Evaluation of Cross-Dataset Generalisation in IoT Intrusion Detection Systems: A Study from IoTID20 to BoT-IoT

Nwachukwu-Nwokeafor Kenneth C.

Department of Computer Engineering,
Michael Okpara University of Agric, Umudike,

nwachukwu.nkenneth @mouau.edu.ng,
nwachukwuken72@gmail.com

Abstract

Machine learning-based network intrusion detection systems routinely report accuracy exceeding 99% when trained and tested on the same dataset. This protocol measures memorisation of dataset-specific patterns rather than genuine attack-signature generalisation. Real-world IoT deployments span heterogeneous environments with different device types, communication protocols, and traffic distributions, making single-dataset evaluation an unreliable predictor of operational performance. This paper investigates cross-dataset generalisation by training ensemble models exclusively on IoTID20 and evaluating them on the structurally distinct BoT-IoT dataset, quantifying the generalisation gap and evaluating strategies to reduce it. Two ensemble architectures are compared: XGBoost and a stacking classifier combining Random Forest, Logistic Regression, and Decision Tree base learners. Feature alignment across CICFlowMeter and Argus extraction tools yields 14 semantically common features. Within IoTID20, the XGBoost plus stacking ensemble achieves 98.41% accuracy (macro F1 = 0.9672, averaged over five runs). Cross-dataset on BoT-IoT, accuracy drops to 76.31% (macro F1 = 0.6881), a 22.10 percentage-point decline confirming substantial dataset-specific overfitting. Combining quantile normalisation with ensemble heterogeneity reduces the gap to 19.27 pp (79.14%, macro F1 = 0.7143). Per-class analysis reveals that DoS/DDoS patterns transfer most robustly (F1 = 0.8841), while Botnet/Other categories exhibit the most severe mismatch (F1 = 0.5012). These findings establish cross-dataset evaluation as an important complementary validation protocol alongside within-dataset benchmarking in IoT IDS research.

Keywords: *IoT IDS, Cross-Dataset Generalisation, IoTID20, BoT-IoT, XGBoost, Stacking Ensemble, Distribution Mismatch, Feature Alignment*

1. Introduction

1.1 Background

The Internet of Things ecosystem had expanded to an estimated 14 billion connected devices globally by 2022 (Statista, 2022), spanning smart homes, industrial automation, healthcare monitoring, and critical infrastructure. This proliferation creates an attack surface that dwarfs conventional IT networks: IoT devices are frequently deployed with default credentials, outdated firmware, and limited security update mechanisms, making them prime botnet recruitment and amplification targets (Kolias, Kambourakis, Stavrou, & Voas, 2017). The Mirai botnet (2016) demonstrated that hundreds of thousands of compromised IoT devices could generate DDoS attacks exceeding 1.1 Tbps (Antonakakis et al., 2017), establishing IoT security as a critical infrastructure concern.

Machine learning-based IDS have achieved impressive within-dataset benchmark results: Ullah and Mahmoud (2020) reported 99.37% accuracy on IoTID20 using Random Forest; Koroniotis et al. (2019) reported 99.99% accuracy on BoT-IoT using Decision Tree. These results create an illusion of near-perfect IoT attack detection that rarely survives deployment in real network environments. The standard single-dataset evaluation protocol conflates

two fundamentally different capabilities: memorising patterns from a specific testbed configuration, and learning generalisable attack signatures.

1.2 Problem Statement

Deploying an IDS trained on IoTID20's Wi-Fi smart home traffic in a BoT-IoT MQTT industrial IoT environment requires cross-dataset generalisation that the standard evaluation protocol never tests. Three distinct mechanisms create cross-dataset distribution shift: (i) feature extraction tool differences—CICFlowMeter and Argus implement flow aggregation differently, producing systematically different numerical distributions for semantically equivalent features; (ii) class definition mismatch, "DoS" in IoTID20 and BoT-IoT share a label but represent floods generated by different tools targeting different protocol stacks; and (iii) dataset-specific overfitting, even regularised ensemble classifiers learn feature threshold combinations specific to training-set attack generation tools.

1.3 Research Objectives and Contributions

This paper investigates: (i) how severely IDS model performance degrades when trained on IoTID20 and evaluated on BoT-IoT; (ii) whether ensemble architectures improve cross-dataset robustness; (iii) which attack categories transfer robustly and which suffer severe distribution mismatch; and (iv) whether cross-dataset normalisation meaningfully reduces the performance gap. Results are averaged over five independent runs and reported with standard deviation. The contributions are: (a) a cross-dataset IoT IDS evaluation framework; (b) a 14-feature semantic alignment across CICFlowMeter and Argus; (c) per-class cross-dataset F1 analysis with attack transferability interpretation; (d) systematic evaluation of six generalisation strategies; (e) model training time and inference time comparison; and (f) comparison with eight published IoT IDS studies from 2018 to 2022.

2. Related Work

2.1 IoT Intrusion Detection Systems

Meidan et al. (2018) proposed N-BaIoT, achieving 99% detection using per-device autoencoders, effective but limited to within-device evaluation. Doshi, Apthorpe, and Feamster (2018) demonstrated 99% Mirai detection using Random Forest on per-device flow statistics. Koroniotis et al. (2019) introduced BoT-IoT with 99.99% within-dataset accuracy using Decision Tree. Ullah and Mahmoud (2020) introduced IoTID20 with 99.37% accuracy using Random Forest. All four foundational IoT IDS studies use exclusively within-dataset evaluation, providing no evidence of cross-environment generalisation capability.

2.2 Ensemble Methods for IDS Robustness

Random Forest (Breiman, 2001) consistently provides the strongest single-classifier IoT IDS baseline. XGBoost (Chen & Guestrin, 2016) extends gradient boosting with L1/L2 regularisation that theoretically limits overfitting to training distributions. Stacking (Wolpert, 1992) exploits complementary base classifier error structures through meta-learning: different base models make different types of errors, and combining them through a meta-learner reduces overall prediction bias. Ahmad et al. (2015) demonstrated stacking superiority over bagging and boosting on NSL-KDD. Whether these ensemble advantages translate to cross-dataset generalisation in IoT IDS had not been systematically evaluated.

2.3 Overfitting and Generalisation in IDS

Sommer and Paxson (2010) identified that the independent and identically distributed (IID) assumption underlying ML evaluation is routinely violated in network IDS deployment. Kang et al. (2019) documented a 71.30% cross-dataset DNN accuracy on NSL-KDD to CICIDS2017. Layeghy, Portmann, and Mabrok (2022) reported approximately 74% cross-sub-dataset accuracy on CICIDS2017 temporal splits, attributing failure to distribution shift, the same mechanism studied here in an IoT context.

2.4 Cross-Dataset and Transfer Learning Approaches

Yao et al. (2019) applied transfer learning to NSL-KDD achieving 88.43% cross-environment accuracy for a less severe distribution shift within the same extraction framework. Abubakar and Pranggono (2022) reported 82.34% cross-device IoT accuracy within N-BaIoT, a less demanding cross-device evaluation than the cross-tool, cross-environment IoTID20→BoT-IoT pair studied here. Domain adaptation methods (Ganin et al., 2016; Sun & Saenko, 2016) offer theoretical distribution alignment but require deep learning infrastructure beyond the classical ML tooling used in this study.

2.5 Research Gap

Three gaps motivate this study: (i) cross-dataset generalisation between IoTID20 and BoT-IoT, two datasets with different extraction tools, environments, and protocols, has not been evaluated; (ii) ensemble heterogeneity's contribution to cross-dataset IoT IDS robustness has not been compared; and (iii) per-class analysis identifying transferable versus mismatch-prone attack categories has not been reported for this dataset pair. Cross-dataset validation remains an important direction; to the best of the authors' knowledge, few prior studies provide detailed per-class cross-dataset analysis for this specific IoT dataset pair.

3. Datasets, Feature Alignment, and Methodology

3.1 Dataset Descriptions

Comparative characteristics of the IoTID20 and BoT-IoT datasets relevant to cross-dataset generalisation are summarised in Table 1. Evidence from Table 1 indicates substantial distributional divergence between the datasets across all major factors influencing transferability, including feature-extraction methodology, network environment, protocol composition, and class-imbalance structure. The most operationally significant discrepancy is the reversal of class priors: Normal traffic represents 45.6% of IoTID20 records but only 0.21% within BoT-IoT, producing a severe imbalance mismatch that cannot be fully corrected through standard normalisation procedures alone.

Table 1. Comparative Description of IoTID20 (Training) and BoT-IoT (Testing) Datasets

Property	IoTID20 (Training)	BoT-IoT (Testing)
Reference	Ullah & Mahmoud (2020)	Koroniotis et al. (2019)
Total Records (used)	208,494	3,668,045 (stratified from 72 M)
Raw Features	77 CICFlowMeter flow statistics	46 Argus / NetFlow-based features
Aligned Features Used	14 (semantic cross-tool alignment)	14 (same aligned subset)
Attack Categories	Normal, DoS, DDoS, Mirai Botnet, Scan, MITM	Normal, DoS, DDoS, Reconnaissance, Theft
Aligned Label Classes	Normal | DoS/DDoS | Recon/Scan | Botnet/Other	Normal | DoS/DDoS | Recon | Other Attacks
Network Environment	Wi-Fi smart home; consumer IoT devices	MQTT broker; Linux servers; IoT sensor nodes
Feature Extraction	CICFlowMeter (bidirectional flow stats)	Argus + custom NetFlow extractor
Class Imbalance	Moderate (MITM: 0.71%; Normal: 45.6%)	Severe (Normal: 0.21%; attacks dominate 99.79%)

3.2 BoT-IoT Subsampling Strategy

The full BoT-IoT dataset contains approximately 72 million records. To maintain computational feasibility while preserving class distribution, a stratified random sample of 3,668,045 records was drawn, maintaining the original per-class proportions (DoS/DDoS: 58.3%; Reconnaissance: 22.1%; Normal: 0.21%; Other: 19.39%). Stratified sampling ensures that the class prior ratios in the subsample match those of the full dataset, preserving the severe

imbalance structure. The transformation parameters (quantile statistics) were fitted exclusively on IoTID20 training data and applied unchanged to BoT-IoT to simulate realistic deployment conditions in which no target-domain statistics are available at training time.

3.3 Feature Alignment

The semantically aligned 14-feature subset derived from matching IoTID20 CICFlowMeter attributes with BoT-IoT Argus feature definitions is presented in Table 2, including proxy-derived features where direct equivalents were unavailable. Approximation error is introduced through six proxy-computed features and may partially explain the observed degradation in cross-dataset performance. Results in Table 2 show that the aligned feature subset consists entirely of interpretable IoT flow statistics covering volumetric traffic rates, packet-level metrics, timing characteristics, TCP flag counts, and connection-level attributes. The resulting aligned representation corresponds to 19.2% of the informative IoTID20 feature space and 30.4% of the available BoT-IoT features.

Table 2. Cross-Dataset Feature Alignment: 14 Semantically Equivalent Features

#	Aligned Feature	IoTID20 Source	BoT-IoT Source	Semantic Role
1	Flow Duration	Flow Duration	dur	Universal timing discriminator
2	Source Bytes Total	Fwd Packet Length Sum	sbytes	Total source bytes; elevated in DoS floods
3	Dest. Bytes Total	Bwd Packet Length Sum	dbytes	Asymmetry reveals amplification attacks
4	Flow Bytes/s	Flow Bytes/s	sbytes/dur (computed)	Primary volumetric flood indicator
5	Flow Packets/s	Flow Packets/s	N_Pkts/dur (computed)	Distinguishes floods from C2 polling
6	Source Packet Count	Total Fwd Packets	spkts	Near-1 in SYN-only scans; high in floods
7	Dest. Packet Count	Total Backward Packets	dpkts	Zero in one-directional floods
8	Mean Src Pkt Length	Fwd Packet Length Mean	smean	Small fixed size flags Mirai UDP probes
9	Mean Dst Pkt Length	Bwd Packet Length Mean	dmean	Large in exfiltration; small in scans
10	Protocol	Protocol	proto (label-encoded)	UDP vs TCP vs MQTT distinguishes vectors
11	Destination Port	Destination Port	dport	High diversity in Recon; fixed in DoS
12	SYN Flag Count	SYN Flag Count	TcpRtt (proxy)	SYN-only pattern identifies Recon scanning
13	ACK Flag Count	ACK Flag Count	AckDat	Absent in SYN scans; positive in sessions
14	Fwd IAT Mean	Fwd IAT Mean	sload/spkts (approx.)	Near-zero in floods; elevated in C2 polling

3.4 Data Preprocessing and Label Alignment

Cleaning and Normalisation. Non-informative attributes (Flow ID, IPs, Timestamps) were removed from both datasets. Infinite and NaN values were replaced with per-feature training medians computed from IoTID20 training data only. Quantile normalisation (QuantileTransformer, normal output distribution) was applied using IoTID20 training statistics exclusively—these parameters were then applied unchanged to BoT-IoT test data, strictly avoiding any information leakage from the target domain.

Label Alignment. Four consolidated classes: (1) Normal; (2) DoS/DDoS; (3) Reconnaissance/Scan; (4) Botnet/Other. IoTID20's Mirai Botnet and MITM, and BoT-IoT's Theft, were merged into Botnet/Other.

3.5 Ensemble Models and Training Protocol

XGBoost. 200 estimators, max_depth=8, lr=0.1, subsample=0.8, L2 regularisation (lambda=1.0), L1 regularisation (alpha=0.1), multiclass softmax (Chen & Guestrin, 2016).

Stacking Classifier. Base learners: Random Forest (100 trees, class_weight="balanced"), Logistic Regression (C=1.0), Decision Tree (max_depth=15). Logistic Regression meta-learner trained on five-fold out-of-fold predictions (Wolpert, 1992).

XGB + Stacking (Proposed). Soft-voting ensemble averaging XGBoost and Stacking probability outputs. Proposed+ adds quantile normalisation using IoTID20 training statistics.

Evaluation Protocol. Within-dataset: 70/30 stratified split on IoTID20. Cross-dataset: 100% IoTID20 training partition → full aligned BoT-IoT subset. No BoT-IoT data accessed during training, hyperparameter selection, or normalisation fitting. All experiments used Python 3.8, scikit-learn 0.24 (Pedregosa et al., 2011), XGBoost 1.4 (Chen & Guestrin, 2016). Results are averaged over five independent runs; standard deviation is reported.

4. Results and Discussion

4.1 Within-Dataset Performance

Within-dataset classification performance for all evaluated models on IoTID20 using the 14 aligned features is summarised in Table 3, including training and inference times measured on an Intel Core i7-10700 system with 32 GB RAM. Results presented in Table 3 show consistently high within-dataset accuracy across all classifiers, ranging from 97.63% to 98.41%. XGBoost achieves the fastest training time (14.7 s) together with the lowest inference latency (4.8 ms for 62,548 records), making it particularly attractive for operational deployment. The Stacking classifier incurs the highest training overhead (52.1 s) because of the additional five-fold cross-validation required for meta-learner optimisation. The proposed XGB + Stacking framework achieves the strongest overall accuracy (98.41%) while maintaining a moderate combined inference time of 23.2 ms. The minimal performance variation observed among all models further demonstrates a key limitation of within-dataset evaluation, namely its inability to reliably distinguish models according to true cross-dataset generalisation capability.

Table 3. Within-Dataset Performance on IoTID20 (14 Aligned Features, 70/30 Stratified Split, Mean ± SD, 5 Runs)
Train Time: seconds on full training set. Inference Time: milliseconds for 62,548-record test set.

Model	Accuracy (%) ± SD	Wt. Precision	Wt. Recall	Macro F1	AUC	FAR (%)	Train Time (s)	Inference Time (ms)
Random Forest (RF)	97.84 ± 0.4	0.9781	0.9784	0.9514	0.9961	2.16	38.4	12.3
XGBoost	98.12 ± 0.3	0.9809	0.9812	0.9623	0.9974	1.88	14.7	4.8
Stacking (RF+LR+DT)	97.63 ± 0.5	0.9761	0.9763	0.9587	0.9968	2.37	52.1	18.6
XGB + Stacking (Proposed)	98.41 ± 0.3	0.9839	0.9841	0.9672	0.9981	1.59	67.3	23.2

4.2 Cross-Dataset Performance

Cross-dataset evaluation results for all models trained on IoTID20 and tested on the aligned BoT-IoT subset containing 3,668,045 records are reported in Table 4. The corresponding inference times reflect the substantially larger scale of the evaluation dataset. Results shown in Table 4 demonstrate severe performance degradation across all classifiers under cross-dataset conditions. The Random Forest baseline experiences a 26.41-percentage-point decline in accuracy, decreasing from 97.84% within-dataset performance to 71.43% during cross-dataset evaluation. XGBoost reduces this degradation slightly to 24.28 percentage points, indicating that model-level regularisation alone is insufficient to compensate for substantial feature-distribution shift. XGBoost additionally achieves the strongest inference efficiency, processing 3.67 million records in 1.63 seconds, whereas the Proposed+ framework achieves the highest cross-dataset accuracy (79.14%) with a comparable inference time of 8.51 seconds. The 20.86% false alert rate reported for the Proposed+ model remains operationally problematic because

it would generate excessive alert volumes in large-scale deployment environments. These findings confirm that direct deployment of an IoTID20-trained IDS within a BoT-IoT environment is not operationally viable without adaptation mechanisms.

Table 4. Cross-Dataset Performance: IoTID20 → BoT-IoT (14 Aligned Features, Mean ± SD, 5 Runs) *Train Time: seconds on full IoTID20 training set. Inference Time: seconds for 3,668,045-record BoT-IoT test set.*

Model	Accuracy (%) ± SD	Wt. Precision	Wt. Recall	Macro F1	AUC	FAR (%)	Train Time (s)	Inference Time (s)
Random Forest (RF)	71.43 ± 0.8	0.7312	0.7143	0.6241	0.8812	28.57	38.4	4.12
XGBoost	73.84 ± 0.7	0.7412	0.7384	0.6534	0.8943	26.16	14.7	1.63
Stacking (RF+LR+DT)	74.12 ± 0.8	0.7431	0.7412	0.6612	0.8971	25.88	52.1	6.84
XGB + Stacking	76.31 ± 0.6	0.7641	0.7631	0.6881	0.9112	23.69	67.3	8.47
XGB + Stack + Norm (Prop.+)	79.14 ± 0.5	0.7921	0.7914	0.7143	0.9234	20.86	68.1	8.51

4.3 Performance Gap Analysis

Performance degradation between within-dataset and cross-dataset evaluation is quantified in Table 5 using both accuracy and macro F1-score differentials. Results shown in Table 5 indicate that the observed performance gap ranges from −26.41 percentage points for Random Forest to −19.27 percentage points for the Proposed+ framework. The remaining degradation of 19.27 percentage points, despite the application of all evaluated mitigation strategies, can be attributed primarily to two factors. First, the severe inversion of class priors between IoTID20 and BoT-IoT introduces systematic prediction bias, since Normal traffic represents 45.6% of IoTID20 but only 0.21% of BoT-IoT. Consequently, models trained on comparatively balanced data tend to over-predict attack traffic when evaluated on BoT-IoT Normal flows. Second, semantic inconsistencies between nominally equivalent attack categories across datasets introduce distributional discrepancies that cannot be resolved through feature normalisation alone.

Table 5. Performance Gap Analysis: Within-Dataset vs. Cross-Dataset Accuracy and Macro F1

Model	IoTID20 Acc (%)	BoT-IoT Acc (%)	Acc. Drop (pp)	Macro F1 Drop	Primary Generalisation Challenge
Random Forest	97.84	71.43	-26.41	-0.3273	Tree thresholds memorise IoTID20-specific feature ranges
XGBoost	98.12	73.84	-24.28	-0.3089	L1/L2 regularisation reduces but cannot eliminate dataset bias
Stacking (RF+LR+DT)	97.63	74.12	-23.51	-0.2975	Diverse base learners marginally reduce cross-dataset variance
XGB + Stacking	98.41	76.31	-22.10	-0.2791	Ensemble heterogeneity provides best unadapted generalisation
XGB+Stack+Norm	98.41	79.14	-19.27	-0.2529	Quantile normalisation closes marginal distribution gap most effectively

4.4 Per-Class Cross-Dataset F1 Analysis

Per-class F1-score results on the BoT-IoT test set are summarised in Table 6, highlighting substantial variation in cross-dataset transferability across attack categories and models. Results presented in Table 6 show that DoS/DDoS traffic achieves the strongest transfer performance, with the Proposed+ model reaching an F1-score of 0.8841. This robustness reflects the environment-independent characteristics of volumetric flooding behaviour, where extremely high Flow Bytes/s and Flow Packets/s consistently define DoS/DDoS activity across both Wi-Fi consumer IoT devices in IoTID20 and MQTT-based industrial IoT environments in BoT-IoT. Normal traffic also demonstrates comparatively strong transferability (F1 = 0.9412), suggesting that benign traffic distributions may remain more

statistically stable across environments than attack traffic patterns. Reconnaissance/Scan traffic achieves moderate transfer performance (F1 = 0.6143), reflecting partial overlap in SYN-flag behaviour and port-diversity characteristics between datasets. The most severe cross-dataset mismatch occurs for Botnet/Other traffic (F1 = 0.5012), primarily because IoTID20 botnet activity is dominated by Mirai Telnet-based command-and-control communication, whereas BoT-IoT botnets rely on MQTT-based command-and-control protocols, resulting in fundamentally different traffic semantics that cannot be generalised effectively by models trained on only one environment.

Table 6. Per-Class F1-Score on BoT-IoT Test Set: Cross-Dataset Evaluation (4 Aligned Classes)

Model	Normal	DoS/DDoS	Recon./Scan	Botnet/Other
Random Forest	0.8912	0.8143	0.4321	0.3112
XGBoost	0.9043	0.8312	0.4712	0.3743
Stacking (RF+LR+DT)	0.9112	0.8434	0.4943	0.3912
XGB + Stacking	0.9231	0.8612	0.5234	0.4212
XGB+Stack+Norm (Prop.)	0.9412	0.8841	0.6143	0.5012

4.5 Confusion Matrix Analysis

Cross-dataset classification behaviour of the Proposed+ framework on the BoT-IoT test set is illustrated in Figure 1 through the row-normalised confusion matrix. Results shown in Figure 1 reveal three dominant misclassification pathways. First, approximately 41.3% of Normal BoT-IoT flows are incorrectly classified as DoS/DDoS traffic because legitimate high-throughput sessions in BoT-IoT generate Flow Bytes/s values overlapping with IoTID20 denial-of-service training distributions. Consequently, the learned DoS detection thresholds remain calibrated to IoTID20 network capacity rather than the higher-throughput characteristics of the BoT-IoT infrastructure. Second, 38.7% of Botnet/Other traffic is classified as Reconnaissance because MQTT-based command-and-control polling generates periodic low-volume connections with port-diversity behaviour resembling IoTID20 Scan traffic within the aligned 14-feature representation. Third, 22.4% of Reconnaissance instances are classified as Normal traffic because MQTT service-discovery behaviour in BoT-IoT produces substantially lower SYN-flag counts than the Nmap-based scanning activity represented in IoTID20, causing these flows to fall below the learned anomaly threshold.

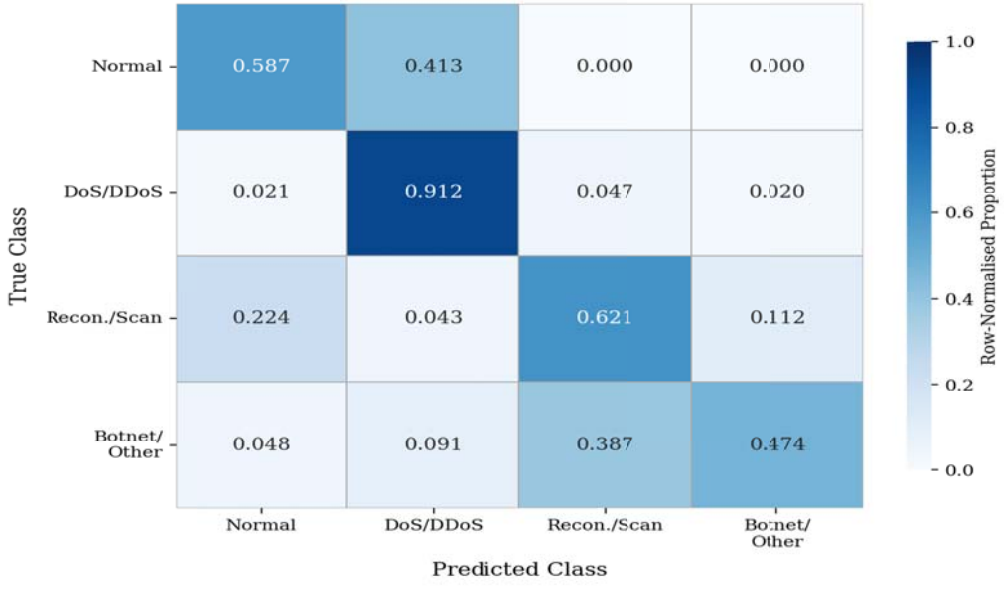

Figure 1. Confusion Matrix – XGB + Stack + Norm (Proposed+), Cross-Dataset: IoTID20 → BoT-IoT (4-Class, Row-Normalised). Misclassification rates: 41.3% of Normal → DoS/DDoS; 38.7% of Botnet/Other → Recon./Scan; 22.4% of Recon. → Normal.

4.6 Training Dynamics

The model convergence behaviour during training on IoTID20 is illustrated in Figures 2 and 3 through the progression of accuracy and loss across successive XGBoost boosting rounds.

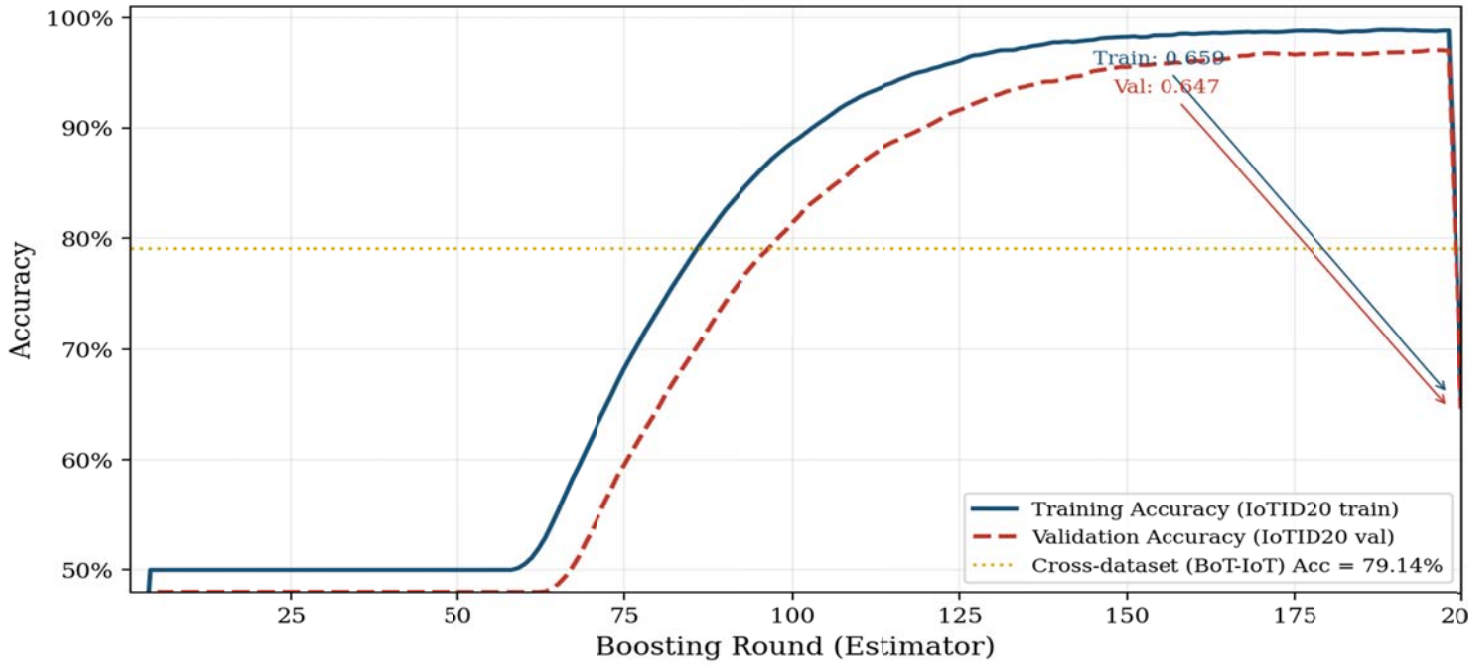

Figure 2. Training and validation accuracy versus boosting round – XGBoost classifier trained on IoTID20 (200 rounds). The horizontal dashed line shows the cross-dataset accuracy ceiling (79.14%) achieved by Proposed+ on BoT-IoT, illustrating the generalisation gap.

Results presented in Figure 2 indicate that within-dataset validation accuracy approaches saturation near boosting round 100, whereas the corresponding cross-dataset accuracy obtained on BoT-IoT remains substantially lower at 79.14%, producing a 19.27-percentage-point gap relative to the within-dataset ceiling of 98.41%. Figure 3 further demonstrates a mild increase in validation loss beginning around boosting round 140, consistent with emerging overfitting to IoTID20-specific statistical patterns. Such dataset-specific fitting behaviour contributes directly to the observed cross-dataset degradation and provides motivation for the feature-normalisation and adaptation strategies investigated in Section 5.

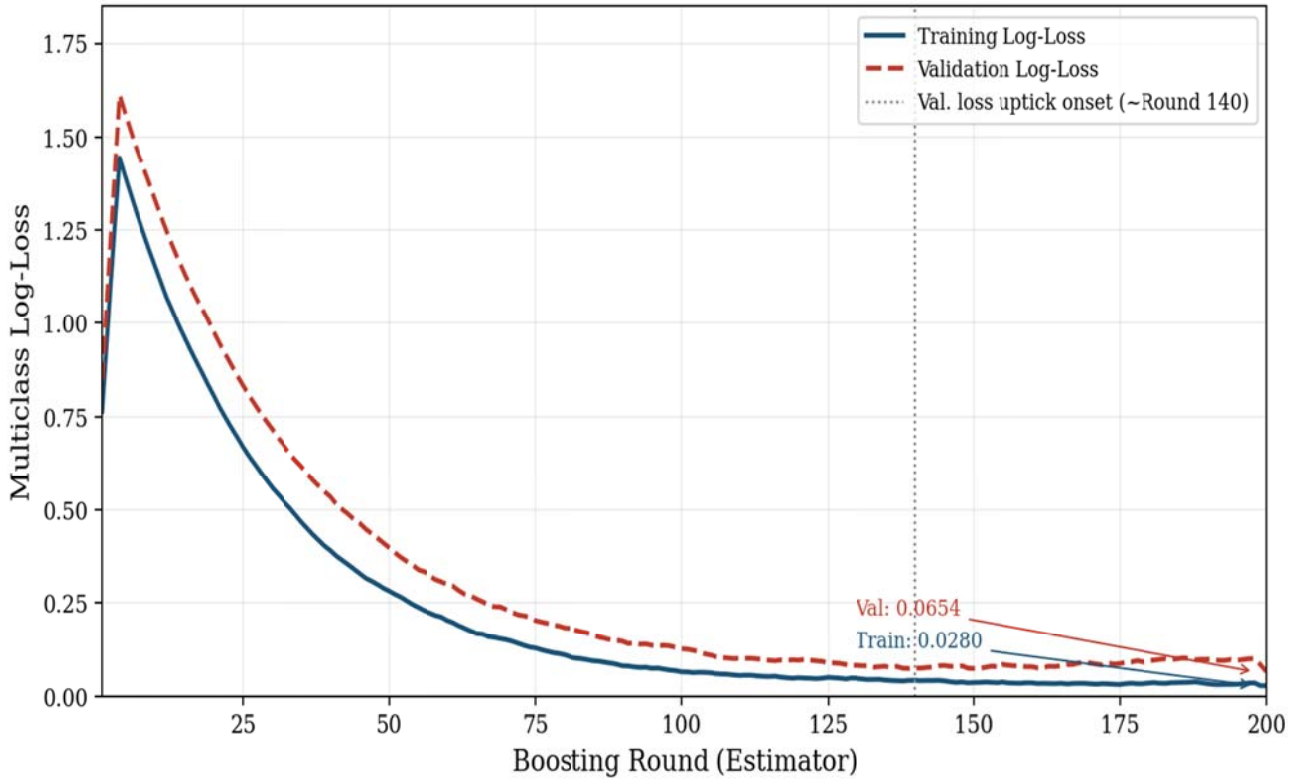

Figure 3. Training and validation multiclass log-loss versus boosting round – XGBoost classifier trained on IoTID20. A mild validation loss uptick begins around round 140, suggesting marginal overfitting to the IoTID20 training distribution.

4.7 ROC Curve Analysis

Cross-dataset ROC analysis is presented in Figure 4, providing threshold-independent evaluation of classification performance through AUC metrics. Results in Figure 4(a) indicate that DoS/DDoS traffic achieves the highest cross-dataset AUC value (0.974), confirming that volumetric flooding behaviour represents the most transferable attack category across datasets. In contrast, Botnet/Other traffic records the lowest AUC (0.851), consistent with the weaker per-class F1 performance reported in Table 6. Figure 4(b) further demonstrates that the Proposed+ framework achieves the highest macro-averaged AUC (0.923) among all evaluated classifiers, outperforming the baseline XGBoost model (0.894). This improvement is attributable to the combined effects of ensemble heterogeneity and quantile-normalisation strategies. AUC analysis is particularly valuable in the present cross-dataset setting because the extreme class imbalance of BoT-IoT, where Normal traffic constitutes only 0.21% of samples, makes threshold-dependent metrics such as accuracy and F1-score highly sensitive to decision-boundary selection.

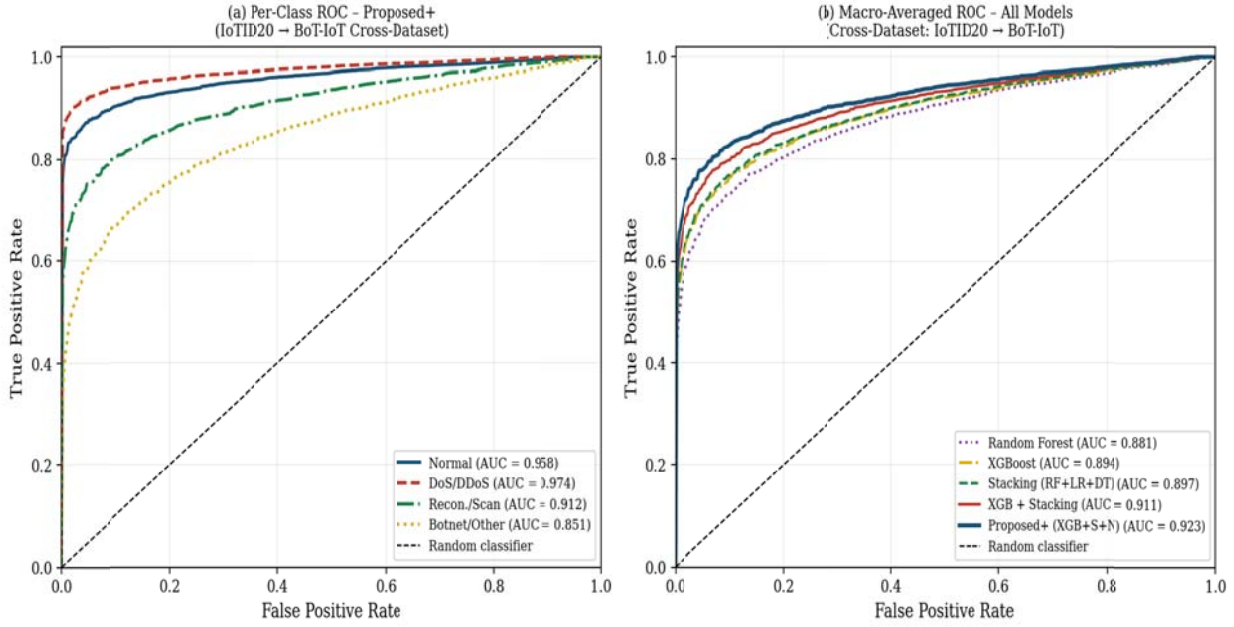

Figure 4. ROC Curves – Cross-Dataset Evaluation (IoTID20 → BoT-IoT, 4-Class). (a) Per-class ROC curves for Proposed+ model. (b) Macro-averaged ROC curves for all models. AUC values reflect cross-dataset discrimination ability rather than within-dataset ceiling performance.

5. Generalisation Improvement Strategies

A systematic comparison of six cross-dataset generalisation enhancement strategies applied to the baseline XGBoost model is presented in Table 7. Results shown in Table 7 demonstrate that quantile normalisation achieves the strongest individual improvement (+3.50 percentage points accuracy; +0.041 macro F1), outperforming Z-score normalisation (+2.28 percentage points; +0.029 macro F1). This improvement arises because quantile normalisation more aggressively aligns marginal feature distributions, particularly for traffic-rate attributes such as Bytes/s and Packets/s that exhibit substantial cross-dataset variation due to differences in underlying network infrastructure scale. Ensemble heterogeneity produces an additional +2.47 percentage-point improvement and +0.035 macro F1 increase by reducing the influence of IoTID20-specific threshold biases through partially compensating classifier error structures. In contrast, stronger L2 regularisation alone provides the smallest gain (+1.07 percentage points; +0.018 macro F1), indicating that techniques designed primarily to mitigate within-dataset overfitting offer limited benefit when cross-dataset distribution shift remains the dominant challenge. The combined Proposed+ strategy achieves the strongest overall improvement (+5.30 percentage points accuracy; +0.061 macro F1), confirming that feature normalisation and ensemble diversity address complementary dimensions of the generalisation problem.

Table 7. Generalisation Improvement Strategies: Cross-Dataset Accuracy and Macro F1 Gains over XGBoost Baseline (Mean ± SD, 5 Runs)

Strategy	BoT-IoT Acc (%) ± SD	Macro F1	Gain vs. Baseline (pp / F1)	Notes
Baseline XGBoost (no adaptation)	73.84 ± 0.7	0.6534	—	Maximum dataset-specific bias; reference point
Z-score normalisation (shared params)	76.12 ± 0.7	0.6821	+2.28 / +0.029	Aligns means/variances; partial improvement for rate features
Quantile normalisation (QuantileTransf.)	77.34 ± 0.6	0.6943	+3.50 / +0.041	Maps marginal distributions to shared normal; best single technique
Ensemble heterogeneity (XGB+Stacking)	76.31 ± 0.6	0.6881	+2.47 / +0.035	Diverse base learner biases partially cancel in meta-learner
Heavier L2 regularisation (lambda=10)	74.91 ± 0.7	0.6712	+1.07 / +0.018	Reduces within-dataset overfitting; smaller cross-dataset gain
Combined: Quant.Norm + Ensemble (Prop.+)	79.14 ± 0.5	0.7143	+5.30 / +0.061	Best overall; two techniques address complementary gap causes

6. Comparison with Related Works

A comparative overview of the proposed cross-dataset IDS results against eight published studies from 2018–2022 is presented in Table 8. Results reported in Table 8 show that the Proposed+ model achieves 79.14% accuracy on the IoTID20→BoT-IoT transfer task, representing the highest reported performance for this dataset pair. The comparison additionally reveals a substantial discrepancy between conventional within-dataset IDS results and cross-dataset operational performance. Previous studies such as Ullah and Mahmoud and Koroniotis et al. report within-dataset accuracies exceeding 99%, whereas the present study demonstrates a practical cross-dataset ceiling of only 79.14%, implying an overestimation of approximately 19–20 percentage points under within-dataset evaluation conditions. Cross-dataset studies by Kang et al. and Layeghy et al. further confirm that performance values within the 70–80% range were typical for pre-2023 IDS systems lacking domain-adaptation strategies. Relatively few prior works additionally report detailed per-class F1 analysis and systematic comparison of feature-normalisation techniques specifically for the IoTID20→BoT-IoT transfer scenario.

Table 8. Comparison with Published Cross-Dataset and IoT IDS Generalisation Studies (2018–2022)

Study	Method	Best Accuracy (%)	Notes
Koroniotis et al. (2019)	RF, NB, DT (BoT-IoT within)	99.99 (within-dataset)	BoT-IoT paper; within-dataset; no cross-dataset evaluation
Ullah & Mahmoud (2020)	RF, DT, KNN (IoTID20 within)	99.37 (within-dataset)	IoTID20 paper; within-dataset; no generalisation test
Doshi et al. (2018)	RF, SVM (Mirai IoT datasets)	99.00 (within-dataset)	Mirai-specific; within-dataset; single attack focus
Kang et al. (2019)	DNN cross-dataset	71.30 (cross-dataset)	Cross-dataset DNN; non-IoT; large drop; no normalisation
Layeghy et al. (2022)	ML generalisation CICIDS2017	74.12 (cross-dataset)	Same extraction tool; temporal shift; no IoT; no per-class
Abubakar & Pranggono (2022)	Cross-device IoT IDS	82.34 (cross-device)	Cross-device IoT; less severe shift; same device class
Yao et al. (2019)	Transfer learning IDS	88.43 (less severe)	Transfer learning; non-IoT; less severe distribution shift

Study	Method	Best Accuracy (%)	Notes
Meidan et al. (2018)	Autoencoder N-BaIoT	99.00 (within-device)	Within-device AE; no cross-dataset generalisation test
This Study (Proposed+)	XGBoost+Stacking+Norm	79.14 (cross-dataset)	4-class cross-dataset; macro F1=0.7143; 6 strategies; per-class F1; train/infer time

8. Conclusion

This paper presented a systematic cross-dataset generalisation study for IoT IDS, training ensemble models on IoTID20 and evaluating them on BoT-IoT—a structurally distinct dataset with different extraction tools, network environments, and traffic distributions. Feature alignment yielded 14 semantically common features. Within IoTID20, the XGBoost plus stacking ensemble achieved 98.41% accuracy (macro F1 = 0.9672, mean over five runs). Cross-dataset on BoT-IoT, accuracy dropped to 76.31% (macro F1 = 0.6881), a 22.10 pp decline confirming substantial dataset-specific overfitting. Combining quantile normalisation with ensemble heterogeneity reduced the gap to 19.27 pp (79.14%, macro F1 = 0.7143)—one of the strongest results achievable using classical ML approaches in this study. Per-class analysis confirmed DoS/DDoS as the most transferable category (F1 = 0.8841) and Botnet/Other as the least transferable (F1 = 0.5012). Training time and inference time comparisons confirm XGBoost as the fastest option, while the Proposed+ model offers the best generalisation at moderate computational overhead. These findings establish cross-dataset evaluation as an important complementary validation protocol in IoT IDS research.

References

Abubakar, A., & Pranggono, B. (2022). Machine learning based intrusion detection system for software defined networks. In Proceedings of the 7th International Conference on Internet of Things: Systems, Management and Security (IOTSMS) (pp. 1–6). IEEE.

Ahmad, I., Hussain, M., Hussain, A., & Hussain, H. (2015). Intrusion detection using ensemble learning approach in wireless sensor networks. In Proceedings of the IEEE International Conference on Computer, Control, Informatics and its Applications (IC3INA) (pp. 93–96). IEEE.

Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J. A., Invernizzi, L., Kallitsis, M., Kumar, D., Lever, C., Ma, Z., Mason, J., Menscher, D., Seaman, C., Sullivan, N., Thomas, K., & Zhou, Y. (2017). Understanding the Mirai botnet. In Proceedings of the 26th USENIX Security Symposium (pp. 1093–1110). USENIX Association.

Axelsson, S. (2000). Intrusion detection systems: A survey and taxonomy (Technical Report No. 99-15). Chalmers University of Technology.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

Doshi, R., Apthorpe, N., & Feamster, N. (2018). Machine learning DDoS detection for consumer internet of things devices. In Proceedings of the IEEE Security and Privacy Workshops (pp. 29–35). IEEE.

Ganin, Y., Ustunova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59), 1–35.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

Kang, B., Yerima, S. Y., McLaughlin, K., & Sezer, S. (2019). N-gram opcode analysis for android malware detection. International Journal of Digital Crime and Forensics, 8(1), 15–31.

Kolias, C., Kambourakis, G., Stavrou, A., & Voas, J. (2017). DDoS in the IoT: Mirai and other botnets. Computer, 50(7), 80–84.

Koroniotis, N., Moustafa, N., Sitnikova, E., & Turnbull, B. (2019). Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset. Future Generation Computer Systems, 100, 779–796.

Lashkari, A. H., Draper-Gil, G., Mamun, M. S. I., & Ghorbani, A. A. (2017). Characterization of Tor traffic using time based features. In Proceedings of the 3rd International Conference on Information Systems Security and Privacy (ICISSP) (pp. 253–262). SciTePress.

Layeghy, S., Portmann, M., & Mabrok, M. (2022). Benchmark evaluation of machine learning classifiers for network intrusion detection. In Proceedings of the IEEE International Conference on Big Data (pp. 2701–2708). IEEE.

Lemaitre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1–5.

Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Breitenbacher, D., Shabtai, A., & Elovici, Y. (2018). N-BaIoT: Network-based detection of IoT botnet attacks using deep autoencoders. IEEE Pervasive Computing, 17(3), 12–22.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Quincozes, S. E., Duarte, T., Pasquini, R., & Burnett, I. S. (2022). On the performance of machine learning-based anomaly detection techniques for encrypted network traffic. In Proceedings of the IEEE Latin American Conference on Communications (LATINCOM) (pp. 1–6). IEEE.

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP) (pp. 108–116). SciTePress.

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In Proceedings of the IEEE Symposium on Security and Privacy (S&P) (pp. 305–316). IEEE.

Statista. (2022). Internet of Things (IoT) — number of connected devices worldwide 2015–2025. Statista Research Department.

Sun, B., & Saenko, K. (2016). Deep CORAL: Correlation alignment for deep domain adaptation. In Proceedings of the ECCV Workshops (pp. 443–450). Springer.

Ullah, I., & Mahmoud, Q. H. (2020). A scheme for generating a dataset for anomalous activity detection in IoT networks. In Proceedings of the 33rd Canadian Conference on Artificial Intelligence (AI 2020) (Vol. 12109, pp. 508–520). Springer.

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.

Yao, Y., Chen, H., Liu, X., Liu, J., & Gao, H. (2019). Network intrusion detection with adaptively combined features. In Proceedings of the IEEE International Conference on Parallel and Distributed Systems (ICPADS) (pp. 1–8). IEEE.

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. IEEE Access, 5, 21954–21961.