Vol. 08, No. 2 (2026) 1017-1026, doi: 10.24874/PES08.02A.015

Proceedings on Engineering Sciences

www.pesjournal.net

ROBUST FLOW-BASED IOT INTRUSION DETECTION FOR SUSTAINABLE SMART INFRASTRUCTURE: CROSS-CAPTURE EVALUATION AND INTERPRETABLE ATTACK FINGERPRINTS

Soorya S Suresh
Pragati Sachdeva
Amarjit Malhotra
Deepika Kukreja[1]

Received 11.03.2026
Revised 24.04.2026
Accepted 15.05.2026

Keywords:

Internet of Things (IoT), Intrusion detection, Flow-based network traffic, Cross-capture generalization, Permutation importance, Sustainable smart infrastructure

A B S T R A C T

Internet of Things (IoT) devices are crucial in sustainability programs for monitoring energy, air, water, and transportation systems. These systems need reliable, timely data, which can be interrupted by compromised networks. This study examines whether an artificial intelligence model can still perform well when network conditions change. Using flow-based tabular data from the CIC IoT-IDAD dataset, traffic is divided into normal and three attack types: DoS-UDP, DDoS-ICMP, and Mirai. Two practical machine learning models, Logistic Regression and Random Forest, are assessed. The results show that while the best model performs well in a standard random data split (Macro-F1 up to 0.84), its performance drops sharply when trained on one capture group and tested on another (average Macro-F1: 0.34-0.39). This study highlights a significant generalization gap in real deployment scenarios. Additionally, feature importance shows that timing patterns and the sizes of packets or flows consistently serve as reliable indicators across changing network conditions.

© 2026 Published by Faculty of Engineerin

1. INTRODUCTION

IoT devices are a part of the infrastructure in many sustainability-oriented systems, such as smart grids, climate sensing, water monitoring, and intelligent transportation systems. These systems are continuously gathering information in order to make decisions in the real world, such as balancing the amount of electric power, detecting pollution, managing water usage, and managing traffic flow. Data accuracy is a necessity, not a luxury, in such systems. If the system is under attack, the flow of data may be delayed, congested, or even compromised, resulting in incorrect decisions and wastage of precious resources, thereby compromising the very idea of climate stewardship, energy efficiency, and sustainable operations.

In recent times, many researchers have successfully implemented AI/ML models in detecting IoT attacks, achieving promising results (Soni et al., 2024; Soni et al.,

[1] Corresponding author: Pragati Sachdeva
Email: pragati.schdv@gmail.com

2025; Sangaraju et al., 2025). Nevertheless, the results are mostly based on random train-test splits, in which the train and test sets are mixed in a way that is easy to perform experiments on. In contrast, in real-world applications, the network is constantly dynamic, meaning that the state of the network may change in both space and time. For example, in the case of intelligent transportation systems, the model may be trained using data collected in a single capture, only to be exposed to different types of devices, diverse background traffic, or even different network loads in a later capture. Under such circumstances, the model may perform well in random tests, only to perform poorly in real applications. In the context of sustainability-oriented applications, such failures are critical, in that attacks may go unnoticed, thereby compromising the integrity of the data used in making decisions in the real world. Thus, robustness becomes an essential requirement in the context of sustainability-oriented systems.

In this study, the subject of flow-based intrusion detection is explored using the CIC IoT-IDAD dataset. Specifically, the present study focuses on four types of network traffic: normal traffic (Benign) and three types of attacks that have been identified as the most frequent: DoS-UDP, DDoS-ICMP, and Mirai. Two lightweight and easily implementable approaches have been selected for the present study: the Logistic Regression approach with standard scaling, known for its simplicity and interpretability, and the Random Forest classifier, a popular and powerful alternative. The selection of these two approaches is based on the fact that they can be implemented in IoT environments. In order to ensure the best results from the experiment, the balanced sampling method is used with 18,000 samples for each class.

Contributions of this study are as follows:

1. Capture-wise cross-capture Evaluation: In the present study, we introduce the protocol of the cross-capture evaluation method, in addition to the random split method. This is based on the assumption that the two captures have different characteristics. In the cross-capture method, the model is trained on one file and tested on the other file.
2. Generalization Gap: In the present study, the authors have measured the generalization gap between the two methods. This is based on the assumption that the random split method may not be the best approach for the task at hand.
3. Interpretability: In the present study, the authors have used the permutation importance method to determine the importance of the features. In addition, the authors have identified the feature-based "attack fingerprints" by determining the features that have the highest importance in the two scenarios.
4. Simple and Reproducible Pipeline: In the present study, the authors have designed a simple and reproducible pipeline based on the two lightweight and easily implementable approaches.

The rest of this paper is organized as follows: Section II provides related works and backgrounds, Section III describes the data and the data pre-processing methods, Section IV introduces the methodology and the evaluation protocols, Section V shows the experimental results, and Section VI concludes this paper, highlighting some important points and possible future works.

2. RELATED WORK

Infrastructure is now more and more built with sustainability in mind and utilizes Internet of Things (IoT) networks such as smart grids, water monitoring, air quality monitoring, and transportation systems. These are not just for monitoring; they are for leveraging the data for actions and resource optimization as well. Thus, the need for cybersecurity as a sustainability factor is born. Interception of communication in IoT networks can cause incorrect and delayed data, thereby causing inefficient and wasteful outcomes. Thus, the need for AI and ML-based intrusion detection systems for IoT networks arose as a means of automatic attack detection (Chen et al., 2022; Meyer-Berg et al., 2020; Liu et al., 2023; Neto et al., 2023; Abbasi et al., 2021; Noura et al., 2019; Elhadi et al., 2018).

One of the main reasons for the rapid progress in this field is the availability of considerable and realistic public datasets. CICIoT2023 provides large-scale IoT traffic, including many types of attacks such as DoS, DDoS, and Mirai botnets. It is commonly used as a benchmark for IoT intrusion detection systems (Neto et al., 2023). CICIoMT2024 extends this benchmarking concept for Internet of Medical Things scenarios, providing many protocols and device interactions so that the reliability of the model outside the usual IoT traffic scenarios can be tested (Dadkhah et al., 2024). Moreover, the need for realistic evaluation is emphasized in the research by Rabbani et al., where they show the importance of device behavior and network traffic when the model is deployed outside the usual test scenarios (Rabbani et al., 2025).

With the availability of such datasets, many recent research studies focus on improving the accuracy of the model through the use of deep learning and hybrid models. For example, the use of deep sequence models and transfer learning is proposed for improving the performance in the context of the Industrial Internet of Things (Nandanwar & Katarya, 2024). A hybrid feature selection model is proposed for improving the performance and reducing redundancy in the context of IoT NIDS (Walling & Lodh, 2024). All such research studies show the feasibility of good performance on the features extracted in the flow-based model; however, the majority of the studies use a random split approach where the data for the model is taken from the mixed traffic in the capture session. This approach, although convenient, overlooks the fact that the model could end up learning the characteristics specific to the capture session and not the attacks themselves.

This concern is not merely hypothetical; it has manifested in the form of empirical work that directly tests generalization. Doménech et al. compare the performance of the model across the IoT and IoMT datasets and report significant reductions in F1 score when the model is trained and tested across different datasets, suggesting that the model is reliable when operating in the same environment but not when the environment changes (Doménech et al., 2025). This is evidence that supports the argument that the main problem is not accuracy, as the F1 score is merely reduced, and reliability is the main challenge. Another study by Apruzzese et al. highlights the importance of cross-evaluation to determine the reliability of the results obtained using the flow-based IDS and proposes a framework to conduct such reliable cross-evaluations (Apruzzese et al., 2022). These studies demonstrate the validity and importance of the problem as a research direction.

However, it is worth noting that the problem of testing across datasets is not trivial because there are many factors that can cause the reduction in performance, such as protocol, device, and labeling differences. This places our problem at a different level because we introduce a new type of realism test using the same dataset family through the proposed method of cross-capture evaluation per capture. This is because we want to determine the reliability of the IDS when it is deployed with the capability to detect the same family of attacks that it has seen during training in a different capture file.

Attack-centered studies help to illustrate the challenge associated with this problem domain. Flood-style DoS and DDoS attacks often include changing timing and burst characteristics, but their rates, background levels, and device behaviors can differ from one capture file to another. More recent frameworks utilizing deep learning for wireless IoT sensor networks and IIoT environments report favorable attack detection results, yet often fail to report how well the same attack type is recognized in different capture sessions (Nandanwar & Katarya, 2024; Phalaagae et al., 2025). The above aspects directly relate to the attack classes used in this study: DoS-UDP and DDoS-ICMP represent flooding-based DoS/DDoS attacks, which can be difficult to differentiate when timing or packet size characteristics change. Mirai-family botnet communications can also change from one infected endpoint or background loading scenario to another.

A related research avenue discusses this problem domain in terms of concept drift, which is characterized by changes in communication patterns over time even in the absence of DoS or DDoS attacks. Research on drift-aware intrusion detection systems (IDS) concludes that a model learned from a single capture file can degrade in reliability when conditions change over time in a deployment scenario, advocating for adaptability in recognizing such changes (Seth et al., 2024), which directly relates to our rationale for cross-capture evaluation in a manner similar to concept drift in a deployment scenario.

In addition to the accuracy, interpretability has gained prominence in recent years because decision-making in monitoring systems needs interpretability. Surveys show that modern research in IDS systems is moving in the direction of machine learning (ML) and deep learning (DL). It has cautioned the need for more robust evaluation practices in real-world deployments because of the variability involved in it (Hozouri et al., 2025). In smart city and sustainability-based deployments, privacy and trust are also critical factors in addition to interpretability and are being addressed through the use of explainable IDS systems such as federated learning and feature selection (Farooq et al., 2025). Cross-domain explainable IDS work further supports the idea that real IoT environments vary and that models should be evaluated beyond a single capture setting (Alabbadi & Bajaber, 2025).

One notable practical limitation is common in most literature, where many studies are able to identify the features that are important in a given test setting, while few studies are able to identify whether the features that are important in a given setting are also important in other capture settings. This study is able to bridge this gap in knowledge, where we are able to assess the robustness using capture-wise cross-capture tests and also identify stable feature patterns that are important in both directions, which are then used in this study as a form of attack fingerprint. This is because, by using this form of link between interpretability and robustness, we are able to identify features that are important in both directions, meaning that this is a form of evidence that is applicable in the context of sustainability, where the cost of a failed attack is not just security-related, but also includes data reliability issues that affect energy, climate, and infrastructure decisions.

3. DATASET AND PREPROCESSING

We use the CIC IoT IDAD dataset's flow-based subset. In this dataset, each entry is a network flow, which is a compact representation of the network traffic statistics. It does not require the analysis of the payload. It provides counts of the number of packets and bytes transmitted, statistics on the packet sizes, statistics on the inter-arrival times, and counts of the number of TCP flags. It is a lightweight and efficient representation of the network traffic and is appropriate for use in intrusion detection systems.

The dataset comes in the format of several CSV files. Each file corresponds to a specific capture session and type of traffic. It includes several types of attacks and several capture files for several types of attacks. However, for the purpose of this research, we are using

four classes of traffic that are commonly discussed in the context of IoT intrusion detection and are well-represented in the dataset with enough data: Benign, DoS UDP, DDoS ICMP, and Mirai Greeth. We are keeping the original class name Mirai Greeth as it is because this is the class in the network flow.

Our goal is more than just reporting performance when training and testing with the same capture environment. The study verifies if the model still maintains its reliability when the capture environment is different. For this purpose, we will select one capture file for each of the classes for group A and one for each of the classes for group B. Training with group A and testing with group B (A → B) verifies the model's ability to resist changing conditions, and vice versa: training with group B and testing with group A (B → A). Although the training and test sets are different, they are still derived from different capture files; thus, this approach will help verify if the model is learning the attack behavior or is dependent on a certain capture.

Here are the selected files and their flow counts.

Group A:
- Benign: BenignTraffic.pcap_Flow.csv, 183,630 rows
- DoS UDP: DoS-UDP_Flood14.pcap_Flow.csv, 64,565 rows
- DDoS-ICMP: DDoS-ICMP_Flood1.pcap_Flow.csv, 8,263 rows
- Mirai Greeth: Mirai-greeth_flood1.pcap_Flow.csv, 10,856 rows

Group B:
- Benign: BenignTraffic2.pcap_Flow.csv, 91,279 rows
- DoS UDP: DoS-UDP_Flood16.pcap_Flow.csv, 35,173 rows
- DDoS-ICMP: DDoS-ICMP_Flood.pcap_Flow.csv, 10,415 rows
- Mirai Greeth: Mirai-greeth_flood4.pcap_Flow.csv, 9,012 rows

Each of the CSV files contains some identifying information as well as some numeric flow statistics. To study the dataset in a more generic manner, we are removing the identifying information from the flow data. Thus, we are removing the following columns: Flow ID, Src IP, Src Port, Dst IP, Dst Port, Protocol, Timestamp, and Label. After removing the identifying information, we are left with 76 flow statistics for training and testing purposes.

We then clean and prepare the data to ensure all input values are consistent across all files. Infinite values are replaced with missing values, and missing values are imputed based on the median of each feature. All features are converted to a numeric data type to ensure consistency in parsing the data across all capture files. For Logistic Regression, features are standardized based on the StandardScaler because flow features are of different scales, including counts, bytes, and time-based statistics. Random Forest uses unscaled, raw features.

The raw data consists of class imbalance, where benign traffic dominates compared to some of the attack classes. In the random split experiment, a balanced dataset was created by randomly sampling 18,000 flows per class from the combined selected files, resulting in a total of 72,000 flows across the four classes. This helps to reduce bias toward the majority classes, ensuring that Macro-F1 score is a good measure of the model's performance.

For the cross-capture experiment, only classes A and B are used as per the predefined capture groups. Since each direction of the cross-capture experiment consists of disjoint capture files, the number of flows varies per class per group. To ensure all classes contribute equally to the final score, both Macro-F1 score and Accuracy are used as evaluation metrics.

4. METHODOLOGY

We assume that the problem of intrusion detection is a supervised classification problem of four classes, which is defined by flow features in network flows. Each flow is associated with a feature vector x of 76 numeric features after preprocessing. The label y of each flow is classified as either Benign, DoS-UDP, DDoS-ICMP, or Mirai Greeth. The goal of this paper is to train a classifier that performs well on all datasets, regardless of whether training and testing data come from different capture files.

To keep things simple and deployable in real-world scenarios, we explore two effective solutions to the problem of tabular intrusion detection:

- Logistic Regression: A simple and interpretable baseline classifier for this problem. Due to the different scales of flow feature values, we use StandardScaler to normalize all feature values.
- Random Forest: A powerful nonlinear classifier for this problem. It is also an effective and popular choice for flow-based network intrusion detection systems, which are known to be robust to moderate levels of noise in tabular data.

An important part of this research is the evaluation method. The results are provided for two cases: random split and cross file. In the former case, the dataset is split into two subsets using the 80/20 ratio and random allocation. This method is in line with the majority of the literature. However, the results obtained in this way are likely to be biased since the training and test datasets are sampled from the same population.
In the cross file case, the change in the working conditions is simulated by splitting the dataset into two

subsets and allocating them to two capture groups. In the A→B case, the model is trained on group A and tested on group B. In the B→A case, the model is trained on group B and tested on group A. The results are provided for both cases and the average value as well. This configuration is in line with the real-world scenario where the detector is trained in one environment and then tested in the other where the capture properties are different.

Primary metrics: Accuracy and Macro-F1. Accuracy is the measure of the number of correct classifications over the total number of instances. Macro-F1 is the measure where all classes are treated equally. Precision, recall, and F1 are provided for each class separately in order to show the classes where the performance is the worst. We utilize confusion counts to concisely describe the most common confusions between classes in the capture shift. To build trust and provide insights into the learned features, we also report interpretability results and stable attack fingerprints. For the Random Forest classifier, we calculate the permutation importance using the random split test set. The method starts with the baseline test score calculation. Then, we permute the features column-wise and calculate the decrement in the test score. The larger the decrement, the higher the reliance of the classifier on the feature column to make correct predictions. This process is repeated multiple times, and the mean and standard deviation values of the decrement in the test score are reported.

To determine the stable attack fingerprints, we calculate the top twenty important features for A→B and B→A separately and then calculate the overlap between these two sets. The features that appear in the overlap set are the stable attack fingerprints because they remain important even when the two captures are switched. Features that appear in only one direction are likely due to capture-specific information that does not generalize across captures. In addition, we report the rank of the features in both A→B and B→A to show the level of consistency with which the features appear as important indicators.

5. RESULTS AND DISCUSSION

In this section, results are presented for two evaluation paradigms. The first paradigm is random split, in which training data and test data are randomly split from the same pool of capture files. This is arguably the most common evaluation paradigm in intrusion detection systems literature. The second evaluation paradigm is cross file testing, in which training is performed using capture group A and then tested on capture group B, followed by training performed on capture group B and then testing performed on capture group A. The evaluation metrics used are Accuracy and Macro-F1. Accuracy is defined as the number of accurate predictions divided by all predictions made. Macro-F1 is defined as equal weight to all classes, making it suitable for use in this experiment since we have a four-class problem with classes Benign, DoS-UDP, DDoS-ICMP, and Mirai Greeth. The class designation of Mirai Greeth is preserved as it is in the flow files.

An overview of the performance achieved by Logistic Regression and Random Forest is provided in Table 1.

Table 1. Overall classification performance using Accuracy and Macro-F1. Cross-avg is the average over A→B and B→A

Setting	Model	Accuracy	Macro-F1
Random split	Logistic Regression	0.5069	0.5073
Random split	Random Forest	0.8403	0.8404
Cross A→B	Logistic Regression	0.3161	0.3207
Cross B→A	Logistic Regression	0.4589	0.4565
Cross-avg	Logistic Regression	0.3875	0.3886
Cross A→B	Random Forest	0.3955	0.3778
Cross B→A	Random Forest	0.344	0.3005
Cross-avg	Random Forest	0.3697	0.3391

In the random split scenario, the Macro-F1 achieved by the Random Forest is 0.8404, compared to the Macro-F1 achieved by the Logistic Regression, which is 0.5073. This is also illustrated in Figure 1.

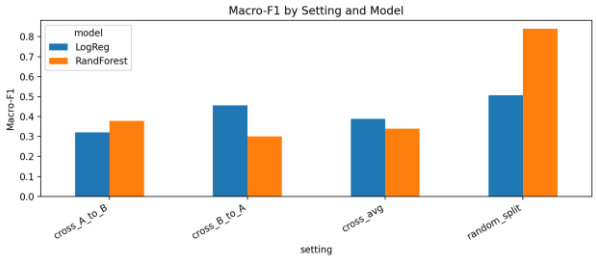


Figure 1. Macro-F1 across evaluation settings for Logistic Regression and Random Forest. Random split yields high performance for Random Forest, while cross-capture testing reveals a clear generalization gap when capture conditions change

The first observation is noted when transitioning from the random split scenario to the cross-capture test scenario. In the cross-capture test scenario, the performance is reduced for the two classifiers. The Macro-F1 achieved by the Random Forest is reduced to 0.3391, which is the average Macro-F1 achieved across the scenarios, while

the Macro-F1 achieved by the Logistic Regression is reduced to 0.3886. This is an important aspect when considering the IoT networks that are sustainability-oriented and may have variable conditions across devices and locations over time, which require the detector to perform reliably.

The cross-capture directions show varying degrees of difficulty for the model. For Random Forest, in the A to B direction, Macro-F1 is 0.3778, whereas in the B to A direction, Macro-F1 is 0.3005. These results show that there are variations in flow statistics for different captures. In other words, the same type of attack may look different in different captures due to varying traffic rate, background traffic, endpoints, or device behavior. To understand which classes are easy or hard for the random split, we also report per-class performance for Random Forest in Table 2.

Table 2. Random split per-class performance for Random Forest

Class	Precision	Recall	F1
Benign	0.8267	0.8656	0.8457
DDoS ICMP	0.8116	0.8353	0.8233
DoS UDP	0.8632	0.8114	0.8365
Mirai Greeth	0.8631	0.8492	0.8561
Macro avg	0.8412	0.8403	0.8404

It is observed that all classes show similar F1 values, which are high for all classes. Hence, it may be concluded that flow features are highly discriminative when training and test traffic are from similar conditions.

To further understand prediction errors for different classes of attacks, we also analyzed prediction errors for different classes of attacks to understand which phenomena are problematic for different captures. When we performed cross-capture testing, we observed that prediction errors show some patterns. A large number of DDoS-ICMP flows are misclassified as Benign or another type of attack in the cross-capture directions. Likewise, Mirai Greeth flows are also frequently classified as DoS-UDP in one direction and as Benign in another direction. These are the reasons for the sharp decline in Macro-F1 performance for cross-capture directions. The model is not only making more errors but also making very specific errors, which are risky in nature. When attack traffic is classified as Benign, it may fail to detect attacks and continue to use collected IoT data.

We support this explanation with the cross-capture confusion counts reported in Tables 3 and 4.

Table 3. Confusion counts for Random Forest under cross-capture A→B. Rows are true classes and columns are predicted classes

	Benign	DDoS ICMP	DoS UDP	Mirai Greeth
Benign	4564	188	578	703
DDoS ICMP	4294	1792	1422	2549
DoS UDP	561	1080	4330	392
Mirai Greeth	2123	348	4271	1422

Table 4. Confusion counts for Random Forest under cross-capture B→A. Rows are true classes and columns are predicted classes

	Benign	DDoS ICMP	DoS UDP	Mirai Greeth
Benign	5569	2503	2897	998
DDoS ICMP	549	1645	5441	308
DoS UDP	2136	2831	6247	423
Mirai Greeth	5739	1859	1464	774

Each row is the true class and each column is the predicted class; the significant values off the diagonal indicate the confusions that occur with the capture shift effect.

To gain some insight into what the model is using, we have employed the permutation importance method with the Random Forest algorithm. The method is quite intuitive; we first calculate the test score. Then we choose one column of the features and randomly permute the values across the test set, leaving the other features unaltered. A significant change in the test score when the values are permuted is indicative that the model relies heavily on that feature; conversely, minimal change indicates that the feature is not very useful. The experiment is repeated multiple times with the average change reported to ensure that the results are stable. The top features reported in the table are presented in Table 5.

Note that many of these features are related to the timing and burst characteristics, such as the minimum forward inter-arrival and the minimum flow inter-arrival. Others are related to the behavior of the connections, such as the initial window bytes and the number of TCP flags. This is intuitive because the type of flooding attacks will tend to disrupt the timing and behavior patterns with respect to benign traffic. The larger the change in the test score, the more useful the feature is in detecting the attacks.

Table 5. Top Random Forest permutation importance features on the random split. Higher means the feature is more important

Feature	Importance	Std
Fwd IAT Min	0.01465	0.0006
Flow IAT Min	0.00858	0.00081
FWD Init Win Bytes	0.00248	0.00026
RST Flag Count	0.00087	0.00003
Bwd IAT Min	0.00071	0.00035
FIN Flag Count	0.00061	0.00023
Bwd Bulk Rate Avg	0.00045	0.00017
Fwd IAT Std	0.00038	0.0002
Packet Length Std	0.00032	0.00023
Active Max	0.00028	0.00009

Finally, we identify the stable attack fingerprints, which persist across different capture variations. The goal is to identify the features that are significant despite the interchange of the capture groups. For this purpose, the top twenty features based on importance are identified for A → B and, separately, for B → A. The intersection of these two sets of features is the set of stable fingerprints, defined as those features where significance is maintained for both directions of analysis. The features unique to one direction are identified as the capture-specific cues with limited transferability. In Table 6, the intersection of the top twenty features for Logistic Regression and Random Forest is shown along with the corresponding feature ranks.

Table 6. Stable fingerprint overlaps across cross directions. Ranks show the A→B rank and the B→A rank within the top twenty feature lists

Model	Stable features with ranks
Logistic Regression	Flow IAT Max 4 and 14; Average Packet Size 6 and 8; Fwd IAT Total 7 and 2; Fwd IAT Std 8 and 9; Fwd Header Length 10 and 1; Bwd Packet Length Std 16 and 19; Idle Min 17 and 3
Random Forest	FWD Init Win Bytes 5 and 1; Total Length of Bwd Packet 11 and 8; Fwd Header Length 18 and 15; Total Length of Fwd Packet 20 and 17

Seven stable features are identified for Logistic Regression, and four stable features are identified for Random Forest, thus validating the robustness results. Although Random Forest is capable of delivering high performance in random splits, a subset of its top features is observed to have reduced stability across different captures. The identified stable features are beneficial for transferability to different IoT scenarios.

6. CONCLUSION AND FUTURE WORK

In this paper, the problem of flow-based intrusion detection in IoT devices is addressed in the context of sustainability-oriented smart infrastructure, in which reliable data streams are critical to the successful deployment of applications such as energy monitoring, environmental monitoring, water management, and intelligent transportation systems. In this paper, four types of traffic were tested using the CIC IoT IDAD dataset: Benign, DoS-UDP, DDoS-ICMP, and Mirai Greeth. Two popular intrusion detection methods were compared and evaluated in this paper, including the scaled Logistic Regression method and the Random Forest method, which is a popular nonlinear method. However, the two methods were also compared and evaluated using two types of evaluation methods: the random train/test split method and the stringent cross-capture protocol method, in which the operating environment changes from A to B and vice versa.

The results show that there is a significant generalization gap between the two methods. The Random Forest method works well with the random train/test split method, with the Macro F1 value being about 0.84; however, the value is significantly reduced when the stringent cross-capture protocol method is used, with the average Macro F1 value being about 0.34 using the Random Forest method and about 0.39 using the scaled Logistic Regression method. The results show that the random train/test split method can be unreliable in the evaluation of the reliability of the intrusion detection system because the operating environment changes over time, devices, and networks. The confusion matrix results show that the errors using the stringent protocol method are systematic rather than random, with the majority of the errors being misclassifications between the attacks and the benign traffic and misclassifications between the different types of attacks.

To improve trust and usability, this study also incorporated interpretability analysis. Permutation importance was used to identify the flow features that contributed most strongly to the Random Forest predictions. In addition, a stability-based attack fingerprint was constructed by identifying the overlap between the top-ranked features in the A→B and B→A cross-capture directions. The results indicate that timing-related and flow/transport-related features provide useful evidence of attack behavior across changing capture conditions. However, the limited overlap among the most

important features, particularly for Random Forest, also shows that some features are capture-specific and may not generalize reliably across different network environments.

There are three directions for future work: First, we will consider more capture files and types of attacks in order to analyze the robustness of the approach on a broader scale. Second, we will consider approaches for distribution shift, such as domain adaptation, calibration, and robust feature selection, in order to reduce cross-capture degradation. Finally, we will consider deployment-scale data splits, such as temporal splits and device-wise splits, where we will consider training and testing on different sets of devices. These approaches will help develop more reliable and trustworthy intrusion detection for IoT systems in climate, energy, and sustainability applications.

References:

Abbasi, F., Naderan, M., & Alavi, S. E. (2021). Anomaly detection in internet of things using feature selection and classification based on logistic regression and artificial neural network on n-baiot dataset. *2021 5th International Conference on Internet of Things and Applications (IoT)*, 1–7. Isfahan, Iran: IEEE. https://doi.org/10.1109/IoT52625.2021.9469605

Alabbadi, A., & Bajaber, F. (2025). X-fuserlstm: A cross-domain explainable intrusion detection framework in iot using the attention-guided dual-path feature fusion and residual lstm. *Sensors*, *25*(12), 3693. https://doi.org/10.3390/s25123693

Apruzzese, G., Pajola, L., & Conti, M. (2022). The cross-evaluation of machine learning-based network intrusion detection systems. *Iris Unimore (University of Modena and Reggio Emilia)*. https://doi.org/10.48550/arxiv.2203.04686

Chen, Z., Chen, D., Zhang, X., Yuan, Z., & Cheng, X. (2022). Learning graph structures with transformer for multivariate time-series anomaly detection in iot. *IEEE Internet of Things Journal*, *9*(12), 9179–9189. https://doi.org/10.1109/JIOT.2021.3100509

Dadkhah, S., Neto, E. C. P., Ferreira, R., Molokwu, R. C., Sadeghi, S., & Ghorbani, A. A. (2024). CICIoMT2024: A benchmark dataset for multi-protocol security assessment in IoMT. *Internet of Things*, *28*, 101351. https://doi.org/10.1016/j.iot.2024.101351

Doménech, J., León, O., Siddiqui, M. S., & Pegueroles, J. (2025). Evaluating and enhancing intrusion detection systems in IoMT: The importance of domain-specific datasets. *Internet of Things*, *32*, 101631. https://doi.org/10.1016/j.iot.2025.101631

Elhadi, S., Marzak, A., Sael, N., & Merzouk, S. (2018). Comparative study of iot protocols. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3186315

Farooq, M. S., Saleem, M., Khan, M. A., Khan, M. F., Siddiqui, S. Y., Aslam, M. S., & Adnan, K. M. (2025). Interpretable federated learning model for cyber intrusion detection in smart cities with privacy-preserving feature selection. *Computers, Materials & Continua*, *85*(3), 5183–5206. https://doi.org/10.32604/cmc.2025.069641

Hozouri, A., Mirzaei, A., & Effatparvar, M. (2025). A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. *Discover Artificial Intelligence*, *5*(1), 314. https://doi.org/10.1007/s44163-025-00578-1

Liu, Y., Wang, H., Zheng, X., & Tian, L. (2023). An efficient framework for unsupervised anomaly detection over edge-assisted internet of things. *ACM Transactions on Sensor Networks*, 3587935. https://doi.org/10.1145/3587935

Meyer-Berg, A., Egert, R., Böck, L., & Mühlhäuser, M. (2020). IoT dataset generation framework for evaluating anomaly detection mechanisms. *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 1–6. Virtual Event Ireland: ACM. https://doi.org/10.1145/3407023.3407036

Nandanwar, H., & Katarya, R. (2024). Deep learning enabled intrusion detection system for Industrial IOT environment. *Expert Systems with Applications*, *249*, 123808. https://doi.org/10.1016/j.eswa.2024.123808

Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., & Ghorbani, A. A. (2023). Ciciot2023: A real-time dataset and benchmark for large-scale attacks in iot environment. *Sensors*, *23*(13), 5941. https://doi.org/10.3390/s23135941

Noura, M., Atiquzzaman, M., & Gaedke, M. (2019). Interoperability in internet of things: Taxonomies and open challenges. *Mobile Networks and Applications*, *24*(3), 796–809. https://doi.org/10.1007/s11036-018-1089-9

Phalaagae, P., Zungeru, A. M., Yahya, A., Sigweni, B., & Rajalakshmi, S. (2025). A hybrid cnn-lstm model with attention mechanism for improved intrusion detection in wireless iot sensor networks. *IEEE Access*, *13*, 57322–57341. https://doi.org/10.1109/ACCESS.2025.3555861

Rabbani, M., Gui, J., Nejati, F., Zhou, Z., Kaniyamattam, A., Mirani, M., … Ghorbani, A. A. (2025). Device identification and anomaly detection in iot environments. *IEEE Internet of Things Journal*, *12*(10), 13625–13643. https://doi.org/10.1109/JIOT.2024.3522863

Sangaraju, V. V., Rao, S. A., Shaik, A., Prasuna, V. G., & Nagaraju, P. B. (2025). A survey on zero trust security – Applications and challenges. *Proceedings on Engineering Sciences*, *7*(1), 437–446. https://doi.org/10.24874/PES07.01D.001

Seth, S., Chahal, K. K., & Singh, G. (2024). Concept drift–based intrusion detection for evolving data stream classification in ids: Approaches and comparative study. *The Computer Journal*, *67*(7), 2529–2547. https://doi.org/10.1093/comjnl/bxae023

Soni, V., Kukreja, D., & Malhotra, A. (2024). Impact of class balancing on intrusion detection system for wsn-bfsf dataset. *SN Computer Science*, *5*(8), 1105. https://doi.org/10.1007/s42979-024-03468-y

Soni, V., Kukreja, D., & Malhotra, A. (2025). Enhancing security of iot enabled smart healthcare clinics using mud. *Journal of Mobile Multimedia*. https://doi.org/10.13052/jmm1550-4646.213417

Walling, S., & Lodh, S. (2024). Network intrusion detection system for IoT security using machine learning and statistical based hybrid feature selection. *Security and Privacy*, *7*(6), e429. https://doi.org/10.1002/spy2.429

Soorya S Suresh
Netaji Subhas University of Technology,
Delhi,
India
soorya.suresh.pg24@nsut.ac.in
ORCID 0009-0002-0125-6369

Pragati Sachdeva
Mtree Software Pvt Ltd,
Noida,
India
pragati.schdv@gmail.com
ORCID 0000-0001-6349-4739

Amarjit Malhotra
Netaji Subhas University of Technology,
Delhi,
India
amarjit.malhotra@nsut.ac.in
ORCID 0000-0003-4169-8022

Deepika Kukreja
Netaji Subhas University of Technology,
Delhi,
India
deepika.kukreja@nsut.ac.in
ORCID 0000-0001-6460-8355