Building an Intrusion Detection Model To Secure IoT Network Traffic

BATUHAN ORHON
n24111251@cs.hacettepe.edu.tr

Abstract

The rapid proliferation of Internet of Things (IoT) devices has introduced unprecedented security challenges, necessitating robust intrusion detection systems (IDS) tailored to resource-constrained environments. In this work, we explore two complementary approaches for AI-driven IDS in IoT networks. First, we replicate the ARP-Probe method proposed by Alani *et al.* (1), reproducing its feature extraction pipeline and model training procedure to match the reported training accuracy. Recognizing that the original study validated the approach on a single, unspecified subset, we extend its evaluation to the publicly available CIC IoT 2023 dataset (2). The baseline ARP-Probe model yields near-random performance (60% accuracy), motivating hyperparameter and architectural refinements that elevate accuracy to 88.51%, representing a key contribution. Second, we address the limitations of packet-level ARP spoofing detection by developing a multi-class flow-level classifier. Utilizing the CIC Flowmeter's 40 features, a baseline MLP fails to surpass 75% accuracy. We then introduce a novel feature extraction workflow based on NFStream, generating a richer flow-feature representation. With the same MLP architecture, this enhanced dataset achieves 94.82% accuracy. Finally, we perform SHAP-based interpretability analysis, presenting per-class and overall feature importance to assess model robustness and potential overfitting.

1 Introduction

The integration of Internet of Things (IoT) devices into critical infrastructure and consumer applications has revolutionized connectivity but also exposed networks to a wide range of security threats. IoT devices often lack sophisticated security mechanisms and operate in heterogeneous environments, making them prime targets for adversaries exploiting vulnerabilities at both the packet and flow levels. Artificial intelligence (AI) techniques, particularly machine learning models, have demonstrated promise in detecting anomalous behavior and classifying network attacks with high accuracy and adaptability. However, IoT-specific constraints—such as limited computational resources and diverse attack scenarios—demand tailored IDS solutions that balance detection performance with practical deployability.

One prominent approach for detecting ARP spoofing in IoT networks is the ARP-Probe framework introduced by Alani *et al.* (1). This method employs packet-level feature extraction from ARP messages followed by a supervised classifier to distinguish legitimate from malicious ARP traffic. We implement the ARP-Probe pipeline end-to-end, replicating its feature extraction and model training steps to confirm the training set accuracy reported in the original work. Noting that validation was limited to an unspecified portion of a single dataset, we evaluate the same model on the CIC IoT 2023 dataset (2). The baseline ARP-Probe model yields only about 50% accuracy on this new dataset, revealing poor generalization. To address this, we perform systematic hyperparameter tuning and minor architectural adjustments, resulting in an improved model that achieves 88.51% accuracy. This extension demonstrates the necessity of cross-dataset validation and highlights our first contribution: an enhanced ARP-Probe variant with robust performance on unseen IoT traffic.

CMP656 Advanced Computer And Network Security, Spring 2025, Hacettepe University, Ankara, TURKEY.

While packet-level analysis is effective for ARP spoofing, many IoT attacks manifest across multi-packet flows and involve diverse protocols beyond ARP. To capture these broader patterns, we develop a multi-class flow-level IDS using the CIC Flowmeter's 40-feature CSV exports. A simple multilayer perceptron (MLP) trained on these features attains at most 75% accuracy, indicating limited expressiveness. We therefore propose a novel feature extraction approach using NFStream, which processes raw `.pcap` files to generate richer flow-based features. Training the same MLP on the NFStream-enhanced dataset yields a substantial accuracy increase to 94.82%, marking our second contribution: a flow-level IDS workflow that leverages detailed network flow characteristics for multi-class attack detection. To ensure model reliability, we compute SHAP values to interpret feature importance, reporting mean SHAP values in Figure 4. This interpretability analysis verifies that the model captures meaningful attack signatures without overfitting.

Finally, to assess domain-shift robustness, we applied our NFStream pipeline and MLP to the BoT-IoT benchmark dataset (3), achieving only 9% accuracy in the 34-class classification task. This result underscores the severity of dataset bias in IDS research and motivates future work on domain-adaptation techniques.

2 Key Concepts

2.1 Intrusion Detection Systems (IDS)

An *Intrusion Detection System* (IDS) monitors network or host activities to identify malicious behavior. IDS approaches are commonly classified as **signature-based**, which detect known attack patterns, and **anomaly-based**, which model normal behavior and flag deviations. In IoT contexts, network-based IDS (NIDS) are favored for their ability to inspect traffic without modifying constrained devices.

2.2 Packet-Level vs. Flow-Level Analysis

Packet-level detection examines individual packets, extracting features such as header fields, flags, and payload statistics. This granularity enables precise detection of protocol-specific attacks (e.g., ARP spoofing) but can incur high overhead in high-bandwidth environments. **Flow-level** analysis aggregates sequences of packets sharing common 5-tuple identifiers (source/destination IP, ports, protocol, timestamps) into *flows*, summarizing them with statistical features (e.g., byte counts, inter-arrival times). Flow-based IDS scale better to large volumes and capture multi-packet attack patterns.

2.3 Supervised, Unsupervised, and Hybrid Learning

Supervised learning trains classifiers on labeled benign and attack data, achieving high accuracy for known threats but requiring comprehensive labelled datasets. **Unsupervised** methods (e.g., clustering, autoencoders) learn normal traffic patterns and detect anomalies without labels, enabling zero-day threat discovery but often suffering from higher false alarms. **Hybrid** frameworks combine both paradigms—e.g., using supervised models for known classes and anomaly detectors for novel variants—to balance detection coverage and adaptability.

2.4 Novelty Detection

Novelty detection refers to identifying previously unseen attack types without retraining. Techniques include embedding into specialized latent spaces that preserve neighborhood structure (e.g., Null Foley–Sammon Transform) and statistical outlier scoring. Effective novelty detection enhances IDS robustness against zero-day threats.

2.5 Explainable AI and SHAP Values

Explainable AI (XAI) techniques aim to make model decisions transparent. SHAP (SHapley Additive exPlanations) values quantify each feature's contribution to individual predictions, facilitating trust and diagnostic insight in security-critical environments.

2.6 Feature Extraction Tools

Popular tools for network flow feature generation include:

- **Wireshark/Tshark**: Packet capture and manual feature extraction for packet-level analysis.
- **CIC Flowmeter**: Generates 40 statistical flow features widely used in IDS benchmarks.
- **NFStream**: Processes raw `.pcap` files to produce richer flow representations (e.g., per-flow time series, protocol statistics), enabling improved detection performance.

2.7 Common Model Architectures

Intrusion detection models range from lightweight **Multilayer Perceptrons (MLP)** to deeper **Convolutional Neural Networks (CNN)** and **Recurrent Neural Networks (RNN)** (e.g., LSTM) for temporal modeling. Ensemble methods (e.g., Random Forest, Gradient Boosting) and hybrid architectures (e.g., CNN+LSTM) further enhance detection accuracy and robustness.

3 Related Work

Intrusion detection in IoT networks has leveraged both packet-level and flow-level data, under supervised and unsupervised paradigms, to address known threats and uncover unknown attack variants.

3.1 ARP-PROBE: An ARP spoofing detector for Internet of Things networks using explainable deep learning

Alani et al. propose **ARP-PROBE**, an explainable deep neural network model designed specifically to detect ARP spoofing attacks in IoT environments (1). The system extracts 21 handcrafted packet-level features using Wireshark and related tools, and trains a lightweight multilayer perceptron (MLP) classifier. The model achieves outstanding results, including 99.98% accuracy and an F_1-score of 0.999, across two distinct ARP spoofing datasets. Furthermore, SHAP (SHapley Additive exPlanations) values are used to provide feature-level interpretability, confirming that the model's decision-making aligns with known protocol behaviors such as IP header anomalies and ARP flag misuse. The authors emphasize the system's suitability for resource-constrained IoT deployments, as it maintains both computational efficiency and high transparency. ARP-PROBE stands out as one of the few studies addressing a specific Layer 2 threat in IoT networks using explainable deep learning, making it a strong candidate for deployment in real-time, trust-sensitive environments.

3.2 Robust detection of unknown DoS/DDoS attacks in IoT networks using a hybrid learning model

Nguyen and Le introduce a **hybrid learning framework** that combines supervised and unsupervised machine learning techniques to detect both known and unknown DoS/DDoS attacks in IoT environments (4). The approach leverages a Soft-Ordering Convolutional Neural Network (SOCNN) for high-precision supervised classification, complemented by unsupervised detectors such as Local Outlier Factor (LOF) and Improved Nearest Neighbour Ensemble (iNNE) to capture novel attack variants. Experiments conducted on CIC-IDS-2017 and BoT-IoT datasets demonstrate an accuracy exceeding 98% for known classes, while achieving over 90% unknown detection rate (UDR) under adversarial settings. This dual-layer architecture improves generalization and robustness against zero-day threats without requiring retraining. The study highlights the importance of integrating anomaly detection into conventional IDS pipelines, making it particularly relevant for the dynamic and evolving threat landscape in IoT networks.

3.3 nNFST: A single-model approach for multiclass novelty detection in network intrusion detection systems

Nguyen and Le propose **nNFST**, a novel deep learning-based framework for multiclass novelty detection tailored to network intrusion detection systems (NIDS) (5). Unlike traditional hybrid models that separate supervised and unsupervised phases, nNFST adopts a single-model design

based on the neighbor-aware Null Foley–Sammon Transform. This technique projects feature vectors into a latent space that preserves local and global neighborhood structures, allowing the model to distinguish both known and unknown attack categories in a unified fashion. Evaluated on the CIC-IDS-2017 dataset, nNFST achieves 92–99% accuracy and 95–99% F_1-scores across multiple known classes, while successfully identifying novel attack types without retraining. Its efficient inference and scalability make it well-suited for real-time IDS deployments. The authors argue that integrating novelty detection directly into the primary classification pipeline enables better robustness against zero-day threats in evolving IoT and network environments.

3.4 A high precision intrusion detection system for network security communication based on multi-scale convolutional neural network

Yu *et al.* propose a high-precision intrusion detection system (IDS) leveraging a **Multi-Scale Convolutional Neural Network (MSCNN)** architecture designed to improve both detection accuracy and convergence speed (6). Their model incorporates multiple convolutional layers with varying kernel sizes to extract features at different spatial resolutions, simulating a human visual-like multiscale recognition mechanism. The architecture also integrates batch normalization and dropout strategies to prevent overfitting and accelerate training.

Evaluated on a dataset of 5 million records, MSCNN achieves a peak accuracy of 98.3%, surpassing traditional Recurrent Neural Networks (RNN) and Adaboost baselines by up to 4.37% in accuracy and 4.02% in false alarm reduction. Notably, the model maintains high detection rates across diverse attack types, including DoS, Probe, U2R, and R2L. Additional experiments demonstrate MSCNN's superior early-warning capability and robustness to increasing attack speed. These results highlight MSCNN's suitability for real-time network security applications, particularly where low latency and high precision are paramount.

3.5 Feature-Optimized Fusion Neural Networks for Intrusion Detection

Wang *et al.* propose a comprehensive IDS framework combining feature selection and neural architecture optimization to enhance detection performance across varying network conditions (7). The method begins with a joint symmetric uncertainty-based feature selection process that evaluates both individual and combined feature relevance using an approximate Markov blanket approach. This step significantly reduces feature dimensionality while preserving class-discriminative information.

The core classification module is a fusion neural network that integrates **Convolutional Neural Networks (CNN)** for spatial feature extraction and **Long Short-Term Memory (LSTM)** layers for modeling temporal dependencies in traffic flows. Furthermore, the architecture is fine-tuned using an improved Particle Swarm Optimization (PSO) algorithm, which optimizes hyperparameters for better generalization.

Tested on benchmark datasets such as KDD99, UNSW-NB15, and CIC-IDS-2017, the system consistently exceeds 98% accuracy. It also demonstrates strong adaptability to different attack types and environments due to its data-driven feature selection and automated tuning. The study illustrates the value of joint optimization in both feature space and model space for constructing scalable and robust IDS solutions.

3.6 Fed-ANIDS: Federated Learning for Anomaly-Based Network Intrusion Detection Systems

Idrissi *et al.* propose Fed-ANIDS (8), a distributed anomaly-based NIDS that combines autoencoder-based intrusion scoring with federated learning to preserve client data privacy. Local clients first preprocess raw PCAPs using CICFlowMeter to extract 87 statistical flow features, then train one of three autoencoder variants—Simple AE, Variational AE (VAE), or Adversarial AE (AAE)—solely on benign traffic. After local training, model weights are shared with a central server, which aggregates them using both FedAvg and the heterogeneous-aware FedProx algorithm, with FedProx yielding superior stability and accuracy under non-IID data distributions.

Comprehensive experiments on USTC-TFC2016, CIC-IDS2017, and CSE-CIC-IDS2018 demonstrate that Fed-ANIDS consistently achieves high detection performance—up to 99.95 % accuracy and 99.94% F1-score with AAE—and low false discovery rates (as low as 0.18 %). Autoencoder-based

Fed-ANIDS outperforms GAN-based federated baselines (e.g., FEDGAN-IDS) and matches or even exceeds centralized learning benchmarks, illustrating the viability of lightweight, privacy-preserving federated autoencoder approaches for scalable intrusion detection in heterogeneous IoT and edge environments

4 Challenges and Open Issues

Despite significant advances in the field of intrusion detection systems (IDS) for IoT networks, several critical challenges remain open, requiring further investigation to enhance security measures.

4.1 Detection of Unknown Attacks

A major unresolved issue in IDS is the effective detection of unknown or zero-day attacks, as current models often depend heavily on previously identified attack signatures. Hybrid models combining supervised and unsupervised learning methods have shown promise in addressing this gap but still require refinement to consistently detect novel threats (4; 5).

4.2 High False Alarm Rates

Intrusion detection systems commonly struggle with high false positive and negative rates, reducing their practical reliability. Ensemble methods, such as combining bagging with gradient boosting decision trees, have improved accuracy but have yet to fully solve the issue of false alarms (9).

4.3 Explainability and Interpretability

Explainable AI (XAI) methods have gained attention for enhancing the transparency of IDS, aiding stakeholders in understanding model decisions. Approaches like SHAP values have been applied successfully to specific attacks such as ARP spoofing, but expanding these techniques to broader multi-class intrusion detection scenarios remains challenging (1).

4.4 Computational Efficiency

Advanced deep learning approaches, while accurate, often entail significant computational overhead. Techniques like multi-scale convolutional neural networks and deep residual convolutional neural networks show promise but require optimization to be suitable for resource-constrained IoT devices (6; 10).

4.5 Generalizability Across Different Datasets

Many intrusion detection methods perform inconsistently across varying datasets, indicating limited generalizability. Approaches using hyperparameter optimization and federated learning attempt to address this, yet achieving reliable cross-dataset performance is still an active research area (11; 8).

4.6 Privacy Concerns

Centralized IDS architectures raise significant privacy concerns. Federated learning has emerged as a solution to preserve data privacy while enabling collaborative training across distributed networks. Nonetheless, federated IDS approaches must still overcome challenges such as communication overhead and secure aggregation of models (8).

4.7 Robustness Against Adversarial Attacks

Recent studies highlight that IDS based on machine learning models can be vulnerable to adversarial attacks, wherein minor perturbations can bypass security measures. Developing models resilient to such adversarial attacks remains a significant challenge, demanding further investigation and innovation (12; 13; 4). These open issues collectively underline the complexity and ongoing nature of research in IoT intrusion detection systems, emphasizing the need for continued development of robust, efficient, explainable, and privacy-preserving detection methods.

5 The Approach

5.1 ARP-Probe Implementation

List of selected features at the packet level.

Feature	Description
frame.len	Frame length in bytes
ip.proto	Protocol name
ip.len	IP packet length in bytes
ip.ttl	TTL field in IP packet
ip.flags	Flags within IP header
ip.hdr_len	IP header length
arp	A field identifying whether a frame is an ARP frame
tcp.flags.syn	SYN flag in TCP header
tcp.flags.ack	ACK flag in TCP header
tcp.flags.reset	RESET flag in TCP header
tcp.window_size	Window size in TCP header
icmp	A field identifying whether a packet is an ICMP packet
tcp.checksum.status	TCP header checksum status
tcp.dstport	TCP destination port number
tcp.srcport	TCP source port number
tcp.flags	TCP header flags
tcp.len	TCP segment length
tcp.time_delta	TCP inter-segment time spacing
tcp.urgent_pointer	TCP urgent pointer
udp.srcport	UDP source port
udp.dstport	UDP destination port

Figure 1: Extracted feature set (1).

As depicted in Figure 1, we first reproduce the ARP-Probe feature extraction pipeline (1). Raw `.pcap` files from the IoT Network Intrusion Dataset (14) are processed with `tshark` to separate malicious (attacker) and benign packets. Each packet is then pre-processed—removing checksum fields, normalizing header values, and filtering irrelevant metadata—and transformed into the same set of 21 handcrafted features defined in the original ARP-Probe study. This ensures an apples-to-apples comparison of model performance before we introduce our subsequent hyperparameter and architectural enhancements.

As shown in Figure 2, we modify the original hidden layer dimensions from 24–16–16–8 to 42–16–8–8, adopt the Adam optimizer for improved convergence, and reduce the minimum number of training epochs from 25 to 10. These adjustments effectively mitigate overfitting while maintaining high detection performance. Newly proposed model has an accuracy of 99.98% on IoT Network Intrusion Dataset and 88.51 % on ARP Spoofing examples of CIC IoT 2023.

5.2 Flow-Level Intrusion Detection

As an independent second approach, we leverage the CIC IoT 2023 dataset, which was collected in the University of New Brunswick's lab environment using 105 heterogeneous IoT devices and simulates 33 distinct attack scenarios plus benign traffic (34 classes in total) as shown in the Table 5.2. This public dataset includes hundreds of gigabytes of raw `.pcap` captures and provides out-of-the-box CSV files with 40 flow-level features extracted via the dpktr.

We trained a deep neural network whose architecture consists of an input layer matching the 40 features, followed by three hidden layers of sizes 128, 64, and 32. Each hidden layer is followed by batch normalization, ReLU activation, and a 30 % dropout, with a final output layer using log-softmax over the 34 classes. Despite extensive tuning, this model failed to exceed 75 % accuracy, and no further improvement was observed in later epochs, indicating that the out-of-the-box feature set lacked sufficient discriminative power for multi-class IoT attack detection.

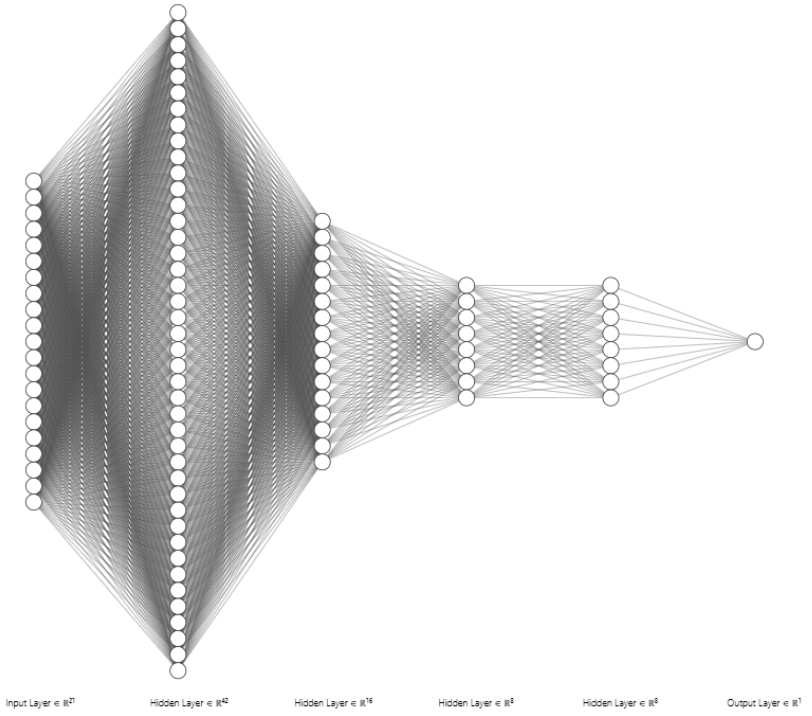

Figure 2: Enhanced ARP-Probe model architecture with modified hidden layers.

Labels	
DDOS-PSHACK_FLOOD	RECON-PORTSCAN
MIRAI-GREIP_FLOOD	DDOS-ACK_FRAGMENTATION
DOS-UDP_FLOOD	DDOS-UDP_FRAGMENTATION
DNS_SPOOFING	RECON-OSSCAN
DDOS-ICMP_FLOOD	BACKDOOR_MALWARE
DDOS-TCP_FLOOD	DOS-HTTP_FLOOD
DDOS-SYN_FLOOD	XSS
DDOS-UDP_FLOOD	DDOS-HTTP_FLOOD
MITM-ARPSPOOFING	BROWSERHIJACKING
DDOS-SYNONYMOUSIP_FLOOD	SQLINJECTION
DOS-TCP_FLOOD	DICTIONARYBRUTEFORCE
VULNERABILITYSCAN	COMMANDINJECTION
DOS-SYN_FLOOD	RECON-PINGSWEEP
DDOS-RSTFINFLOOD	UPLOADING_ATTACK
BENIGN	MIRAI-UDPPLAIN
DDOS-SLOWLORIS	MIRAI-GREETH_FLOOD
DDOS-ICMP_FRAGMENTATION	RECON-HOSTDISCOVERY

Table 1: Attack and benign class labels used in the multi-class IDS

5.3 Flow-Level Intrusion Detection with NFStream-Enhanced Features

To overcome the limited discriminative power of the 40 dpkt features, we reprocessed the raw `.pcap` captures from the CIC IoT 2023 dataset using the NFStream library, which generates a richer set of flow-level statistics. Since some attack classes in the public dataset comprised over 50 GB of `.pcap` files, we trimmed each class to at most four files (8 GB total), creating a representative yet computationally manageable subset. NFStream initially produced 87 features; to prevent overfitting, we removed 26 identifier- and timestamp-related attributes:

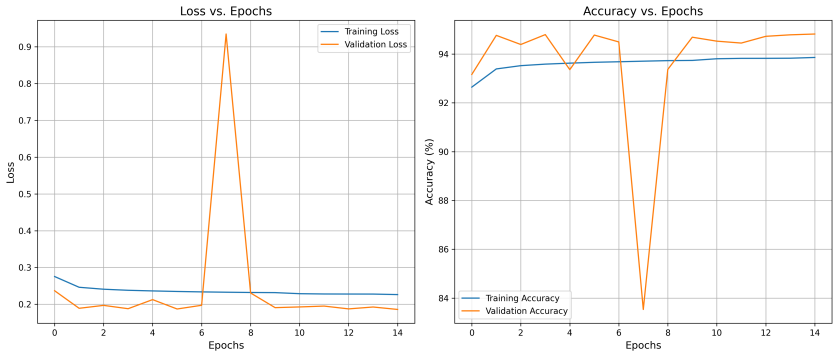

Figure 3: Training and validation accuracy and loss over epochs.

- `id`, `expiration_id`, `src_ip`, `src_mac`, `src_port`, `dst_ip`, `dst_mac`, `dst_port`,
- `src_oui`, `dst_oui`, `ip_version`, `vlan_id`, `tunnel_id`,
- `bidirectional_first_seen_ms`, `bidirectional_last_seen_ms`,
- `src2dst_first_seen_ms`, `src2dst_last_seen_ms`,
- `dst2src_first_seen_ms`, `dst2src_last_seen_ms`,
- `application_name`, `application_is_guessed`, `requested_server_name`,
- `client_fingerprint`, `server_fingerprint`, `user_agent`, `content_type`

After one-hot encoding the remaining categorical features, the final dataset comprised 61 attributes. Training the same MLP architecture on this NFStream-enhanced feature set yielded a substantial accuracy increase to 94.82 %, demonstrating that richer flow-level representations significantly improve multi-class IoT attack detection. The training process is shown on the Figure 3 which hits the maximum accuracy after 15th epoch.

5.4 Evaluation on the BoT-IoT Benchmark Dataset

To further evaluate the robustness and generalization capability of our NFStream-enhanced model, we tested it on the widely used BoT-IoT benchmark dataset (3). We applied the same NFStream feature-extraction pipeline to the raw `.pcap` captures, selecting representative files for the predominant DoS and DDoS attack classes and limiting each class to at most four `.pcap` files (8 GB total) to maintain computational tractability. After identical preprocessing steps—removing identifier and timestamp features and one-hot encoding categoricals—the final BoT-IoT feature set comprised 61 attributes.

Despite the model achieving 94.82 % accuracy on the CIC IoT 2023 dataset, its performance on BoT-IoT was markedly lower, with a maximum accuracy of only 9 % in the 34-class multi-class classification scenario. This result highlights the significant dataset bias and domain shift challenges inherent in IDS research: network flow patterns are highly dependent on the specific environment and simulation settings, making it difficult to produce models that generalize well across heterogeneous IoT domains.

5.5 SHAP Values

As shown in Figure 4, the bar chart presents the mean absolute SHAP values computed across all test samples, indicating which input features the model relies on most heavily to produce its predictions. SHAP values quantify the contribution of each feature to the model's output, enabling a clear interpretation of feature importance.

Importantly, the fact that multiple features exhibit similarly high mean SHAP values suggests that the model does not depend on a single "shortcut" feature to distinguish attack classes. Instead, it integrates information from a broad set of attributes, which reduces the risk of easy or spurious learning and enhances robustness against overfitting.

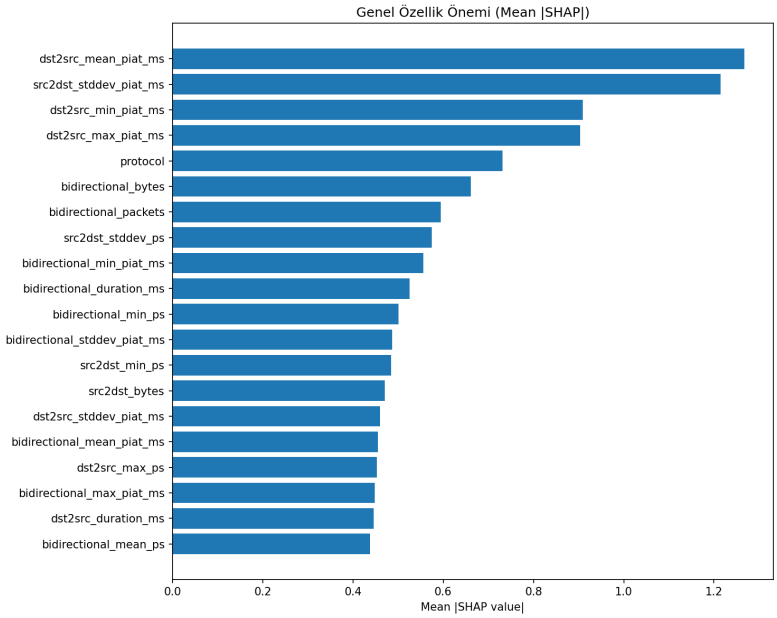

Figure 4: Bar graph of mean SHAP values for each input feature.

Between the original dpkt-generated CSVs and our NFStream-generated CSVs there are several shared columns—protocol types and counts, flag counts, and a handful of others—but NFStream delivers much richer detail. As Table 2 and Table 3 demonstrate, dpkt only provides a single bi_directional or mean IAT feature, whereas NFStream separately tracks src2dst and dst2src inter-arrival times with multiple statistics (min, max, mean, std, etc.). IAT is just one example, but if you examine the SHAP values in Figure 4 you'll see that while bidirectional_mean_piat_ms (dpkt's IAT) has a noticeable effect, features such as dst2src_min_piat_ms, src2dst_stddev_ps, or dst2src_max_piat_ms exert far greater influence. This clearly shows that our more granular NFStream features provide more informative and discriminative insights when comparing attack types.

6 Future Directions

While our work demonstrates the efficacy of NFStream-enhanced flow features and SHAP-guided interpretability, several avenues remain to further strengthen and extend these results:

- **Feature Independence & Enrichment.** Richer flow-level representations easily elevate baseline accuracy from 75% to 95%. We must decouple features from dataset-specific bias through advanced extraction methods (e.g. payload embeddings) and domain-agnostic transformations. Investigating token- or word-embedding techniques on raw packet sequences (e.g., Byte Pair Encoding on payloads or header fields) followed by sequence modeling under real-time constraints can uncover semantic patterns in protocol behaviors and further enrich feature sets. This ensures models learn fundamental attack signatures rather than overfitting to a single environment.
- **Cross-Dataset Benchmarking & Federated Learning.** To validate generalization and preserve privacy, evaluate the proposed pipeline on multiple public IDS benchmarks (e.g., CIC-IDS-2017, BoT-IoT, UNSW-NB15) while also exploring federated learning setups. By training collaboratively across edge sites without sharing raw traffic, we can both measure domain-shift robustness and benefit from a broader, privacy-preserving aggregation of diverse network profiles.
- **Advanced Model Architectures.** Explore more complex classifiers such as Transformer-based sequence models, Graph Neural Networks operating on flow graphs, or hybrid CNN–LSTM ensembles. To make such architectures viable on resource-constrained IoT devices, apply model compression techniques (e.g., pruning, quantization) and lightweight

New flow features generated by NFStream		
application_category_name	application_confidence	application_is_guessed
application_name	bidirectional_ack_packets	bidirectional_bytes
bidirectional_cwr_packets	bidirectional_duration_ms	bidirectional_ece_packets
bidirectional_fin_packets	bidirectional_first_seen_ms	bidirectional_last_seen_ms
bidirectional_max_piat_ms	bidirectional_max_ps	bidirectional_mean_piat_ms
bidirectional_mean_ps	bidirectional_min_piat_ms	bidirectional_min_ps
bidirectional_packets	bidirectional_psh_packets	bidirectional_rst_packets
bidirectional_stddev_piat_ms	bidirectional_stddev_ps	bidirectional_syn_packets
bidirectional_urg_packets	client_fingerprint	content_type
dst2src_ack_packets	dst2src_bytes	dst2src_cwr_packets
dst2src_duration_ms	dst2src_ece_packets	dst2src_fin_packets
dst2src_first_seen_ms	dst2src_last_seen_ms	dst2src_max_piat_ms
dst2src_max_ps	dst2src_mean_piat_ms	dst2src_mean_ps
dst2src_min_piat_ms	dst2src_min_ps	dst2src_packets
dst2src_psh_packets	dst2src_rst_packets	dst2src_stddev_piat_ms
dst2src_stddev_ps	dst2src_syn_packets	dst2src_urg_packets
dst_ip	dst_mac	dst_oui
dst_port	expiration_id	id
ip_version	label	protocol
requested_server_name	server_fingerprint	src2dst_ack_packets
src2dst_bytes	src2dst_cwr_packets	src2dst_duration_ms
src2dst_ece_packets	src2dst_fin_packets	src2dst_first_seen_ms
src2dst_last_seen_ms	src2dst_max_piat_ms	src2dst_max_ps
src2dst_mean_piat_ms	src2dst_mean_ps	src2dst_min_piat_ms
src2dst_min_ps	src2dst_packets	src2dst_psh_packets
src2dst_rst_packets	src2dst_stddev_piat_ms	src2dst_stddev_ps
src2dst_syn_packets	src2dst_urg_packets	src_ip
src_mac	src_oui	src_port
tunnel_id	user_agent	vlan_id

Table 2: Flow-level features generated by NFStream. Those features are generated by us on a subset of the CIC IoT 2023 dataset, and used to train the second multi-class intrusion detection model.

Flow features generated by dpkt in the original CIC IoT 2023	
ARP	AVG
DHCP	DNS
HTTP	HTTPS
Header_Length	IAT
ICMP	IGMP
IPv	IRC
LLC	Label
Max	Min
Number	Protocol_Type
Rate	SMTP
SSH	Std
TCP	Telnet
Time_To_Live	Tot_size
Tot_sum	UDP
Variance	ack_count
ack_flag_number	cwr_flag_number
ece_flag_number	fin_count
fin_flag_number	psh_flag_number
rst_count	rst_flag_number
syn_count	syn_flag_number

Table 3: The list of columns of .csv files provided out of the box in the CIC IoT 2023 dataset. The first multi-class model that couldn't exceed 75% accuracy was trained using these features.

inference optimizations, ensuring that more complex models can be reduced in size and accelerated without compromising detection accuracy.

- **Ablation Studies for Robustness.** Perform systematic ablation by removing the top k highest-SHAP features and re-training the model. A stable accuracy under these perturbations will confirm that the classifier does not rely on a small "shortcut" subset.

- **Concept Drift and Online Adaptation.** Deploy online or continual-learning strategies to adapt to evolving IoT traffic and novel attack techniques.

7 Conclusion

In this work, we presented a comprehensive study of AI-driven intrusion detection for IoT networks, addressing both packet-level and flow-level attack detection. First, we faithfully reproduced the ARP-Probe methodology (1), validated its performance on the CIC IoT 2023 dataset (2), and introduced targeted architecture and hyperparameter modifications that improved generalization accuracy from near-random (50 %) to 88.51 %. Second, we developed a multi-class flow-level IDS by reprocessing raw `.pcap` data with NFStream, curated a representative subset of the data, and demonstrated that our enhanced 61-feature representation boosts MLP accuracy to 94.82 %.

Next, to assess true cross-domain robustness, we applied the same NFStream pipeline and MLP to the BoT-IoT benchmark dataset (3). The model achieved only 9 % accuracy in the 34-class classification task, underscoring the severity of dataset bias and the need for domain-agnostic feature extraction. These results motivate future work on federated learning and advanced, privacy-preserving workflows to train more generalizable IDS models across heterogeneous IoT environments.

To ensure transparency and robustness, we employed SHAP-based interpretability analysis, showing that a diverse set of features—particularly inter-arrival time statistics—collectively drive model decisions and mitigate shortcut learning. Our experiments confirm that richer flow features and careful model design can significantly enhance IoT IDS performance under realistic, multi-device, multi-attack scenarios. Future work will focus on broader benchmark evaluations, advanced sequence models, and adaptive, privacy-preserving deployment strategies for real-world IoT environments.

References

[1] M. M. Alani, A. I. Awad, and E. Barka, "Arp-probe: An arp spoofing detector for internet of things networks using explainable deep learning," *Internet of Things*, vol. 23, p. 100861, 2023.

[2] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "Ciciot2023: A real-time dataset and benchmark for large-scale attacks in iot environment," *Sensors*, vol. 23, no. 13, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/13/5941

[3] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[4] X.-H. Nguyen and K.-H. Le, "Robust detection of unknown dos/ddos attacks in iot networks using a hybrid learning model," *Internet of Things*, vol. 23, p. 100851, 2023.

[5] ——, "nnfst: A single-model approach for multiclass novelty detection in network intrusion detection systems," *Journal of Network and Computer Applications*, p. 104128, 2025.

[6] J. Yu, X. Ye, and H. Li, "A high precision intrusion detection system for network security communication based on multi-scale convolutional neural network," *Future Generation Computer Systems*, vol. 129, pp. 399–406, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X21004143

[7] Q. Wang, H. Jiang, J. Ren, H. Liu, X. Wang, and B. Zhang, "An intrusion detection algorithm based on joint symmetric uncertainty and hyperparameter optimized fusion neural network," *Expert Systems with Applications*, vol. 244, p. 123014, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423035169

[8] M. J. Idrissi, H. Alami, A. El Mahdaouy, A. El Mekki, S. Oualil, Z. Yartaoui, and I. Berrada, "Fed-anids: Federated learning for anomaly-based network intrusion detection systems," *Expert Systems with Applications*, vol. 234, p. 121000, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423015026

[9] M. H. L. Louk and B. A. Tama, "Dual-ids: A bagging-based gradient boosting decision tree model for network anomaly intrusion detection system," *Expert Systems with Applications*, vol. 213, p. 119030, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422020486

[10] G. Sai Chaitanya Kumar, R. Kiran Kumar, K. Parish Venkata Kumar, N. Raghavendra Sai, and M. Brahmaiah, "Deep residual convolutional neural network: An efficient technique for intrusion detection system," *Expert Systems with Applications*, vol. 238, p. 121912, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423024144

[11] M. S. Korium, M. Saber, A. Beattie, A. Narayanan, S. Sahoo, and P. H. Nardelli, "Intrusion detection system for cyberattacks in the internet of vehicles environment," *Ad Hoc Networks*, vol. 153, p. 103330, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570870523002500

[12] D. Schepers, A. Ranganathan, and M. Vanhoef, "On the robustness of wi-fi deauthentication countermeasures," in *Proceedings of the 15th ACM conference on security and privacy in wireless and mobile networks*, 2022, pp. 245–256.

[13] G. Chatzisofroniou and P. Kotzanikolaou, "Security analysis of the wi-fi easy connect," *International Journal of Information Security*, vol. 24, no. 2, pp. 1–11, 2025.

[14] H. Kang, D. H. Ahn, G. M. Lee, J. D. Yoo, K. H. Park, and H. K. Kim, "Iot network intrusion dataset," 2019. [Online]. Available: https://dx.doi.org/10.21227/q70p-q449