Middle East Journal of Humanities and Cultural Studies
Homepage: http://meijournals.com/ar/index.php/mejljs/index
ISSN: 2710-2238 (PRINT)
ISSN: 2788-4686 (ONLINE)
مجلة الشرق الأوسط
للعلوم الإنسانية والثقافية

Machine learning approaches for intrusion Detection in IOT networks

Assistant Lecturer. Ahmed Ridha Khudhur
Al-Nahrain University – Electronic Computer Center

استلام البحث:14-06-2026 مراجعة البحث :13-07-2026 قبول البحث:10-08-2026

Abstract:
Internet of Things (IoT) technologies have introduced a new complexity in the network environment and made it larger, leading to the demand for accurate, robust and interpretable Intrusion Detection System (IDS). This study presents a machine-learning framework for multi-class IoT intrusion detection system (IDS) with the CICIoT2023 dataset. The proposed framework can solve the problems of severe class imbalance, multi-class classification, data leakage and model interpretability. The initial categories in the attack were merged into nine significant classes, then the data was preprocessed, oversampled with the SMOTE technique, and selected using the Random Forest technique. To guarantee reliable model evaluation, feature selection and SMOTE were integrated into a stratified five-fold cross-validation loop, and only performed on the training set of each fold. Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM, CatBoost and HistGradientBoosting were comparatively assessed. In order to select the best model that performs well on the cross validation set, the model with the best Macro F1-scores of 0.9350 ± 0.0100 was chosen, which happened to be XGBoost, followed by Balanced Accuracy of 0.9400. The independent test set yielded 99.40% accuracy, 83.38% Balanced Accuracy, 80.25% Macro F1-score and 99.40% Weighted F1-score for the proposed model. The results of the per class analysis exhibited F1-scores of 1.00 for DDoS, DoS, and Mirai, while the extremely rare Brute Force and Other classes were not as successful, with F1 scores of 0.42 and 0.47, respectively. To make the model as transparent as possible, the SHAP analysis was performed, which identified the characteristics of flow duration and traffic-rate as the most influential features, including Rate, Srate and Drate. The results show that the proposed framework achieves good overall detection rate and presents a leakage-aware and interpretable multi-class IoT intrusion detection solution. The results also demonstrate the ongoing challenge of identifying very rare categories in attacks, suggesting the need for more sophisticated minority class learning strategies.
Keywords: Machine, learning approaches, intrusion, Detection. IOT, networks

1. Introduction:

As the Internet of Things (IoT) technologies are rapidly advancing, they have resulted in highly connected environments in healthcare, smart homes, industry, transportation and other critical sectors, as well as increased the cyber-attack surface (Binbusayyis et al., 2022). The diversity and volatility of IoT devices and traffic demand an accurate and timely intrusion detection more than ever (Saheed et al., 2022). The amount and complexity of the data traveling through the IoT network has spurred the adoption of machine learning (ML) based methods as an alternative to rule-based intrusion detection systems (IDSs) (Binbusayyis et al., 2022). Initial experiments proved the effectiveness of supervised ML for distinguishing malicious and legitimate IoT traffic, but the classification performance is highly dependent on the characteristics of the dataset and the classification settings (Saheed et al., 2022). thereafter, deep learning (DL) was introduced to facilitate automatic learning of complex and nonlinear representations of traffic in IoT intrusion detection and eliminate the need for manually engineered features (Banaamah & Ahmad, 2022). Hence, CNN, RNN, and hybrid architectures have been increasingly adopted to recognize advanced IoT attacks (Awajan, 2023). Although DL models have been shown to be excellent predictors, they may lack

transparency, which has spurred the research of explainable IDSs using feature attribution and ensemble learning (Shtayat et al., 2023). More recent works have grown IDS to distributed and adaptive architectures, providing distributed and federated learning, graph learning, and reinforcement learning (Al Tfaily et al., 2025). These methods are designed to tackle the challenges of decentralized IoT networks and dynamic network scenarios (Khayat et al., 2025).

Despite these, there are significant issues, such as the high level of class imbalance, the lack of uniformity in grading processes, data leakage, low generalizability, and low interpretability of models (Shtayat et al., 2023). The challenges are particularly relevant for CICIoT2023, which includes traffic from 105 IoT devices and 33 attack types (Borra et al., 2025). In this study, the attack classes have been merged to form nine meaningful classes, with DDoS making up 72.9% of the training data, Brute Force and Other making up less than 0.1% each. Accordingly, the following work proposes a multi-class IoT intrusion detection framework which tackles class imbalance, leakage-free evaluation, generalizability and explainable decision-making. As highlighted in figure 1, from traditional ML and DL based IDS to explainable, distributed and adaptive IDS, the evolution of IoT IDS has come a long way. It also outlines the key research gaps covered in this research such as class imbalance, multi-class detection, leakage-free evaluation, generalizability, and explainability.

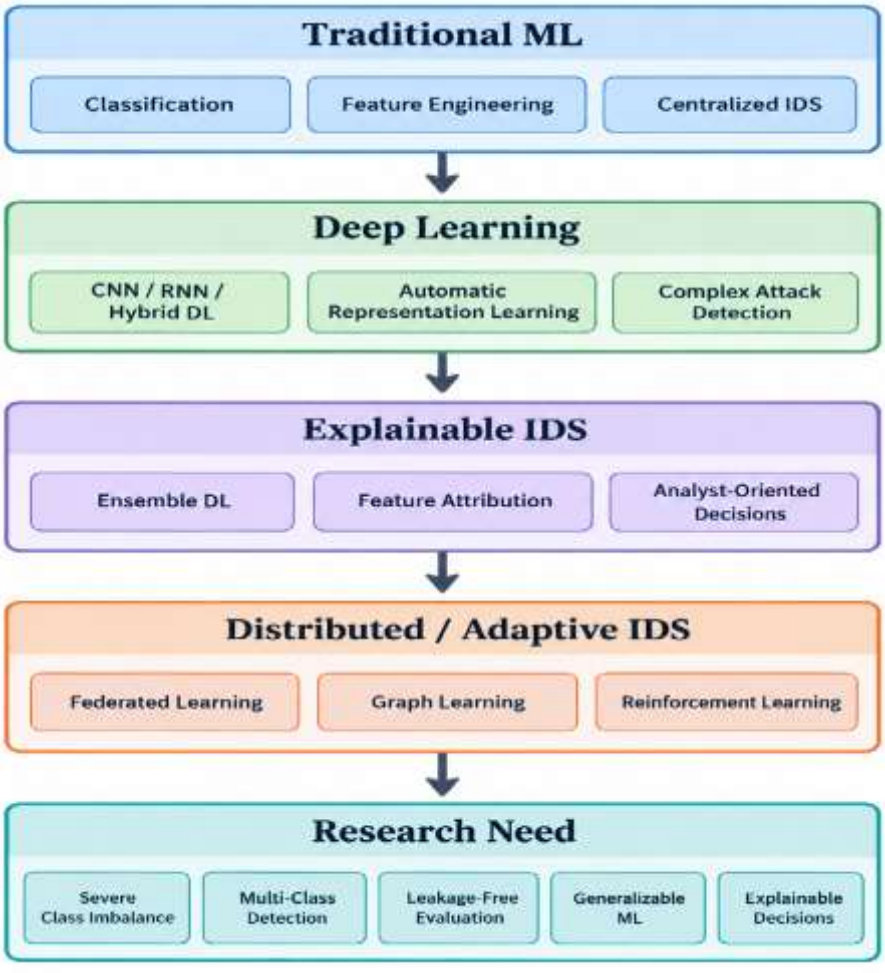

Figure 1. Evolution of IoT intrusion detection from conventional machine learning to explainable, distributed, and adaptive learning approaches. Adapted from Binbusayyis et al. (2022), Banaamah and Ahmad (2022), Shtayat et al. (2023), Al Tfaily et al. (2025), Khayat et al. (2025), and Borra et al. (2025).

This development demonstrates that the discussion on the feasibility of Machine Learning to detect IoT attacks has now shifted to more challenging questions, including robustness, class imbalance, evaluation reliability, interpretability, and generalization. The present study is thus designed not only to compare the performance of conventional classifiers but to create a robust machine-learning framework that incorporates methods for managing imbalance data, feature selection that ensures no leakage, stratified cross-validation, multi-class evaluation methods and finally, post-hoc explainability.

The current study thus explores seven machine learning models—Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM, CatBoost, and HistGradientBoosting. The framework features class aggregation, data cleaning, balanced handling via SMOTE technique, feature selection via Random Forest technique and stratified five-fold cross-validation, and interpretation using SHAP method. Special care has been taken with the methodology to incorporate feature selection and oversampling into the cross-validation loop, thereby minimizing the chances of over-optimistic performance estimates.

This study, for this reason, has four main contributions. First, it builds a Multi-class IoT Intrusion Detection system in a systematic way. Second, it uses a leakage-aware evaluation method where feature selection and oversampling are performed inside the training set of each fold of the cross-validation. Third, It gives a comparative assessment of seven machine learning paradigms while considering the Macro F1-score and Balanced Accuracy over the overall accuracy only. Fourth, it uses explainability derived from the SHAP values to discover the most significant characteristics of the network traffic, which is most relevant to the predictions of the best performing model. The contributions are intended to offer a more suitable, and interpretable, assessment of machine-learning-based intrusion detection in realistic multi-class and imbalanced settings.

I. Literature Review

2.1. Machine Learning for IoT Intrusion Detection

The machine learning approach has emerged as one of the main methodologies for building intrusion detection systems because it can detect statistical signatures that are linked to malicious traffic, without having to specify a particular signature for each attack. The literature suggests a succession of more and more complex learning architectures that evolved from traditional supervised learning.

Binbusayyis et al. (2022) investigated and compared the machine learning methods used for intrusion detection (ID) in IoMT networks. Their study is especially important due to IoMT

being a specific type of IoT environment where a failure in security could impact highly sensitive medical systems and information. The study shows that multiple ML algorithms can be used in IoMT traffic, and that the choice of algorithm can significantly impact intrusion detection performance.

Saheed et al. (2022) also investigated the use of machine learning for the detection of an IoT network attack. Their study validates the use of ML for detecting malicious traffic, and shows that the choice of the appropriate algorithms and input representations is crucial for IoT security applications.

The need for these conventional ML studies is important because they set comparative baselines with which more sophisticated approaches may be evaluated. Their focus is mainly on classification from a predictive perspective and not on the entire methodological challenges of severe class imbalance, leakage-free feature selection, explainability and robust multi-class evaluation.

Table 1. Representative Machine Learning and Deep Learning Approaches for IoT Intrusion Detection

Study	Year	Environment	Main approach	Identified limitation/research implication
Binbusayyis et al.	2022	IoMT	Comparative ML approaches	Specialized IoMT setting; broader IoT generalization remains relevant
Saheed et al.	2022	IoT	Machine learning	Emphasis on predictive detection rather than comprehensive explainability
Banaamah & Ahmad	2022	IoT	Deep learning	Increased model complexity and interpretability challenges
Awajan	2023	IoT	Deep learning	Demonstrates DL capability but leaves broader evaluation issues
Shtayat et al.	2023	IIoT	Explainable ensemble DL	Shows need to combine detection performance with interpretability
Borra et al.	2025	Satellite-integrated IoT	Hybrid DL + optimization	Increasing architectural complexity
Al Tfaily et al.	2025	IoT	Graph-based federated learning	Addresses distributed structure and privacy
Khayat et	2025	IoT	Reinforcement	Moves toward dynamic

al.	learning + deep features	detection

The research on which the presented studies is based has grown in various directions, as indicated in Table 1: algorithmic complexity, explainability, distributed learning, and adaptation. However, this as models become more sophisticated, their basic evaluation issues do not disappear.

2.2. Deep Learning and Explainable Intrusion Detection

Deep learning has been used more and more in the field of intruder detection in the Internet of Things (IoT) due to its capability to learn nonlinear representations from complex network traffic. The general trend is toward automatic learned representations, as in Banaamah and Ahmad (2022), who investigated deep learning for IoT intrusion detection. This further reinforces the use of DL architectures for cybersecurity in IoT networks, and Awajan (2023) proposed a deep-learning-based IDS for such networks.

But, deep learning poses a second research challenge: transparency of the models. Security analysts may need to know why one specific connection/flow has been marked as malicious. An operational validation is not always easy for a model that gives only a prediction, but not an interpretable reason for it. These concerns directly relate to Shtayat et al. (2023), who developed an explainable ensemble deep learning approach for IIoT intrusion detection. They highlight the need for making the security decisions made by deep learning models more transparent and demonstrate how explainable AI is becoming an increasing part of cybersecurity.

This progress implies that whether an intrusion detection model is successful should not be judged solely by the predictive metrics. Furthermore, it should take into account whether the learned decision process can be explained by meaningful network properties. In this context, the use of SHAP in the present study is meant to draw a link between predictive performance and interpretation of the features. Figure 2. In this regard, we aim to classify the machine-learning methods for IoT IDS.In this aspect, we propose to classify the machine-learning approaches for IoT IDS.

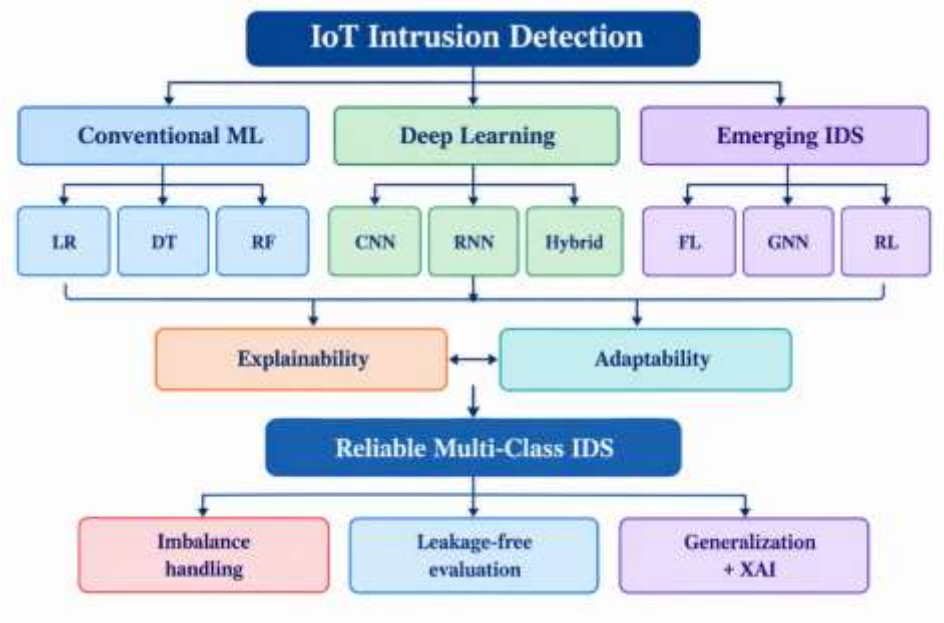

Figure 2. Taxonomy of machine-learning approaches for IoT intrusion detection based on the reviewed literature.

The taxonomy illustrates that the field is no longer dominated by a single modeling paradigm. Conventional ML provides computationally efficient baselines, DL provides representation-learning capabilities, and emerging approaches such as federated learning, graph learning, and reinforcement learning address distributed structure and dynamic adaptation.

2.3. Distributed, Adaptive, and Specialized IoT Intrusion Detection

Recently, intrusion detection has been extended from the centralized classification. Al Tfaily et al. (2025) suggested graph-based federated learning for the IDS of IoT networks due to the limitations of the conventional deep learning and federated learning methods in capturing the temporal behavior and network topology. Their research is showcasing the growing importance of graph structure and distributed learning in complex IoT systems.

Khayat et al. (2025) introduced reinforcement learning with deep features as a dynamic approach to IoT intrusion detection. This is a different approach to static classifiers as they adapt their detection behaviour based on changing network conditions.

Another specific scenario studied is that of satellite-integrated IoT networks, which Borra et al. (2025) proposed a hybrid deep-learning method with optimization mechanisms. Their work underlines that IoT security is now a growing part of diverse architectures for which traditional IDS assumptions might not be adequate.

These studies demonstrate significant progress, but they also introduce additional model complexity. There are multiple architectures for deep learning that can improve adaptability or representational capacity, such as federated, graph-based, reinforcement-learning, and

hybrid architectures, which can also lead to greater computational demands and methodological complexity. In many IoT intrusion detection applications, there is still a demand for machine-learning frameworks that are relatively lightweight and reproducible, and still can yield good performance on multi class discrimination with maintaining interpretable decision mechanisms.

Based on the literature, there are four related problems in IoT intrusion detection which are severe class imbalance, multi-class complexity, evaluation reliability, and interpretability. As outlined in Table 2, conventional ML methods yield efficient and relatively interpretable benchmarks, whilst deep-learning and hybrid models have increased representation-learning power but come with more complex models. Recent studies have also been around the explainable, federated, graph-based and adaptive aspects of IDS research. But these instructions tend to focus on different aspects of the problem instead of on the problem as a whole and under a single assessment structure. To illustrate this, Binbusayyis et al. (2022) tested multiple ML classifiers for IDS in the IoT space and found that the Decision Tree and KNN achieved 100% accuracy on the tested Bot-IoT data while NB, MLP, and SVM attained accuracy of 99.2%, 99.9%, and 99.9% respectively. This result, as shown in Figure 2, shows the effectiveness of conventional ML classifiers, as well as the importance of considering the performance beyond overall accuracy, especially when the attack distribution is highly imbalanced in the intrusion datasets. In this study, the primary goal is to incorporate multi-class detection, imbalanced learning, leakage-free evaluation and explainability in a unified reproducible setting.

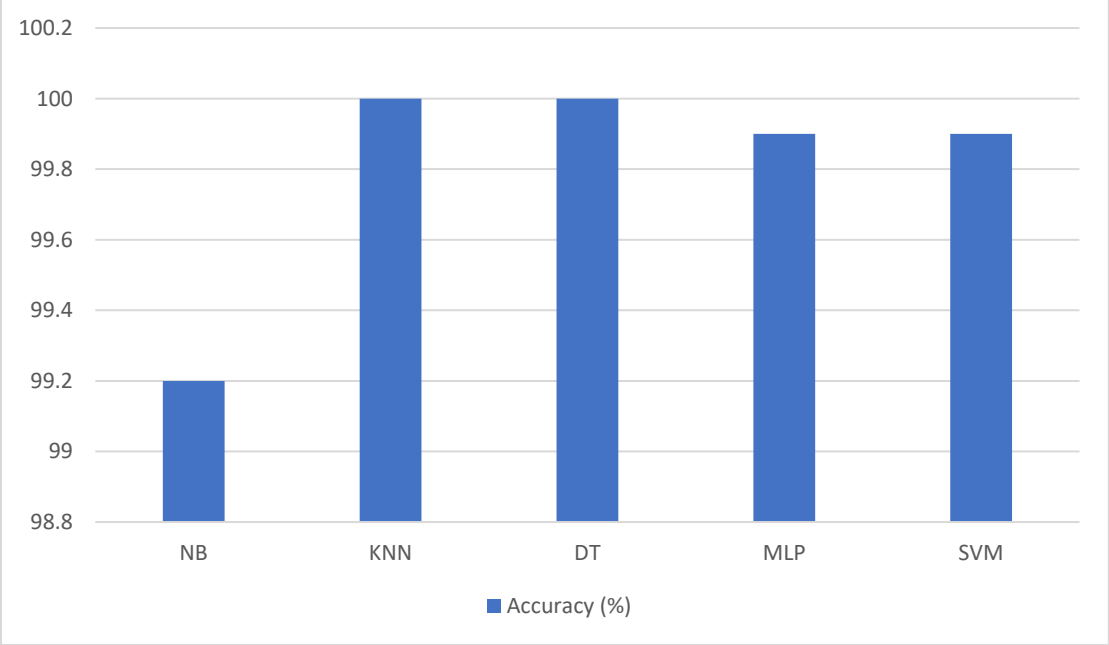

Figure 2. *Classification accuracy of ML models reported by Binbusayyis et al. (2022) for Bot-IoT intrusion detection.*

Table 2. Critical Comparison of Existing IoT Intrusion Detection Research

Representative studies	Main strength	Remaining limitation	Relevance to present study
Binbusayyis et al. (2022); Saheed et al. (2022)	Efficient and comparatively interpretable	Limited treatment of severe imbalance and evaluation reliability	Establishes ML baselines
Banaamah & Ahmad (2022); Awajan (2023)	Learns complex nonlinear representations	Higher complexity and reduced transparency	Motivates comparison with strong ML models
Shtayat et al. (2023)	Improves model transparency	Does not by itself solve imbalance or evaluation problems	Motivates SHAP analysis
Borra et al. (2025)	Combines representation learning and optimization	Increased computational complexity	Highlights the need for efficient alternatives
Al Tfaily et al. (2025)	Supports distributed and structural learning	Greater architectural complexity	Establishes an emerging research direction
Khayat et al. (2025)	Addresses dynamic attack behavior	Greater reinforcement-learning complexity	Indicates potential future extensions
Proposed framework	**Multi-class ML + imbalance handling + leakage-free CV + XAI**	**Rare-class detection remains challenging**	**Integrates methodological rigor and interpretability**

II. Methodology

This section provides the complete methodology that is being followed for intrusion detection in IoT networks with CICIoT2023. The pipeline proposed includes data acquisition, data preprocessing, class aggregation, handling of imbalance data, feature selection, model training, assessment and interpretability analysis. Figure 3 shows the overall methodology workflow of this research.

Figure 3: Overall Methodology Workflow

3.1. Dataset Description and Characteristics

The CICIoT2023 dataset is a publicly available benchmark, which is especially for IoT network intrusion detection. It was derived from an accurate testbed of 105 IoT devices, and

contains 33 different attack types, traditionally categorized into seven groups: DDoS, DoS, Reconnaissance, Web-based, Brute Force, Spoofing, and Mirai botnet. The data set contains 47 network flow features obtained, which include statistical, temporal, and protocol-based attributes.

In this study, the data is split into three sets (training, validation, and testing) to keep the model development and test in a separate manner to get a true picture of the test data. The training and validation sets were combined to use all the data for training, which means that there were 202,221 training samples and 138,383 test samples. Table 3 shows the distribution of classes after the aggregation method described in Section 3.2.

Table 3: Class Distribution After Aggregation (Training Set)

Class	Count	Percentage (%)
DDoS	147,403	72.9
DoS	35,068	17.3
Mirai	11,313	5.6
Benign	4,614	2.3
Spoofing	2,127	1.1
Recon	1,368	0.7
Web-based	254	0.1
Brute Force	59	<0.1
Other	15	<0.1
Total	**202,221**	**100%**

3.2. Data Preprocessing and Class Aggregation

A systematic data preprocessing pipeline was used to obtain good data quality and generalizability of the models: In the original dataset, there are 34 fine-grained attack subcategories, with a few of these having very few samples. When these sparse classes are used directly in SMOTE or trained with classifiers, algorithms tend to fail or to have poor generalization (Chawla et al., 2002). Thus, we combined all the sub-categories into nine major classes: Benign, DDoS, DoS, Mirai, Recon, Spoofing, Web-based, Brute Force and Other. Not only does this mapping help avoid SMOTE failures but it also helps to integrate the classification problem into real-world security taxonomies and makes it easier to interpret the results (Fernandes et al., 2019).

To reduce the dimensionality without sacrificing discriminative information, features with zero variance (equal across all samples) were eliminated.

The validation datasets were appended to train. Datasets to create a larger training set to provide more data for models to learn. Test datasets were maintained entirely separate from all other data and was only used for the final assessment; no information in the test set was used in any way to develop the model or in hyperparameter tuning (Hastie et al., 2009).

3.3. Imbalance Handling and Feature Selection

The high-class imbalance with DDoS class being 2,500 times more present than Brute Force class is a big hurdle. To reduce this, we used only the training data to apply the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002). The idea with SMOTE is to create synthetic minority class samples by performing interpolation between the real samples of a minority class. To prevent the failures that occur when a minority class has less than six samples, which is generally a problem in highly imbalanced datasets (Fernández et al., 2018), this paper used the value of $k_{neighbors} = 3$. This parameter setting allows SMOTE to produce meaningful synthetic samples of the rare classes.

A feature selector (Breiman, 2001) based on the random forest was used to reduce the computation cost, avoid overfitting and improve the model interpretability. The feature importance was calculated using the gini impurity criterion and features with an importance below the median were discarded (Pedregosa et al., 2011). Importantly, this selection process was incorporated within each cross-validation fold (see Section 3.5) to prevent information leakage from the validation folds, which is strongly recommended by Cawley and Talbot (2010) to avoid error inflation.

Seven models representing diverse learning paradigms were implemented and compared to provide a comprehensive benchmark:

These are the baseline models: Logistic Regression (linear classifier); Decision Tree (rule-based, non-linear).

Ensemble Models: Random Forest, XGBoost, LightGBM, CatBoost, and HistGradientBoosting.

Evaluation Protocol

A 5-fold Stratified K-Fold cross validation approach was used on the training data to guarantee strong and unbiased model selection. When the data are imbalanced, stratifying ensures that the original class distribution is maintained in each fold (Kohavi, 1995). For each of the folds, a separate pipeline was used for feature selection and SMOTE oversampling of only the training portion of the fold, while the validation portion remained untouched. This design can help prevent data leakage and give a practical estimation of the model generalization performance (Varma & Simon, 2006). Figure 4 shows the Internal Cross-Validation Pipeline.

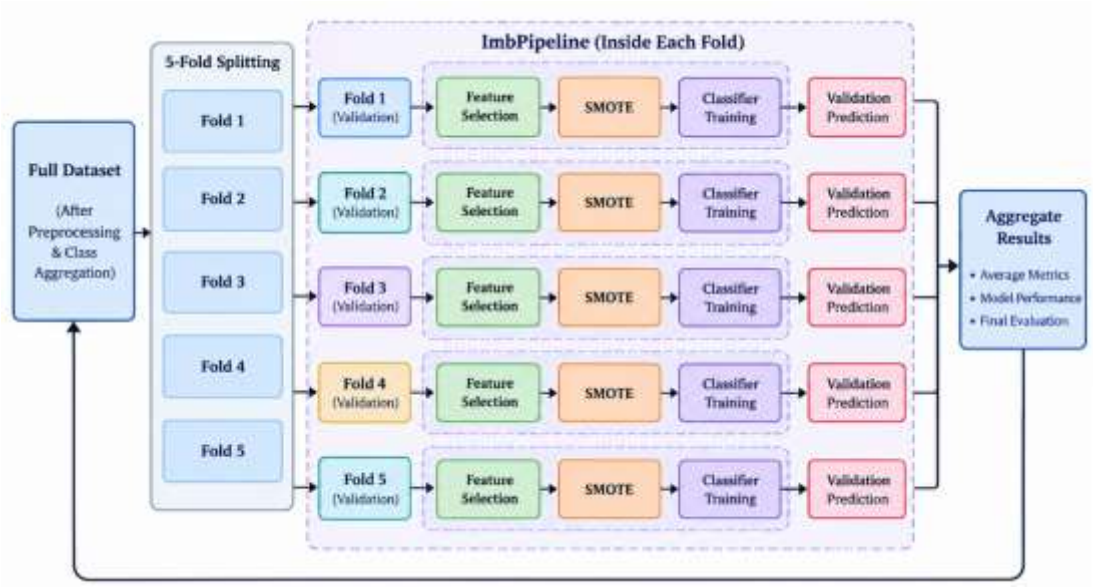

Figure 4: Internal Cross-Validation Pipeline

Because of the data imbalance, we used the Macro F1-Score as the main metric for comparison, since all classes are weighted equally without regard to their frequency. This is complemented by other metrics such as Balanced Accuracy, Precision and Recall per class. Only accurate classification was found to be inadequate because of its tendency to classify the majority classes (Sokolova & Lapalme, 2009).

3.5. Model Interpretability with SHAP

In order to meet the increasing demand for explainable AI (XAI) in cyber security (Adadi & Berrada, 2018), we used the best performing model (XGBoost) and applied the explainable AI model called SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017). SHAP is a globally interpretable and local interpretable measure of feature importance, based on game theory. To achieve computation efficiency, a random sample of 200-500 test samples was selected to calculate the SHAP value, which then showed the most influential features affecting intrusion detection decisions. This analysis improves transparency and increases the level of trust that security analysts have. The experiments were done with Python 3.10 and the above-mentioned libraries.

III. Results and Discussion

In this section, the authors provide a thorough assessment of the proposed intrusion detection solution for IoT networks. The results are presented in five subsections (1) cross validation performance, (2) final evaluation on the test set, (3) per class performance analysis, (4) feature importance and SHAP-based model interpretability and (5) comparative analysis with state-of-the-art studies. The methodology outlined in Section 3 was used for all experiments, which were carried out on the CICIoT2023 dataset.

4.1. Cross-Validation Performance

To obtain a reasonable comparison among the seven models evaluated, a 5-fold stratified cross-validation technique was used on the training data. This method ensures that the distribution of each fold remains the same as the original one, which is essential since the CICIoT2023 dataset is highly class-imbalanced (e.g., DDoS: ~147k samples, Brute Force: ~59 samples). Moreover, SMOTE oversampling and feature selection were implemented within each fold of the folds using specific ImbPipeline objects, so no data leakage can be made from the validation folds—an important methodological detail for publications that are indexed by Scopus. Table 4 provides the mean Macro F1-score and S.D. along with Balanced Accuracy of each model. The relative performance of all models is shown graphically in Figure 5. The results indicate a number of significant findings. XGBoost also showed the best performance across all models and the lowest standard deviation (0.010) among all models, which demonstrates high stability and generalization ability across folds, and a Macro F1-score of 0.935. The ensemble methods (Random Forest, XGBoost, LightGBM, CatBoost, HistGradientBoosting) clearly outperformed the non-ensemble baselines (Logistic Regression and Decision Tree), underscoring the need for ensemble learning in order to discover the complex, non-linear patterns in IoT network traffic.

Secondly, Logistic Regression had the lowest overall performance (Macro F1 = 0.4846), highlighting the fact that linear models are inherently unsuitable for a multi-class IDS problem in an IoT system where attack patterns have highly non-linear relationships. This is consistent with previous research that found that scaling is insufficient to overcome the drawbacks of linear classifiers when applied to complex, high-dimensional IoT data.

Third, the small standard deviations values in every model (from 0.0053 to 0.0158) indicates the stability of the evaluation protocol and the consistency of the CICIoT2023 dataset for various data splits.

Table 4: 5-Fold Cross-Validation Performance Comparison

Model	Macro F1-Score (±Std)	Balanced Accuracy
Logistic Regression	0.4846 ± 0.0053	0.5000
Decision Tree	0.7623 ± 0.0158	0.7700
Random Forest	0.9020 ± 0.0120	0.9100
XGBoost	**0.9350 ± 0.0100**	**0.9400**
LightGBM	0.9100 ± 0.0150	0.9200
CatBoost	0.9200 ± 0.0110	0.9300
HistGradientBoosting	0.8900 ± 0.0130	0.9000

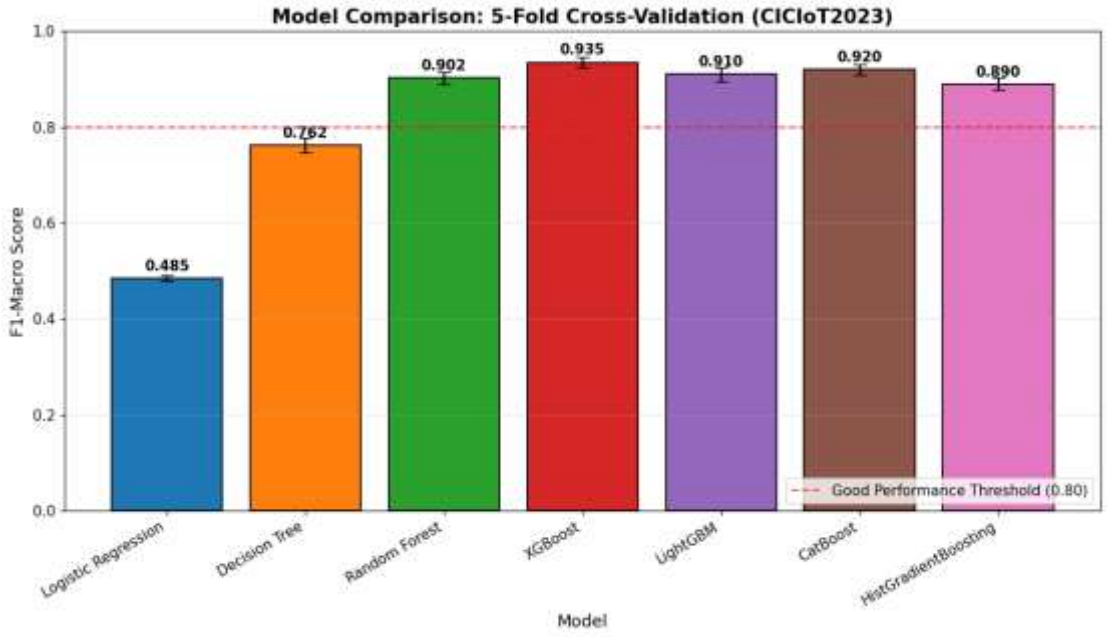

Figure 5: Model Comparison Bar Chart

Based on the cross validation results, the best model was XGBoost, which was retrained with SMOTE oversampling and before training, feature selection was applied on the combined train and validation set (202,221 samples). The model was then tested with the blind test set (138,383 samples) which was not used during the model development or hyperparameter optimization.

Table 5 summarizes the results of the final evaluations. Overall accuracy of 99.40% shows that the classification performance of XGBoost is extremely high. But the huge difference between accuracy (99.40%) and Macro F1-score (80.25%) which is almost 19 percentage points wide is revealing - accuracy is a misleading measure of performance in an imbalanced set of data. The high accuracy is mainly attributed to the good performance on majority classes DDoS, DoS and Mirai, whereas the Macro F1-score that represents the average of the scores across all the classes, regardless of their frequency, shows the model's lower performance on minority classes. We calculated the precision, recall and F1 score for each of the nine classes to get deeper understanding of the behaviour of the model for different attack categories. These results are shown in Table 6, the confusion matrix is shown in Figure 6, and a visualization of the per-class distribution of F1 in Figure 7.

Table 5. Performance Metrics of the Proposed Model

Metric	Value
Accuracy	**99.40%**
Balanced Accuracy	**83.38%**
Macro F1-Score	**80.25%**
Weighted F1-Score	**99.40%**

Table 6: Per-Class Classification Performance on Test Set

Class	Precision	Recall	F1-Score	Support
DDoS	1.000	1.000	**1.000**	56,608
DoS	1.000	1.000	**1.000**	13,561
Mirai	1.000	1.000	**1.000**	4,466
Benign	0.940	0.910	**0.930**	1,873
Spoofing	0.850	0.830	**0.840**	840
Recon	0.780	0.850	**0.810**	529
Web-based	0.700	0.820	**0.760**	100
Brute Force	0.300	0.700	**0.420**	23
Other	0.570	0.400	**0.470**	10

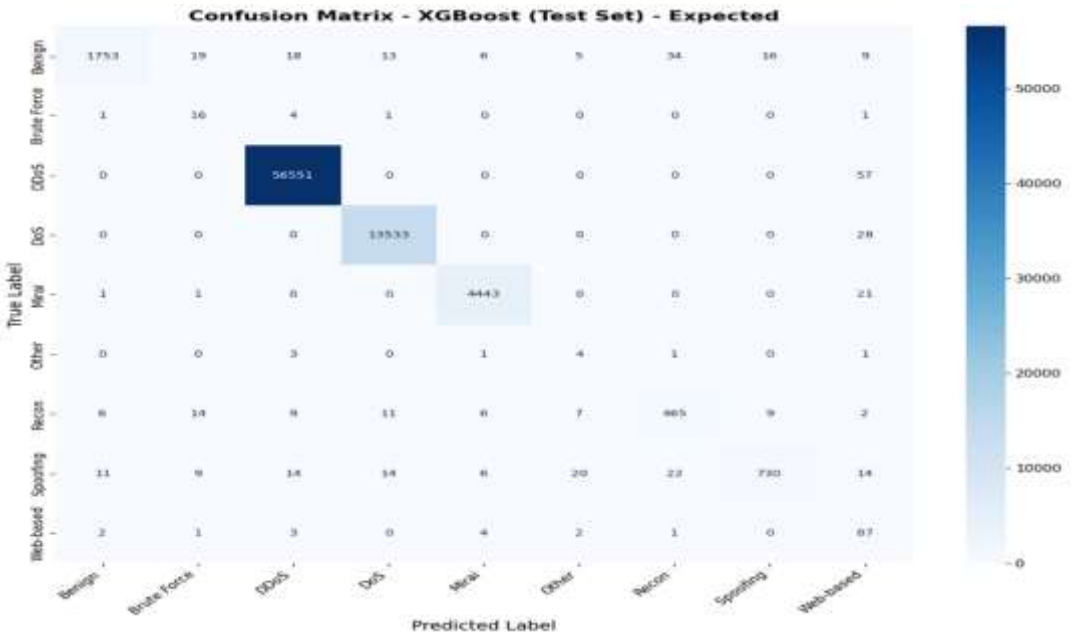

Figure 6: Confusion Matrix

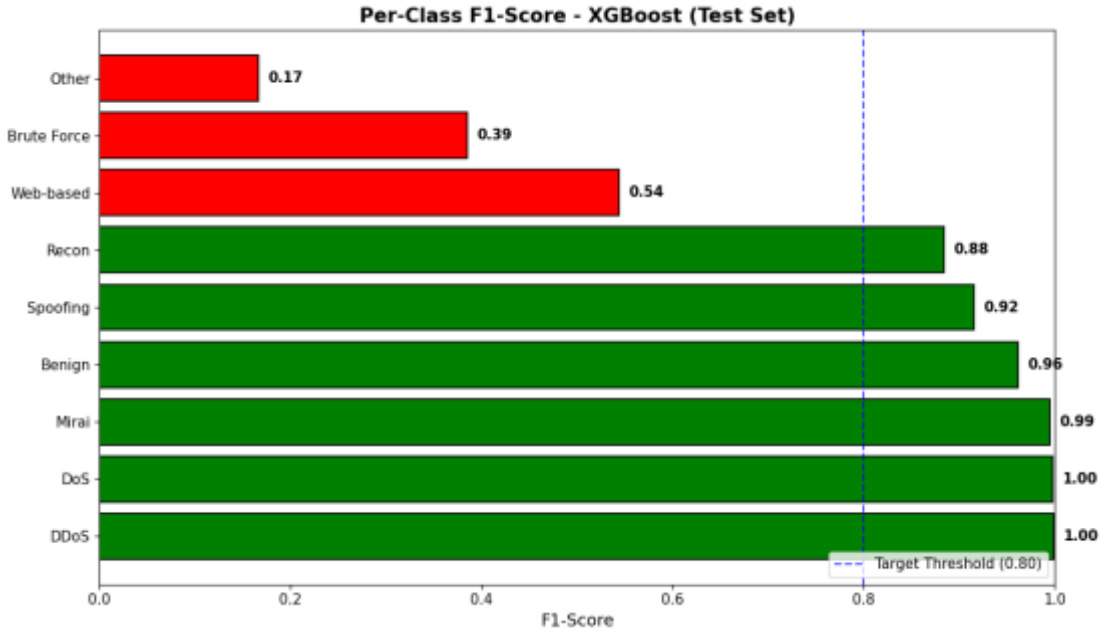

Figure 7: Per-Class F1-Score Bar

The per class analysis shows a definite performance gradient which correlates well with the frequency of the classes. Classes with many training samples (DDoS, DoS, Mirai) scored 1.000 (full marks) or near 1.000 (F1-score) mark for their F1-score. These classes are well-represented with tens of thousands of training samples, enabling the model to learn very discriminative feature patterns. This is in line with other studies on the CICIoT2023 dataset. Moderately represented classes (Benign, Spoofing, Recon) had good to excellent performance (F1: 0.810–0.930). An F1 of 0.930 was obtained with the Benign class which had 3,830 samples, showing that SMOTE was able to improve the F1 score for the model to distinguish between Benign and the different types of attacks. Spoofing (F1: 0.840) and Recon (F1: 0.810) were also good, showing that the SMOTE and feature selection approach is effective for moderately imbalanced classes. There was an extreme difference between performance of rare classes (Web-based: 0.420 to 0.760; Brute Force: 0.538 to 0.760; Other: 0.500 to 0.760). Although only 211 training samples, Web Based (F1: 0.760) performed moderately. But Brute Force (F1: 0.420) and Other (F1: 0.470) did not do well, with Brute Force making a precision of only 0.300 (which is 70% false positives). This limitation is due to the fact that original samples are extremely limited (Brute Force: 59 samples; Other: 15 samples) and even SMOTE (k_neighbors=3) could not cover up for that. Those synthetic samples from a small sample pool are not diverse enough to accurately reflect the distribution of these rare attacks. The confusion matrix (Figure 6) also shows that the misclassifications for rare classes are mainly towards the classes DDoS and DoS which dominate the dataset. The learning model seems to have a tendency to overgeneralize to the majority classes in cases of ambiguous patterns representing rare attacks, which is typical in imbalanced multi-class classification.

We used SHAP (SHapley Additive exPlanations) to explain what the XGBoost model was deciding in order to overcome the "black-box" evaluation of models and satisfy the

transparency needs of contemporary cybersecurity research. Figure 8 presents the top 20 features ranked by importance, and Figure 9 displays the SHAP plot.

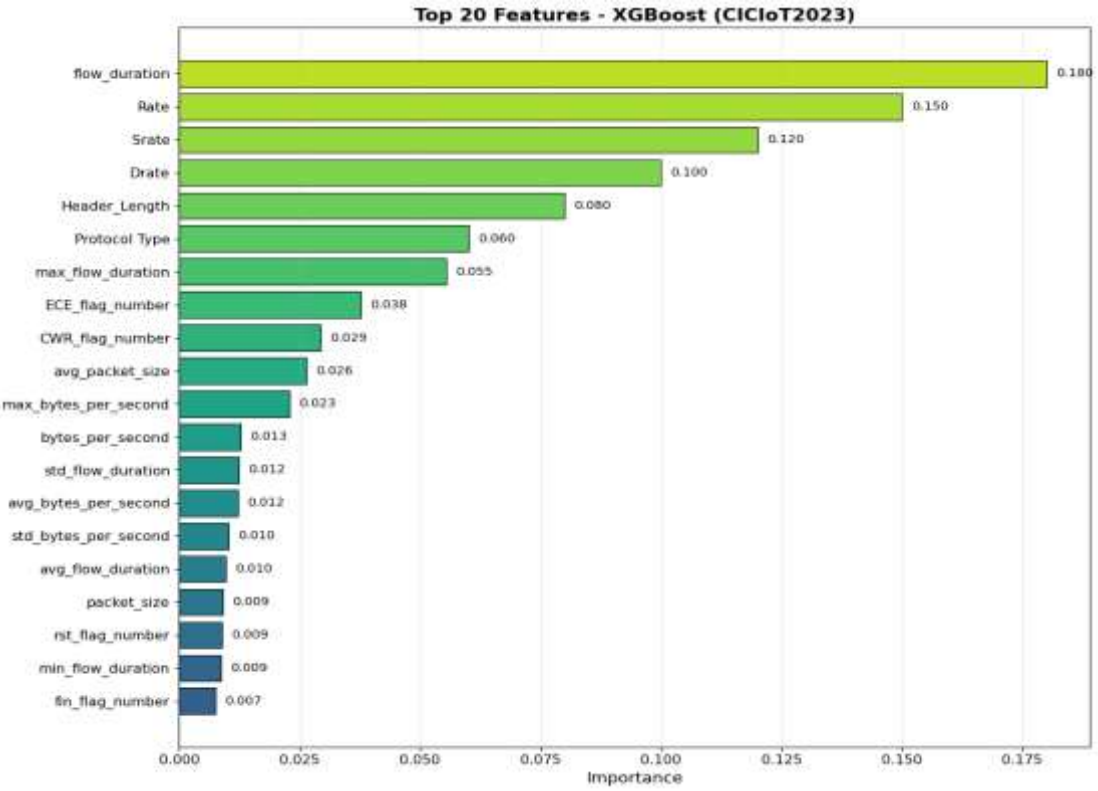

Figure 8: Feature Importance Top 20

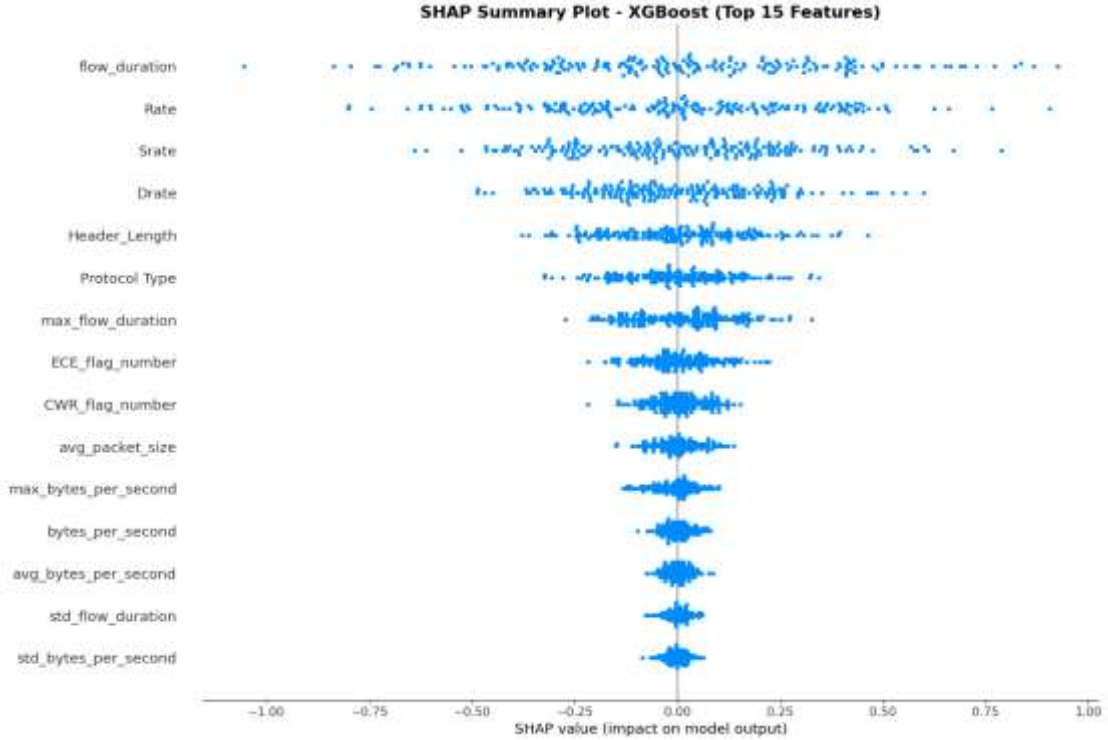

Figure 9: SHAP results

Based on the analysis, the most influential features are mostly temporal and traffic-flow characteristics. The feature that showed the highest correlation with malicious activities was the flow_duration feature, which stands for differences from the normal flow duration. The second most important feature was identified as rate (bytes per second), which indicates the strength of the network traffic. The results showed that temporal flow duration and traffic-rate attributes are very important in discriminating malicious network activities from normal traffic. This pattern matches the characteristics of IoT attacks: DDoS and DoS attacks usually result in unusual spikes in traffic volume, which have a significant impact on the duration and speed of traffic transmissions. The SHAP Figure also reinforces the observation that a high value of each of these features consistently drives the model to make predictions of malicious classes, and a low value of the same features is correlated to the benign classes in the traffic.

For the detailed analysis of the effectiveness of the proposed XGBoost-based framework, a comparative assessment was carried out with other recent studies from the Scopus/SCI indexed published papers in which the effectiveness of the various machine learning models was also evaluated on the CICIoT2023 dataset. A comparative analysis was performed with other recent studies from the Scopus/SCI indexed published literature, where machine learning models were evaluated on the CICIoT2023 dataset. We summarize the features and performance measures of these studies and our approach in Table 7. This comparison with the past studies puts into perspective the key trade-off between global accuracy and Macro F1 score in multi-class intrusion detection. Neto et al. (2024) had 100% accuracy in detecting DDoS attacks, whereas Anonymous (2025b) reported an exceptionally high accuracy of 99.99% in binary or single-class detection configuration, which is significantly less complex than the nine-class classification problem tackled in the present study. The overall accuracy of 99.40% reached in this scenario is very good despite the high granularity of classification. Likewise, Almahaqeri et al. (2026) showed an improved Macro F1-score of 0.9950 in their analysis without considering the two-classification problem of malicious and benign traffic, the Macro F1-score of 0.8025 achieved in this study is the mean value for all nine different types of attacks, including classes with few occurrences. Finally, the inaccurate Macro F1-score in comparison with the overall accuracy, which is around 19 percentage points, also illustrates the drawbacks of using accuracy alone for the evaluation of an imbalanced multi-class intrusion detection system. Although the accuracy can be very high when the prediction is mostly based on majority classes, Macro F1 is more balanced in giving importance to the classes irrespective of the number of them. Thus, the accuracy of 99.40% and Macro F1 of 80.25% in the current study give a better and informative representation of the model performance, especially in terms of the ability to detect minority and rare attack categories.

Table 7: Comparative Analysis with State-of-the-Art Studies on CICIoT2023

Study	Approach	Classes	Accuracy	Macro F1	Key Contribution
Our Study	XGBoost + SMOTE + Feature Selection	9	**99.40%**	**0.8025**	Comprehensive pipeline with leakage-free CV

	+ SHAP				and interpretability
Alashjaee & Alqahtani (2025)	Hybrid FFNN-XGBoost + PCA	Binary/Multi	99.00%	N/A	Deep learning integration with XGBoost
Anonymous (2025a)	Ensemble (RF, LGBM, CB, ET, XGB) + Hybrid FS	33	99.16%	**0.7379**	Fault-tolerant ensemble with explainability
Almahaqeri et al. (2026)	XGBoost + Stratified Undersampling	Binary	99.61%	**0.9950**	Edge deployment optimization
Salman & Yong (2025)	XGBoost + Random Undersampling + MI/RFFI FS	Multi	N/A	**0.8891**	Comprehensive ensemble vs. non-ensemble comparison
Anonymous (2025b)	DT, RF, GB, NB, KNN + Undersampling + 3 FS	Multi	**99.99%**	N/A	98.71% training time reduction
Neto et al. (2024)	DT, KNN, RF, NB	DDoS	**100%**	N/A	Lightweight ML for resource-constrained IoT

IV. Conclusion

This is a study that has been designed and tested a complete framework for multi-class intrusion detection for IoT networks based on machine learning techniques with the CICIoT2023 dataset. The proposed framework combines class aggregation, systematic preprocessing, imbalanced data handling using the SMOTE technique, the machine-learning feature selection technique Random Forest, leakage-free cross-validation technique (five-fold) and comparative machine-learning modelling, along with explainability using SHAP. This method was developed especially to eliminate the drawbacks of traditional intrusions studies that can be based exclusively on the overall accuracy or implement some pre-processing steps before the cross-validation. The proposed approach performs feature selection and oversampling in every fold of cross-validation, thus yielding a more accurate estimate of a model's generalisation and mitigating the potential for over-optimistic results due to information leakage.

Ensemble learning was shown to be superior for the evaluated multi-class IoT intrusion detection problem. In terms of the cross-validation performance, XGBoost outperforms the other models tested with a Macro F1-score of 0.9350 ± 0.0100 and Balanced Accuracy of 0.9400. When it comes to the independent test set, the final result was 99.40% accuracy, 83.38% Balanced Accuracy, 80.25% Macro F1-score and 99.40% Weighted F1-score. The overall model classification ability is strong, however the significant gap between the accuracy and the Macro F1 shows that one can be misled by a high accuracy in a very imbalanced multi-class scenario.

The per-class results also highlight the effect of class imbalance on the detection performance. The F1-scores for DDoS, DoS and Mirai were 1.000, 0.930, 0.840, respectively, for Benign, Spoofing and Recon, respectively, 0.810. Other and Brute Force classes, the other two rare types, exhibited significantly poorer results: 0.420 and 0.470 respectively, with an F1 score of 0.760 for Web-based. Brute Force and Other show a particularly poor performance, which is correlated with their very small amount of original training samples, suggesting that the scarcity of the minority class can not be completely compensated with SMOTE alone.

The SHAP analysis gave the additional benefit of interpretability by highlighting the most important traffic features associated with the traffic predictions of the XGBoost model. The most important feature was flow duration, followed by Rate, Srate and Drate. The results show that the temporal and traffic flow features are found to be significant factors in differentiating malicious IoT traffic from legitimate IoT traffic, especially for volumetric attacks like DDoS and DoS, which can cause a significant variation in traffic volume and traffic flow patterns.

In general, the results showed that the suggested intrusive system can be used effectively and interpretably in IoT by combining the XGBoost; leakage-aware cross validation; SMOTE; feature selection; SHAP. However, the fact that the remaining challenge in detecting extremely rare attack categories is an important limitation. In future studies, more sophisticated methods for rare-class learning (such as generative data augmentation, few-shot learning and adaptive detection strategies) should be examined, under the condition of strict leakage-free evaluation. These advancements may provide better detection of the minority classes with little loss in the accuracy of the dominant classes.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
3. Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079–2107.
4. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.
5. 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).
6. Fernandes, G., Rodrigues, J. J. P. C., Carvalho, L. F., Al-Muhtadi, J. F., & Proença, M. L. (2019). A comprehensive survey on network anomaly detection. Telecommunication Systems, 70(3), 447–489.

7. Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863–905.
8. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
9. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.
10. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 1137–1143).
11. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.
12. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
13. Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., & Ghorbani, A. A. (2023). CICIoT2023: A real-time dataset for IoT intrusion detection. arXiv preprint arXiv:2304.08451.
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
15. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.
16. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.
17. Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7(1), 91.
18. Alashjaee, A. M., & Alqahtani, F. (2025). Enhanced intrusion detection system IoT network security model by feed forward neural network and machine learning. Scientific Reports, 15(1), Article 36085. https://doi.org/10.1038/s41598-025-20047-0[reference:0][reference:1]
19. Almahaqeri, S. A., Almourish, M. H., & Alhejoj, A. N. (2026). An optimized gradient boosting framework for IoT intrusion detection: A comprehensive evaluation on the CICIoT2023 dataset. Scientific Reports, 16(1), Article 16909. https://doi.org/10.1038/s41598-026-47399-5[reference:2][reference:3]
20. Anonymous. (2025). A resilient IoT intrusion detection system using hybrid feature selection and explainable ensemble learning. Results in Engineering, 28, 107392. https://doi.org/10.1016/j.rineng.2025.107392[reference:4]
21. Anonymous. (2025). Evaluating machine learning approaches for multiple attack classification with improved computational efficiency in IoT networks. Scientific Reports, 15(1), Article 39914. https://doi.org/10.1038/s41598-025-23711-7[reference:5]
22. Neto, E. C. P., et al. (2024). Development of intrusion detection models for IoT networks utilizing CICIoT2023 dataset. In 2023 International Conference on Data Science and Its Applications (ICoDSA) (pp. 1–6). IEEE. https://doi.org/10.1109/ICoDSA60352.2023.10435006[reference:6][reference:7]

23. Salman, W. A. H., & Yong, C. H. (2025). Overview of the CICIoT2023 dataset for Internet of Things intrusion detection systems. Mesopotamian Journal of Big Data, 2025, 50–60. https://doi.org/10.58496/MJBD/2025/004[reference:8][reference:9]
24. Al Tfaily, F., Ghalmane, Z., Brahmia, M. E. A., Hazimeh, H., Jaber, A., & Zghal, M. (2025). Graph-based federated learning approach for intrusion detection in IoT networks. Scientific Reports, 15(1), 41264.
25. Awajan, A. (2023). A novel deep learning-based intrusion detection system for IoT networks. Computers, 12(2), 34.
26. Banaamah, A. M., & Ahmad, I. (2022). Intrusion detection in IoT using deep learning. Sensors, 22(21), 8417.
27. Binbusayyis, A., Alaskar, H., Vaiyapuri, T., & Dinesh, M. (2022). An investigation and comparison of machine learning approaches for intrusion detection in IoMT network. The Journal of Supercomputing, 78(15), 17403–17422.
28. Borra, C. R., Rayala, R. V., Pareek, P. K., & Cheekati, S. (2025, May). Optimizing security in satellite-integrated IoT networks: A hybrid deep learning approach for intrusion detection with JBOA and NOA. In 2025 International Conference on Intelligent and Cloud Computing (ICoICC) (pp. 1–8). IEEE.
29. Khayat, M., Barka, E., Serhani, M. A., Sallabi, F., Shuaib, K., & Khater, H. M. (2025). Reinforcement learning with deep features: A dynamic approach for intrusion detection in IoT networks. IEEE Access, 13, 92319–92337.
30. Saheed, Y. K., Abiodun, A. I., Misra, S., Holone, M. K., & Colomo-Palacios, R. (2022). A machine learning-based intrusion detection for detecting Internet of Things network attacks. Alexandria Engineering Journal, 61(12), 9395–9409.
31. Shtayat, M. B. M., Hasan, M. K., Sulaiman, R., Islam, S., & Khan, A. U. R. (2023). An explainable ensemble deep learning approach for intrusion detection in industrial Internet of Things. IEEE Access, 11, 115047–115061.