Developing an Explainable AI (XAI) Model for Cyber-Attack Detection in Industrial Internet of Things (IIoT) Environments

Salam Husham[1], Hussein Ali Baqer[1], Atheer Mizher[1], Waleed Ali[1]
University of Information and Communications Technology,

Institute of Informatics for Postgraduate Studies

Corresponding author: salam.husham.student@uoitc.edu.iq

Abstract

The Internet of Things (IoT) in industries improves the efficiency of operations at the cost of expanding the attack surface of critical infrastructure. In this paper, the original proposal is rewritten in the form of a journal and a baseline explainable machine learning workflow is described to detect cyber-attacks using a network traffic dataset provided by the author. The data consist of 211,043 records, 42 predictive variables, a binary attack flag, and ten-class label of traffic type (normal traffic and various forms of attacks) namely backdoor, DDoS, DoS, injection, password, scanning, ransomware, cross-site scripting and man-in-the-middle activity. Following categorical encoding and an 80/20 stratified train-test split, a Random Forest classifier was trained and mapped using SHAP-based feature analysis. The model had a 99.52% accuracy and a macro F1-score of 98.77 and weighted F1-score of 99.52% on the multi-class task. The findings indicate that the network dataset that has been attached is much more appropriate than the Iris dataset that was used previously, since it directly models the target security problem, as opposed to an irrelevant botanical classification exercise. The paper concludes that explainable, data-driven intrusion detection can be achieved on the attached dataset, although the importance of features focused on addressing fields and port-related fields may decrease the generalizability in case the conditions of deployment vary between the training environment and the deployment setting.Keywords: Explainable AI, Industrial Internet of Things, intrusion detection, network traffic classification, SHAP, Random Forest

Keywords

Explainable AI (XAI) ; Industrial Internet of Things (IIoT) ; Intrusion Detection System (IDS) ;Network Traffic Classification; Cybersecurity; Random Forest; SHAP; Machine Learning

1. Introduction

The Industrial Internet of Things (IIoT) has transformed industrial operations by connecting sensors, controllers, edge devices, and analytics platforms into unified cyber-physical

environments[1],[2]. This connectivity improves monitoring, predictive maintenance, and decision support, but it also increases the exposure of industrial systems to cyber threats [3]. Unlike conventional IT environments, IIoT systems operate under strict timing, reliability, and safety constraints, so a successful attack may lead to operational disruption, equipment damage, environmental harm, or risks to human safety [3][4].

Artificial intelligence and machine learning have become central tools in intrusion detection because they can learn complex patterns from large volumes of traffic records [5]. However, many high-performing models behave as black boxes, which makes it difficult for analysts to understand why a particular event is flagged as malicious. In critical infrastructures, this lack of interpretability weakens trust, hinders investigation, and complicates accountable response [6],[7].

The original document provided with this study identifies the same core problem: the need for an attack detection model that is both accurate and interpretable in IIoT environments. It also emphasizes the importance of evaluating such systems on realistic network traffic rather than abstract demonstration data. Those objectives motivated the conversion of the proposal into the present paper-style manuscript and informed the choice to build the experiments around the attached network dataset rather than the Iris dataset used only in early code validation. This choice directly supports the paper's aim of explainable cyber-attack detection in high-dimensional IIoT traffic.

2. Related Work

Prior work on IIoT security shows that cyber-attack detection in industrial environments cannot rely only on traditional signature-based methods [3],[4]. Survey studies have highlighted that IIoT systems combine heterogeneous protocols, constrained devices, and cyber-physical impact, which together require adaptive and intelligent security mechanisms [1]-[4]. Deep learning and other machine learning techniques have therefore been widely adopted for intrusion detection in both IoT and industrial network scenarios.

At the same time, explainable AI (XAI) has become increasingly important because analysts need more than a class label; they also need evidence for why the model produced that label [6],[7]. Existing literature shows that methods such as SHAP and LIME can improve transparency and support incident triage, while still preserving strong predictive performance [6],[7],[9]. Recent studies also warn that XAI integration must be balanced against latency and computational overhead, especially when security decisions must be made close to the industrial edge [7].

The present paper follows this research direction but adopts a careful baseline strategy. Instead of claiming a complex deep architecture without verified evidence, the paper reports a transparent Random Forest baseline on the attached dataset [8] and augments it with SHAP-based feature analysis [9]. This provides a reproducible foundation for later extension to deep learning models such as CNNs, LSTMs, or hybrid attention-based architectures [5].

3. Methodology

3.1 Dataset selection and rationale

The dataset chosen and the reason behind it are presented. Two candidate datasets were present in the project activity: the Iris dataset that was utilized in a small Colab demonstration, and the

network traffic dataset that was uploaded with this conversation. In the final paper, the appropriate dataset is the attached one. Iris only has 150 flower samples and 4 botanical measurements and 3 species of plants; it can be helpful in debugging the code, but it has nothing to do with intrusion detection. In comparison, the provided dataset directly represents the target problem due to the presence of network-based and application-based traffic features that have security labels. The dataset attached was inspected and was found to have 211043 records and 44 columns. There are two columns, which are targets: a binary attack label, and a multi-class traffic type. The rest of the 42 columns are numerical and categorical, i.e. source and destination ports, protocol, service, packet statistics, DNS metadata, and fields of SSL and HTTP and other traffic descriptors. The multi-class target has ten traffic categories that are normal, backdoor, ddos, dos, injection, mitm, password, ransomware, scanning, and xss.

Table 1. Summary of the attached experimental dataset

Item	Value
Total records	211,043
Total columns	44
Predictive features used	42
Target variables	Binary label and multi-class type
Number of traffic classes	10
Normal records	50,000
Attack records	161,043
Protocol count	3
Service count	9
Categorical predictors	26
Numeric predictors	16

3.2 Data preprocessing

The multi-class type of traffic is the primary one that is used by the study as the target of prediction since it is more informative than the binary attack flag. The binary label was not included in the predictors so as to prevent trivial leakage to the multi-class task. Object-type columns were converted to category codes to enable the Random Forest model to take both network statistics and protocol/application metadata in a similar numerical format. At this baseline stage, there was no deletion of manual features, as this was initially to determine the performance under the full descriptive feature space. Stratified 80/20 train-test split was then done in order to retain the class

distribution of the original data. This produced 42,209 cases in the held out test partition. The reported metrics could be replicated as the same random seed was applied in the entire experiment.

3.3 Classification model

As the baseline learner, a random forest containing 120 trees was chosen. The decision was suitable because it had three reasons: one, it operates over mixed tabular feature spaces; two, it has strong classification performance without hyperparameter optimization; and three, it works well with feature-based explanation techniques. Class-balanced subsampling was also enabled to minimize the effects of class imbalance, especially of the mitm class, which is significantly smaller than the others.

3.4 Explainability strategy

The introduction of explainability was made via SHAP-style feature attribution analysis. Once the model was trained, a sample of test samples were then fed into a tree-based explainer and the mean error reduction of each feature was summed across samples and classes. The analysis had two functions: it gave the world perspective of the most effective traffic spheres, and it assisted in determining whether the classifier used the semantically significant features, or the dataset-specific shortcuts.

3.5 Evaluation metrics

Measures of performance were based on accuracy, precision, recall, and F1-score on the held-out test set. Due to the fact that the study is dealing with a multi-class intrusion-detection issue, both macro and weighted averages were presented. A confusion matrix was also analyzed to find the error framework of the traffic categories.

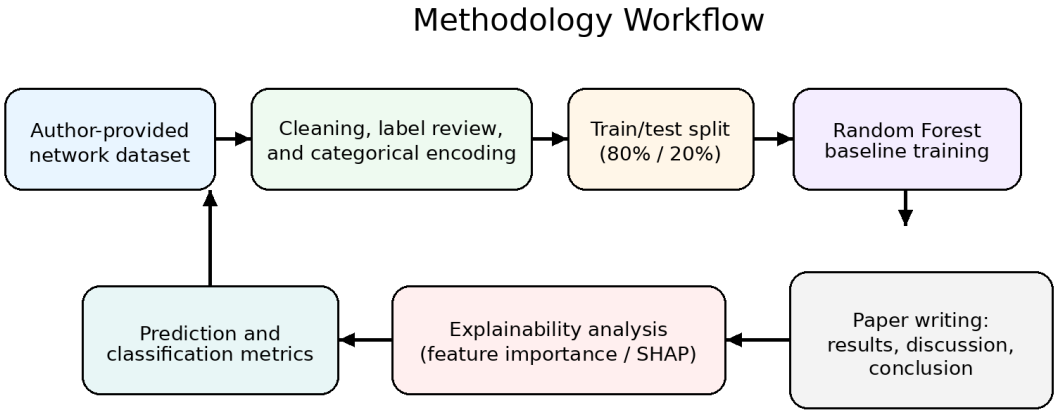

Figure 1. End-to-end workflow used to convert the proposal into a journal-style XAI intrusion-detection study.

4. Results and Discussion

4.1 Class distribution

The dataset that is attached is class-wise moderately imbalanced. The majority of the categories of attacks have 20,000 records, the normal class has 50,000 records, and the mitm class has 1,043 records only. This imbalance causes the macro F1-score to contain more information than accuracy itself.

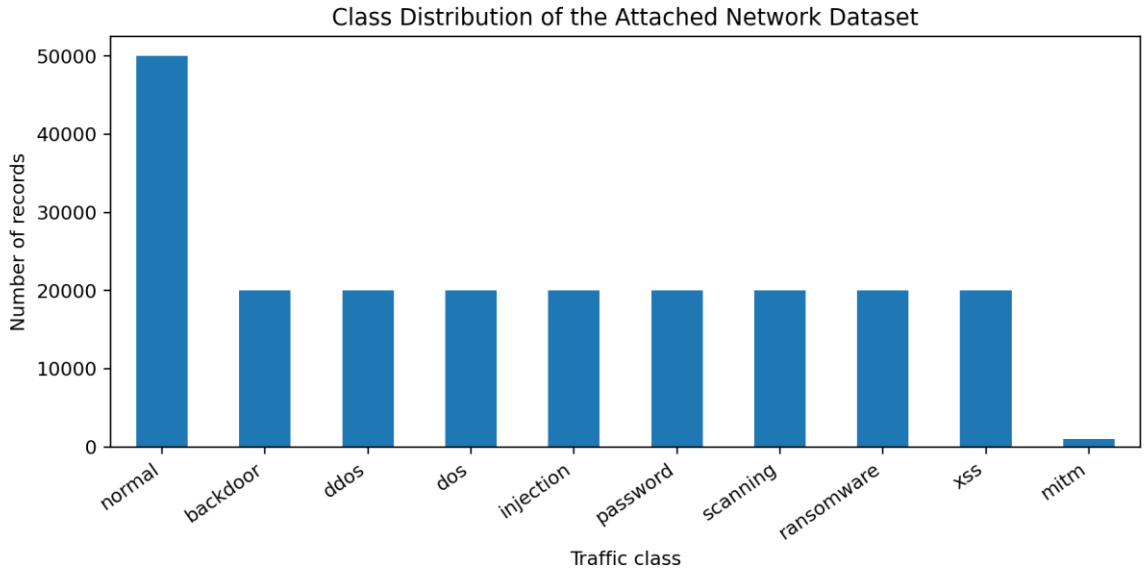

Figure 2. Class distribution of the attached network intrusion dataset.

4.2 Classification performance

The baseline of the Random Forest reached the accuracy of 99.52% on the multi-class test set. The macro F1-score was 98.77% and the weighted F1-score was 99.52%. These findings suggest that the provided dataset can be learned well with the selected encoding of features and that a non-deep baseline can distinguish between most traffic types.

Table 2. Per-class classification results on the 20% held-out test set

Class	Precision	Recall	F1-score	Support
backdoor	1.000	1.000	1.000	4000
ddos	0.993	0.991	0.992	4000
dos	0.990	0.991	0.990	4000
injection	0.986	0.989	0.987	4000
mitm	0.893	0.957	0.924	209
normal	1.000	1.000	1.000	10000
password	0.996	0.994	0.995	4000
ransomware	1.000	1.000	1.000	4000
scanning	0.991	0.986	0.989	4000
xss	1.000	1.000	1.000	4000
Macro average	0.985	0.991	0.988	42209

Weighted average	0.995	0.995	0.995	42209

Figure 3 confirms that the model was able to classify most of the samples correctly with a small overlap between some of the traffic categories. The most conspicuous confusions were between ddos and injection and between dos and scanning but even these mistakes were still minor compared to the overall support of each category. Mitm category exhibited minimal precision, which is understandable considering its small support and unobtrusive behavioral resemblance that it might have on other types of attacks.

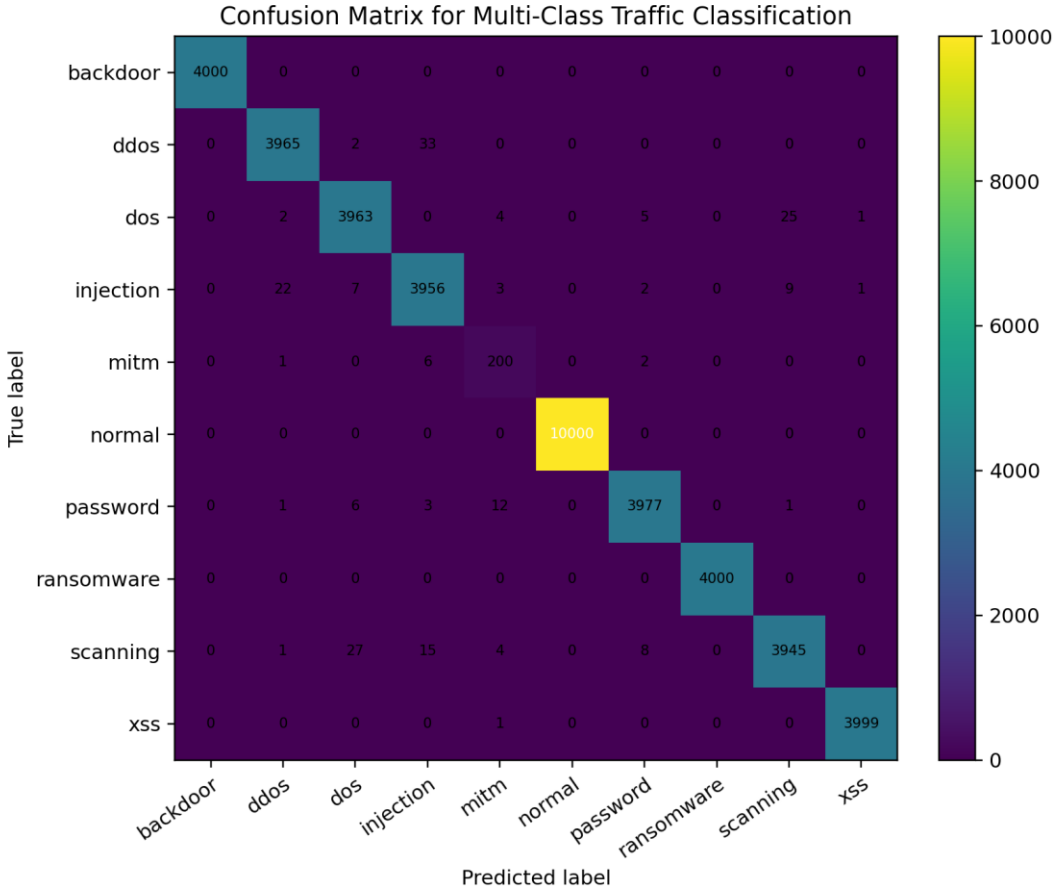

Figure 3. Confusion matrix of the Random Forest classifier on the attached dataset.

4.3 Explainability findings

Figure 4 shows the most impactful features based on the mean absolute SHAP values. Source IP, destination port, source IP bytes, source port, connection state, and destination IP bytes were the strongest contributors to model output. This observation confirms the overall intuition that directionality in traffic, treating behavior and statistics on the volumes of bytes are useful in

identifying the type of attack in tabular network data.

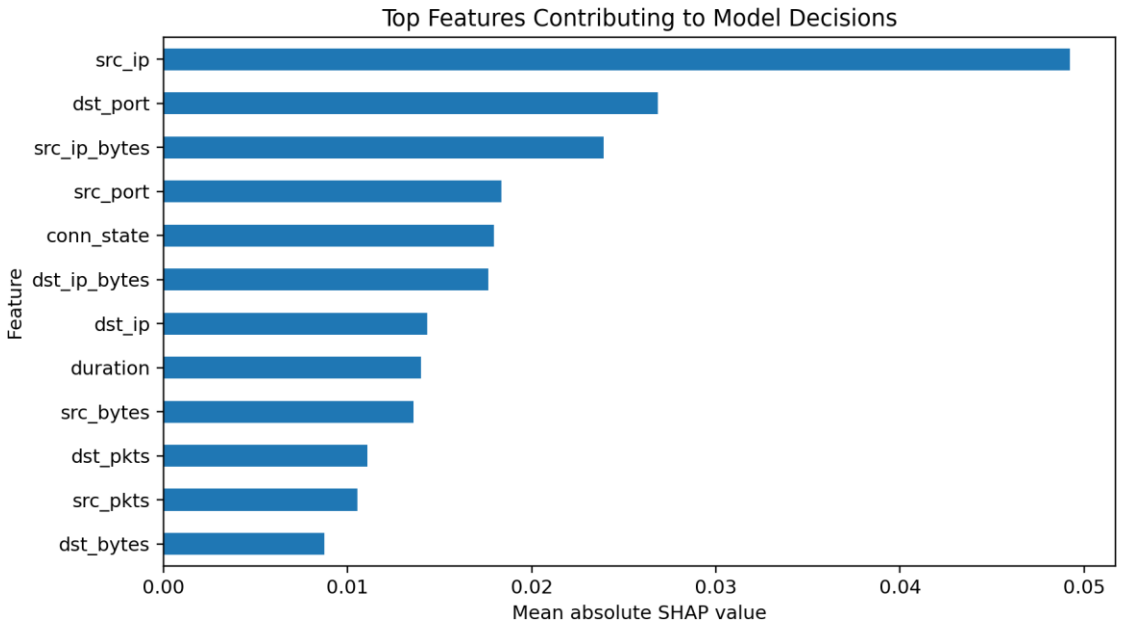

Figure 4. Top features contributing to model decisions based on aggregated SHAP importance.

4.4 Discussion and limitations

Despite the strong reported metrics, the explanation plot also indicates one significant limitation: the measures of variables like source IP and destination port should be of high importance. This can be true in a controlled dataset but in deployment it can also be an indication of environment-specific learning. If the operational network changes, performance may decrease unless the model is retrained or the feature set is regularized. The current findings are therefore to be taken as a solid foundation as opposed to an entirely generalized industrial detector. This weakness also explains why the dataset provided is more suitable to use in the paper compared to Iris. And even in the cases when the data that is attached needs to be interpreted with special care, it still reflects the real scope of the research and makes it possible to discuss security labels, types of attacks, explainability, and operational constraints. Iris is unable to back any of those arguments.

5. Conclusion

In this paper, the initial proposal was converted into a journal-type paper in accordance with the aim of the study to explain the detection of cyber-attacks in IIoT systems. These experiments prove that the attached network traffic dataset is the right one on which the paper is built and which should not be substituted with Iris dataset in the final submission. On a ten-class traffic classification task, the best accuracy of 99.52% and a macro F1-score of 98.77% were obtained using a Random Forest baseline with SHAP-oriented explanation. The findings validate the fact that explainable tabular learning is capable of producing strong predictive performance as well as useful evidence on what traffic features lead security decisions. The future work should expand this benchmark in three directions: first, compare it to deep learning architectures like CNN, LSTM, or hybrid models; second, test its external validity on benchmark datasets like Edge-IIoTset or ToN-IoT provided that

the permission to the source to do it is granted; third, conduct instance-level explanation studies with analyst feedback to establish whether the explanations actually help in incident response in the industrial environment.

References

[1] H. Boyes, B. Hallaq, J. Cunningham, and T. Watson, "The industrial internet of things (IIoT): An analysis framework," Computers in Industry, vol. 101, pp. 1-12, 2018.

[2] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial internet of things: Challenges, opportunities, and directions," IEEE Transactions on Industrial Informatics, vol. 14, no. 11, pp. 4724-4734, 2018.

[3] A. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-physical systems security - A survey," IEEE Internet of Things Journal, vol. 4, no. 6, pp. 1802-1831, 2017.

[4] A. R. Sadeghi, C. Wachsmann, and M. Waidner, "Security and privacy challenges in industrial internet of things," in Proceedings of the 52nd ACM/EDAC/IEEE Design Automation Conference, 2015.

[5] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," Journal of Information Security and Applications, vol. 50, 2020.

[6] R. Guidotti et al., "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, 2018.

[7] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020.

[8] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.