. The evolving complexity of IoT systems has led to an increased adoption of machine learning techniques in the creation of Intrusion Detection Systems. The IoTIDs (2020) study assessed the efficacy of diverse traditional machine learning models, such as Linear Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Random forest (RF), Support vector machine (SVM), and Naive Bayes (NB), using IoT datasets[4]. In another work[1], the researchers developed an advanced deep neural network that outperformed Restricted Boltzmann machine (RBM), Sparse Autoencoder (SAE), and Stacked Denoising Autoencoder (SDAE) frameworks on an IoT dataset.

Despite extensive research into acquiring comprehensive data from IoT devices and the development of robust machine learning models, a thorough analysis of the trained models is frequently absent. For instance, the authors of the Edge-IIoTset (2021) ranked the importance of features for each class without providing any explanatory commentary[3]. The researchers behind WUSTL-IIOT (2021) determined feature importance based on their impact on the model's accuracy rate[12]. Yet, this traditional method's interpretability remains somewhat restricted. In another study[6], the authors concentrated on using feature importance to eliminate redundant features rather than interpreting the model.

3 Methodology

3.1 Explainable Boosting Machine

To comprehend the underpinnings of EBM, we first inspect the standard mathematical form of Generalized Additive Models (GAM)[5]:

$$g(E[y]) = \beta_0 + \sum f_j(x_j) \tag{1}$$

EBM can be viewed as an extension of the GAM, where the prediction is obtained by summing up individual feature functions, f_j. Feature functions encapsulate each feature's contribution to the overall prediction. In contrast to models like logistic regression, which presumes linear relationships, EBM capture non-linear associations, offering superior performance with complex datasets. Furthermore, by plotting the feature functions, f_j, we can identify and visualize the importance of each feature involved in the prediction[5].

In the mathematical formulation below[5], the algorithm for EBM is further improved to detect the pairwise interaction between features (i, j):

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum f_{i,j}(x_i, x_j) \tag{2}$$

3.2 Procedure and Evaluation

This study employs a comprehensive IoT dataset, obtained from real-world devices within a complex network topology[7]. Following feature selection and data processing, our training dataset comprises 2.88 million traffic data entries, spanning six classes: DDoS, DoS, Mirai, Recon, Spoofing, and Benign traffic. We used the utils class from Scikit-learn[8] to calculate the balanced class weight for each class, incorporating these weights into the EBM's parameters.

The processed dataset was then classified by the EBM. Table 1 presents the performance metrics of the resulting EBM model. The model showcased exemplary performance, achieving an impressive accuracy rate of 99.4% and a commendable F1 score of 92.8%. The recall and precision rates generally

Table 1: Results of multi-class classification using EBM

Metric	Explainable Boosting Machine
Accuracy	0.9940481456770219
Recall	0.9367583492627487
Precision	0.9204079536958586
F1-score	0.9276943954810214

maintained equilibrium, attributable to our effective data balance methods. These scores highlight superior predictive capabilities and overall efficiency. A thorough analysis of feature importance based on the EBM's outputs will be presented in the subsequent sections.

4 Results and Discussion

EBM evaluates feature importance through a raw score. In a classification context, the raw score serves as a measure of the evidence supporting a particular class. A high raw score implies strong evidence in favor of that class, while a lower score may indicate weaker support. To comprehend the nuances of feature importance, we must delve into the plots associated with each feature, taking advantage of EBM's data visualization tools. In Figure 1, the vertical axis denotes the raw score, while the horizontal axis represents the feature's value. The description of each curve's color can be found below Figure 1. We will assess features that significantly influence the model's prediction.

4.1 Inter-Arrival Times (IATs)

A crucial feature in the dataset, IATs, measures the time difference between two packets. Our testing designates this as a significant feature, demonstrating that its removal could reduce the machine learning model's accuracy by 15%. The first plot on the left in Figure 1 illustrates the feature importance of IATs. Data predominantly cluster around three intervals: 0s to 0.04s, 0.0829s to 0.0838s, and 0.1664s to 0.1669s. The first interval, from 0s to 0.04s, can be analyzed without the need to zoom in on the plot. The DoS attack class scores the highest, with scores fluctuating between 2.5 and 4, followed by Benign, with scores approximately 1.4. DDoS and Mirai are the least likely classes, with scores ranging from -0.8 to -2.2. Feature importance in the 0s to 0.04s interval is relatively insignificant for the remaining attacks. The second and third plots on the left in Figure 1 offer a detailed view of the IATs feature within the second and third intervals. In the 0.0829s to 0.0838s interval, we

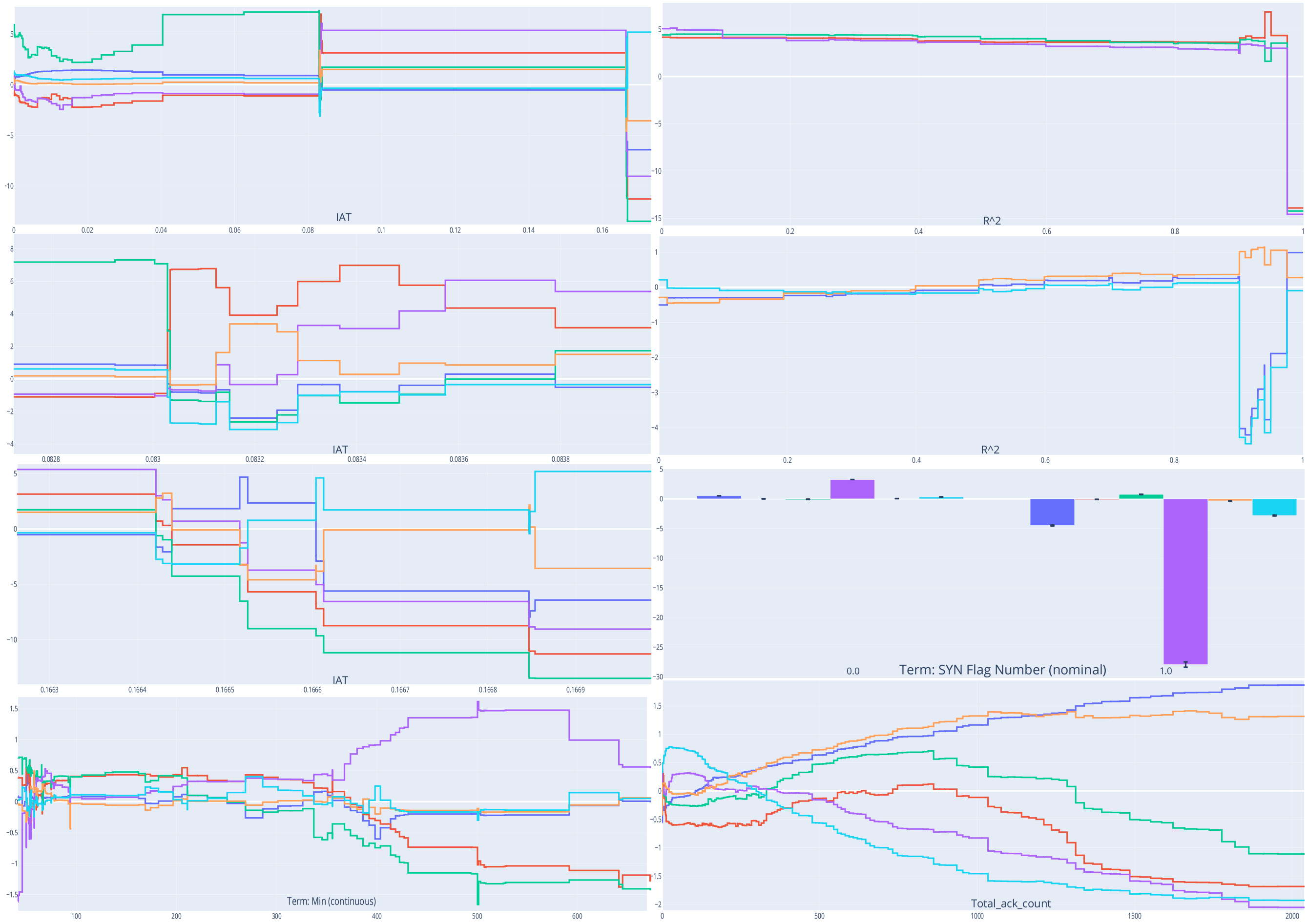

Figure 1: Representations are as follows - DDoS in Red, DoS in Green, Mirai in Purple, Recon in Yellow, Spoofing in Sky Blue, and Benign in Blue.

observe a rising trend in scores for Mirai attacks and consistently high scores for DDoS. Due to the negative feature importance, the machine learning model is less likely to predict DoS attacks within this interval. Lastly, the third plot on the left in Figure 1 shows a declining trend for the scores of DDoS, DoS, and Mirai attacks within the 0.1664s to 0.1669s interval.

These findings can be corroborated with fundamental cybersecurity principles. During DDoS or DoS attacks, attackers aim to overload a network with internet traffic, typically involving sending a large volume of packets in a short span of time. Hence, the IATs during an attack would be exceedingly low, often nearing zero. Mirai attacks involve remote control and device utilization in DDoS attacks, implying similar traffic characteristics to DDoS[2]. Given that these attacks are characterized by shorter IATs, it is improbable for them to reach a duration of 0.16s. This explains why the DoS class has the highest score in the lowest interval (0s to 0.04s) and why there are strong positive scores for DDoS and Mirai attacks from 0.0829s to 0.0838s.

Furthermore, in the second plot on the left in Figure 1, both Benign and Spoofing exhibit a similar pattern of negative feature importance in the 0.0829s to 0.0838 interval. Conversely, strong positive feature importance for Recon suggests a high likelihood of the model predicting this class. In a Recon attack, the attacker might generate noticeable patterns in the network's IATs during scanning. Hence, it is expected to observe high feature importance associated with certain intervals of IATs. In general, IATs analysis can yield valuable insights into understanding the decision-making process of the model.

4.2 The Coefficient of Determination (R^2)

The first and second plots on the right in Figure 1 illustrate the feature importance of R^2. This feature evaluates the coefficient of determination between the lengths of incoming and outgoing packets within a flow. It assesses the correlation between these packet lengths and can yield a value ranging between 0 and 1. During standard network operations, the sizes of incoming and outgoing packets often display a certain correlation. This correlation may, however, change significantly during an internet attack.

The first plot on the right in Figure 1 shows that DDoS, DoS, and Mirai attacks exhibit similar correlations. The feature importance remains consistently high, around 3.5, when R^2 is between 0 and 0.97, but abruptly falls to -14 when R^2 is above 0.97. The model also suggests that a DDoS attack is less likely if the coefficient of determination exceeds 0.97. This observation aligns with basic cybersecurity principles. During DoS, DDoS flooding, the correlation between incoming and outgoing packet sizes may decline as the network becomes inundated with unwanted packets[10].

The second plot on the right in Figure 1 displays the feature importance for Recon and Spoofing attacks. Neglecting the outliers, the peak feature importance of 1.1 for Recon appears when R^2 lies between 0.9 and 0.97. Recon attacks typically involve gathering information about a target system, often as a precursor to a more direct assault. An adept attacker might endeavor to imitate the standard packet size distribution of the network to evade detection[9]. This scenario explains why the feature importance escalates as the coefficient of determination approaches 1. In the case of Spoofing, the feature importance score ranges between -4 and -2 when the R^2 value is between 0.9 and 0.97. The correlation between incoming and outgoing packet sizes is not apparent in such attacks. The machine learning model's generated negative feature importance aligns with this observation. In general, R^2 serves as a strong indicator for the machine learning model to distinguish DDoS, DoS, and Mirai attacks from Recon, Spoofing, and benign traffic.

4.3 TCP Flags (ACK flag and SYN flag)

Transmission Control Protocol (TCP) flags, despite their eight different types, are not all reliable indicators of attacks. Upon analyzing their influence on our model's accuracy, we discovered that the acknowledgment (ACK) and Synchronisation (SYN) flags played a more significant role. The SYN flag, typically used to establish connections, can also be exploited in certain forms of cyberattacks. The third plot on the right in Figure 1 presents the feature importance of the SYN flag for each attack. As a categorical variable, the SYN flags are either 0 for unset (on the left of the plot) or 1 for set (on the right). For the Mirai attack, a set SYN flag signifies a strong negative feature importance, approximately -28. The SYN flag's relevance to Mirai hinges on the types of DDoS attacks the Mirai botnet implements. Our dataset solely contains three forms of Mirai attacks - GRE IP Flood, GRE Ethernet Flood, and UDP Plain attacks. These attacks operate on the Generic Routing Encapsulation (GRE) and User Datagram Protocol (UDP), which belong to a different layer than TCP and do not utilize the SYN flag. Additionally, we observe that Benign and Spoofing also show negative feature importance, with scores of -4.5 and -2.7, respectively. Similar to Mirai, neither Address Resolution Protocol (ARP) nor Domain Name Service (DNS) Spoofing involves the SYN flag. In terms of benign traffic, a high count of SYN packets might suggest malicious activity. The negative feature importance for Benign is likely due to the interaction the model identified between the SYN and ACK flags.

The fourth plot on the right in Figure 1 reveals the feature importance of the total ACK count for each class. Unlike the former feature, the ACK count measures the total number of ACK flags in a flow and is thus a continuous variable. The ACK flag is predominantly used in benign traffic to signal the acknowledgment of packet receipt. Both the fourth and fifth graphs demonstrate a positive correlation between feature importance and benign traffic. During a DDoS attack, the attacker inundates the target's network with excessive internet traffic. Yet, DDoS attacks frequently involve a large quantity of spoofed SYN packets as opposed to ACK packets. A negative feature importance for a DDoS attack is expected for high values of ACK count. As mentioned earlier, DoS and Mirai attacks bear similarities to DDoS attacks; thus, these types of attacks exhibit similar patterns of feature importance[2]. Regarding Recon attacks, ACK flags can be employed in various network scans, such as TCP ACK scanning. The plot also reveals a positive correlation between the ACK flag feature importance and Recon attacks. Lastly, given that our dataset does not include instances of ACK Spoofing, we anticipate a negative feature importance for Spoofing.

In combination with R^2, the SYN and ACK flags allow the machine learning model to classify attacks more effectively. Based on previous features, the model can segregate classes into two groups: DDoS, DoS, Mirai, and Recon versus Spoofing and Benign. A set SYN flag eradicates the possibility of Mirai in the first group and reduces the likelihood of Spoofing and Benign in the second group. The ACK count assists in identifying DDoS and Recon attacks while diminishing the chance of Spoofing.

4.4 Discussion

While the EBM yields an impressive accuracy rate of 99% and an F1 score above 90%, our research still has several limitations and offers potential areas for improvement. The dataset encompasses 105 IoT devices, which, although considerable, might only be representative of specific types or categories of IoT devices encountered in real-world scenarios. Furthermore, it includes only specific instances of DoS, DDoS, Spoofing, Recon, and Mirai attacks. The omission of other types of attacks from the dataset, such as Web-based and BruteForce, could constrain the scope of cybersecurity threats that our research is capable of addressing.

Additionally, the original dataset does not proportionally represent the individual methods employed in each attack, which could affect the feature importance's reliability. For instance, it houses 7 million instances of Internet Control Message Protocol (ICMP) Flood for DDoS attacks, contrasted with a scant 29 thousand instances of Hypertext Transfer Protocol (HTTP) Flood. This unequal representation may skew the EBM's feature importance, as it tends to lean towards the majority group. We have ascertained that the specific characteristics of the dataset heavily influence the predictions made by the EBM. For example, when examining the Mirai attacks, one feature, the minimum packet length, has its feature importance peak at 501 to 591 bytes (the fourth plot on the left in Figure 1). This peak is primarily because our dataset only includes three types of Mirai attacks—GRE IP Flood, Greeth Flood, and UDP Plain—which all display packet sizes within the 554 to 592 bytes range[11]. Other variants of Mirai attacks, like the SYN Mirai attack, which has a packet size of merely 74 bytes[11], are absent from our dataset. As a result, the IoT analysis we suggested, based on feature importance, is limited to the attack types included in our dataset. Encountering new attack types may significantly undermine the machine learning model's accuracy rate. Future research should aim to obtain a sufficiently comprehensive dataset, featuring the most prevalent internet attack forms while maintaining a balanced representation for each class.

It's also important to note that a machine learning-based Intrusion Detection System (IDS) for IoT should be deployed alongside other security measures, such as encryption, authentication, and access control mechanisms, to provide a comprehensive security solution.

5 Conclusions and Future Scope

The utility of machine learning-based intrusion detection systems, while promising, requires careful implementation and evaluation. In our study, we trained an EBM on an extensive IoT dataset to identify potential attacks. Through an analysis of feature importance, we were able to interpret the decision-making process of the machine learning model. Our validation process generally agreed that the predictions made by the machine learning model align with fundamental cybersecurity principles. However, there were instances where the model's predictions were challenging to explain or did not conform with these principles. Furthermore, we identified limitations regarding the model's predictions. Our analysis revealed that the machine learning model's predictions were strongly tied to the specific characteristics of the training IoT dataset, thereby raising concerns about the model's reliability when applied to real-world attack detection.

As we move forward in developing machine learning-based intrusion detection systems, it is vital to ensure the training dataset meets three critical criteria. Firstly, the network topology for IoT devices should be comprehensive, mimicking real IoT operations closely. Secondly, the dataset should encompass a wide variety of attack types, utilizing all prevalent tools and frameworks employed to execute these attacks. Lastly, the dataset should be balanced not only across different attack types but also across the various instances of methods used to execute such attacks. Despite satisfying these criteria, it is important to understand the limitations of these models in real-world applications. The system should ideally be applied only to a similar set of IoT devices as those in the training set, and it should be recognized that attacks not included in the training set may not be detected. By addressing these factors, we believe the utility and reliability of machine learning-based intrusion detection systems for IoT can be significantly improved.

References

[1] Adel Abusitta et al. “Deep learning-enabled anomaly detection for IoT systems”. In: *Internet of Things* 21 (2023), p. 100656.

[2] Manos Antonakakis et al. “Understanding the mirai botnet”. In: *26th USENIX security symposium (USENIX Security 17)*. 2017, p. 1095.

[3] Mohamed Amine Ferrag et al. “Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning”. In: *IEEE Access* 10 (2022), p. 40303.

[4] Hanan Hindy et al. “Machine learning based iot intrusion detection system”. In: *arXiv preprint arXiv:2006.15340* (2020), pp. 73–84.

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

[6] Achmad Akbar Megantara and Tohari Ahmad. “Feature importance ranking for increasing performance of intrusion detection system”. In: *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*. IEEE. 2020, pp. 37–42.

[7] Euclides Carlos Pinto Neto et al. “CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment”. In: (2023).

[8] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[9] Sirikarn Pukkawanna, Youki Kadobayashi, and Suguru Yamaguchi. “Network-based mimicry anomaly detection using divergence measures”. In: *2015 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE. 2015, p. 1.

[10] Zhongmin Wang and Xinsheng Wang. “DDoS attack detection algorithm based on the correlation of IP address analysis”. In: *2011 International Conference on Electrical and Control Engineering*. IEEE. 2011, p. 2951.

[11] Ron Winward. *Iot attack handbook*. 2018. URL: `https://falksangdata.no/wp-content/uploads/2021/04/Mirai-Handbook.pdf`.

[12] M Zolanvari et al. “WUSTL-IIOT-2O2l Dataset for IIoT Cybersecurity Research”. In: *Washington University in St. Louis, USA* (2021).