Journal of
Computer & Electrical and Electronics Engineering Sciences
Original Article
DOI: 10.51271/JCEEES-0036

Explainable AI-driven IoT botnet detection for 6G networks

Merve Güllü*[1], Ercan Atagün[2], Serdar Biroğul[2], Necaattin Barışçı[3]
[1]Department of R&D, Türk Telekom, Ankara, Turkiye
[2]Department of Computer Engineering, Faculty of Engineering, Düzce University, Düzce, Turkiye
[3]Department of Computer Engineering, Faculty of Technology, Gazi University, Ankara, Turkiye

Cite this article as: Güllü, M., Atagün, E., Biroğul, S., & Barışçı, N. (2026). Explainable AI-driven IoT botnet detection for 6G networks. *J Comp Electr Electron Eng Sci, 4*(2), 46-55. https://doi.org/10.51271/JCEEES-0036

Received: 07.11.2025 ◆ **Accepted**: 30.12.2025 ◆ **Published**: 13.10.2026

ABSTRACT

The rapid proliferation of internet of things (IoT) devices in smart homes, healthcare, and industrial automation introduces severe security challenges, especially with the transition toward 6G intelligent networks. Traditional machine learning models achieve high detection accuracy for IoT botnets but often lack interpretability, limiting their trustworthiness in mission-critical 6G environments. This study proposes an explainable AI-driven IoT botnet detection framework utilizing the N-BaIoT dataset to bridge this gap. The framework integrates correlation-based feature selection, mutual information analysis, ANOVA filtering, and low-variance elimination to construct a lightweight yet discriminative feature space. Several ensemble learning algorithms were evaluated, with light gradient boosting machine (LightGBM) achieving the highest accuracy (99.82%) and exceptional stability across cross-validation folds. To enhance transparency, model decisions were explained using SHapley Additive ExPlanations (SHAP) and local interpretable model-agnostic explanations (LIME), revealing key traffic features that differentiate benign and malicious flows. The proposed system demonstrates not only high detection precision but also human-interpretable reasoning, fulfilling essential trust and reliability requirements for next-generation 6G IoT security architectures.
Keywords: IoT botnet detection, explainable Artificial Intelligence, 6G networks, cybersecurity, feature selection, anomaly detection

INTRODUCTION

The proliferation of internet of things (IoT) devices across smart homes, healthcare systems, industrial automation, and urban infrastructure has increased rapidly (Navneet Kaur, 2025). This expansion has introduced unprecedented security vulnerabilities that threaten the integrity of emerging 6G networks (Emmanuel, 2025). With the rapid global growth of device deployments, the attack surface for botnet exploitation has expanded significantly. IoT botnets such as Mirai and Gafgyt have shown how dangerous compromised device networks can be. They are capable of launching large-scale distributed denial of service (DDoS) attacks that disrupt critical services and threaten user privacy (Son, 2024). Unlike traditional computing systems, IoT devices operate with very limited resources. Their restricted processing power, memory, and energy capacity make it difficult to apply conventional security mechanisms.

Furthermore, the heterogeneity of IoT manufacturers, firmware implementations, and communication protocols creates fragmented security landscapes that adversaries systematically exploit. Machine learning (ML) approaches have emerged as promising solutions for automated botnet detection, achieving high classification accuracy on benchmark datasets (Ali, 2024; Abbasi, 2025; Ullah, 2025). However, the black-box nature of these models poses fundamental challenges for operational deployment in security-critical domains (Savant, 2024). Network administrators and incident response teams require transparent explanations of classification decisions to validate model behavior, diagnose false positives, and identify adversarial evasion attempts. The absence of interpretability undermines trust in automated systems and hinders regulatory compliance with emerging AI governance frameworks. This gap between detection accuracy and operational explainability motivates the integration of Explainable AI (XAI) techniques into IoT security architectures (Gummadi, 2024).

This paper addresses these challenges by proposing an explainable AI-driven framework for IoT botnet detection optimized for 6G network environments. This paper makes a meaningful contribution by combining high-accuracy classification with actionable explanations, which are essential for building trustworthy AI systems in next-generation network security.

Related Work

Detecting IoT botnets is critical for ensuring the security and stability of the IoT ecosystem. With billions of interconnected devices operating in diverse environments, botnet attacks can exploit network resources to launch DDoS assaults, steal data, and compromise device integrity. Therefore, the early

*Corresponding Author: Merve Güllü, gullu.merv@gmail.com

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

and accurate identification of malicious IoT traffic is a key requirement for maintaining network integrity and service continuity.

The literature demonstrates the successful application of ML and deep learning (DL) techniques for IoT botnet detection. Publicly available datasets such as N-BaIoT, BoT-IoT, IoT-23, and MedBIoT have significantly accelerated research progress by enabling reproducible and comparable evaluations. Experiments on these datasets have shown that traditional classifiers (e.g., Support Vector Machine, k-NN, decision trees), ensemble methods, and deep learning architectures typically achieve detection accuracies exceeding 95-99%.

Beyond N-BaIoT, several other datasets, such as BoT-IoT, MedBIoT, and IoT-23 have significantly advanced IoT botnet detection research. The BoT-IoT dataset (Nickolaos Koroniotis, 2019), which includes synthetically generated benign and malicious traffic (e.g., DDoS, scanning, and data exfiltration), provides a comprehensive benchmark for algorithm evaluation. Likewise, (Jadel Alsamiri, 2019) reported that the k-NN algorithm achieved up to 99% accuracy in detecting IoT botnet traffic. These results confirm that traditional supervised learning approaches can reach near-perfect detection rates on IoT malware datasets. However, interpretability and deployment efficiency remain key challenges for resource-constrained IoT devices.

Similarly, (Evanson Mwangi Karanja, 2020) converted N-BaIoT network traffic into grayscale images and extracted Haralick texture features to identify IoT malware. When combined with standard classifiers, this image-based approach achieved approximately 95% accuracy, demonstrating the effectiveness of visual feature extraction in representing malicious IoT traffic patterns. These findings highlight that combining deep learning architectures with creative feature extraction strategies can effectively model malicious behaviors in IoT environments.

More recently, the IoT-23 dataset (Sebastian Garcia, 2020) has provided labeled network captures from real IoT devices, encompassing 20 malicious and three benign scenarios. IoT-23 is frequently used alongside N-BaIoT and BoT-IoT for comparative analysis. Ensemble boosting algorithms such as XGBoost and LightGBM have been evaluated on these datasets, achieving over 99% accuracy in multi-class botnet detection. (Saied, 2023) compared tree-based ensemble techniques comprehensively on N-BaIoT and found that Random Forest achieved 99.999% multi-class accuracy. Empirical findings confirm that ensemble learners enhance detection rates and maintain computational efficiency, making them highly practical for real-world IoT security applications (Al-Othman, 2020) applied multiple ML algorithms to this dataset, after addressing class imbalance using Synthetic Minority Oversampling Technique (SMOTE), and found that ensemble decision tree models achieved the highest accuracy.

Trajanovski and Zhang (2021) proposed IoT-BDA, an integrated honeypot and sandbox framework that automates the capture, analysis, and reporting of IoT botnets and reveals advanced evasion techniques used in real attacks.

In one of the earliest studies following the release of the N-BaIoT dataset, Krishna and Thangavelu (2021) used traditional machine learning pipelines with metaheuristic feature selection and a random forest classifier to detect IoT attacks.

Similarly, Abbasi et al. (2021) applied classic learning techniques such as logistic regression and artificial neural networks to the newly introduced N-BaIoT dataset to perform anomaly detection in IoT networks.

In this context, (Alkahtani, 2021) applied a hybrid CNN-LSTM model to the N-BaIoT dataset for classifying botnet traffic across different IoT device types (e.g., webcams, thermostats, doorbells). Separate models were trained for each device category, achieving 88-90% detection accuracy in distinguishing between benign and malicious traffic.

The MedBIoT dataset (Alejandro Guerra-Manzanares, 2022) built from traffic collected from 83 IoT devices under real-world conditions, contains benign and malicious samples from well-known botnets such as Mirai, Bashlite, and Torii. Using this dataset, supervised learning models such as random forest have demonstrated strong performance, achieving 96.2% accuracy in differentiating botnet and benign traffic. The availability of datasets like MedBIoT enables researchers to assess the generalizability of detection models across diverse IoT environments and malware families.

Santhadevi and Janet (2023) introduced an edge-based stacked deep learning system that processes NetFlow data and uses vectorized CNN and LSTM models with optimization methods to detect IoT malware in real time.

Rallapalli et al. (2023) presented a hybrid feature selection approach with a modified CNN that improves IIoT anomaly detection performance on the N-BaIoT dataset compared with existing models.

In recent years, beyond achieving high detection accuracy, the emphasis has shifted toward explainability. Considering the constraints of IoT environments, limited computing resources, energy efficiency, and real-time operational requirements, research has increasingly focused on developing detection systems that are accurate, transparent, interpretable, and deployable at the network edge.

Table 1 summarizes several representative studies that utilized N-BaIoT or similar IoT botnet datasets, highlighting their approaches and key results. While the primary focus in IoT botnet detection has been improving accuracy and efficiency, there is a growing interest in making these detection models more explainable. As complex AI models (e.g., deep neural networks and ensembles) are deployed for security, understanding the reasoning behind their alerts becomes crucial for operator trust and effective mitigation.

While Saied (2023) focused primarily on maximizing detection accuracy through tree-based ensemble techniques and Navneet Kaur (2025) employed XAI methods in 6G-IoT security contexts, the proposed framework distinguishes itself by simultaneously integrating explainability mechanisms with a lightweight LightGBM classifier optimized through multi-stage feature selection, thereby achieving both high detection accuracy and comprehensive interpretability suitable for resource-constrained edge deployments in 6G environments.

Table 1. Comparison of IoT botnet detection studies

Study (year)	Dataset (s)	Approach/model	XAI used?	Performance
Meidan, 2018	N-BaIoT	Deep autoencoder (per-device anomaly detector)	No	100% TPR (binary detection)
Alkahtani, 2021	N-BaIoT	CNN-LSTM hybrid deep neural network	No	~90% accuracy (binary per device)
Evanson Mwangi Karanja, 2020	N-BaIoT	Traffic-to-image+texture feature ML classification	No	80-95% accuracy (varies by class)
Al-Othman, 2020	BoT-IoT	Various ML classifiers; ensemble (J48, RF) best	No	96% accuracy (multi-class)
Jadel Alsamiri, 2019	BoT-IoT	Ensemble of decision trees; feature augmentation	No	99% accuracy (binary)
(Alejandro Guerra-Manzanares, 2022)	MedBIoT	Supervised ML (RF, etc.) on new IoT botnet data	No	96.2% accuracy (binary)
(Saied, 2023)	N-BaIoT	Bagging/boosting ensemble trees (RF, XGBoost)	No	99.99% accuracy (multi-class)
(Do, 2024)	N-BaIoT	CNN, LSTM, GRU	No	**89%** accuracy (multi-class)
(Nazir, 2024)	N-BaIoT	CNN-LSTM hybrid model	No	99.99% accuracy (multi-class)
(Serhane, 2025)	N-BaIoT	Extra Trees Using PCA	No	99.94% accuracy (multi-class)

N-BaIoT, BoT-IoT, and MedBIoT are dataset names. IoT: Internet of things, XAI: Explainable Artificial Intelligence, TPR: True positive rate, CNN: Convolutional neural network, LSTM: Long short-term memory, ML: Machine learning, J48: Java decision tree algorithm (C4.5), RF: Random forest, N-BaIoT: Network-based detection of IoT botnet attacks, BoT-IoT: Botnet Internet of Things, MedBIoT: Medium-sized Botnet internet of things, XGBoost: Extreme gradient boosting, GRU: Gated recurrent unit, PCA: Principal component analysis

METHODS

Ethics

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data. All procedures were carried out in accordance with the ethical rules and principles.

Methodology

In this section, the process steps followed in the problem and the models used are explained and the architecture of the process is shown in **Figure 1**.

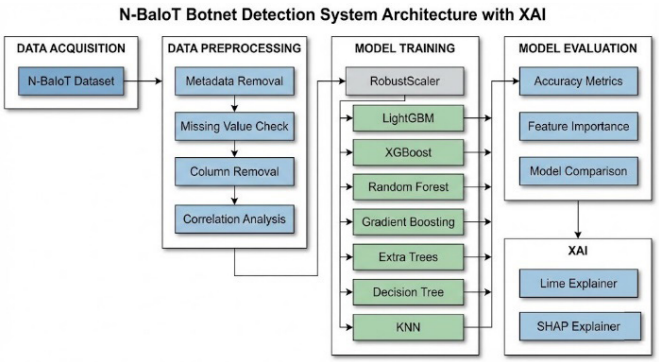


Figure 1. The system architecture for N-BaIot and XAI

N-BaIoT: Network-based detection of IoT botnet attacks, XAI: Explainable Artificial Intelligence, LightGBM: Light gradient boosting machine, XGBoost: Extreme gradient boosting, KNN: K-nearest neighbors, LIME: Local interpretable model-agnostic explanations, SHAP: SHapley additive exPlanations

N-BaIoT Dataset

The N-BaIoT (Meidan, 2018) was introduced to address the lack of public IoT botnet datasets by providing real network traffic from nine commercial IoT devices infected with Mirai and Bashlite malware. The dataset includes benign traffic and 10 attack types (5 Bashlite and 5 Mirai) and can be used for both anomaly detection (benign vs. malicious) and multi-class classification tasks.

Feature Engineering Pipeline

Correlation analysis: Pearson correlation coefficient was computed to identify and remove highly correlated features, reducing multicollinearity. For feature pairs X_i, X_j the correlation coefficient is:

$$\rho_{\{x_i,x_j\}} = \frac{\mathrm{cov}(X_i, X_j)}{\sigma_{\{x_i\}}\,\sigma_{\{x_j\}}}$$
$$\rho_{\{x_i,x_j\}} = \frac{E\left[\left(X_i - \mu_{\{x_i\}}\right)\left(X_j - \mu_{\{x_j\}}\right)\right]}{\sigma_{\{x_i\}}\,\sigma_{\{x_j\}}} \quad (1)$$

Where μ and σ denote mean and standard deviation, respectively. Features with $|\rho|>0.95$ were considered redundant, and one feature from each highly correlated pair was removed (Max Kuhn, 2013).

Mutual information (MI): MI quantifies the mutual dependence between features and target labels, capturing both linear and non-linear relationships.

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (2)$$

Where p (x, y) is the joint probability distribution, and p (x), p (y) are marginal distributions. Features with MI scores below a threshold were eliminated as they provide minimal information gain for classification.

ANOVA F-test: Analysis of Variance F-test assesses the statistical significance of feature means across different classes (Montgomery, 2017). The F-statistic is calculated as:

$$F = \frac{MS_{between}}{MS_{within}}$$
$$F = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 x\,(N-k)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij} - \bar{X}_i\right)^2 x\,(k-1)} \quad (3)$$

Where k is the number of classes, n_i is the sample size of class i and N is the total sample size. Features with higher F-scores demonstrate greater discriminative power between classes.

Low variance elimination: Features with variance below a specified threshold were removed to eliminate near-constant features that contribute minimal discriminative information (Guyon, 2003). The variance of feature X is:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad (4)$$

This technique reduces dimensionality while preserving informative features.

Classification Models

Extra trees classifier: Extremely Randomized Trees (Extra Trees) is an ensemble method that constructs multiple decision trees with random node splits (Pierre Geurts, 2006). Unlike Random Forests, Extra Trees randomly selects split thresholds without searching for optimal values, reducing computational complexity:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (5)$$

Where T is the number of trees, h_t is the prediction of tree t, and x is the input feature vector.

XGBoost: XGBoost (eXtreme Gradient Boosting) implements gradient boosting with regularization to prevent overfitting (Tianqi Chen, 2016). The objective function minimizes:

$$L(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (6)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \quad (7)$$

Where l is the loss function, $\Omega(f_k)$ is the regularization term, T is the number of leaves, and w represents leaf weights. The model builds trees sequentially, with each tree correcting errors from previous iterations.

Random forest: An ensemble learning technique called Random forest uses bootstrap aggregating, also known as bagging, to build several decision trees (Breiman, 2001). A random subset of characteristics and data samples are used to train each tree. The final prediction aggregates individual tree predictions:

$$\hat{y} = model\ \{h_1(x), h_2(x), \ldots, h_T(x)\} \quad (8)$$

For classification tasks, where mode represents the majority vote.

LightGBM: LightGBM is a decision tree-based gradient boosting framework that has focused on speed and scalability since its development (Wang, 2017). It uses leaf-wise tree growth with depth constraints and histogram-based splitting. The objective function is minimized as:

$$L(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (9)$$

Where l is the loss function, Ω (f_k) is a regularization term, and each tree f_k aims to correct the residuals of the previous iteration.

Decision tree: A decision tree recursively partitions the feature space into regions that minimize impurity, using measures like Gini index or information gain. The predicted output is the mean (regression) or mode (classification) of samples within each leaf:

$$\hat{y} = \text{mode}\ (y_i \in R_j) \quad (10)$$

where R_j denotes the j^{th} region (leaf) of the tree.

K-Nearest Neighbors (KNN)

The KNN method determines the class of a data point based on the label most commonly held by its k nearest neighbors (Steinbach, 2009). The distances between these neighbors are typically calculated using the Euclidean distance, which corresponds to the straight-line distance between two points. The distance metric is typically Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{m}(x_j - x_{ij})^2} \quad (11)$$

The predicted class is determined by:

$$\hat{y} = \text{mode}\ [y_i : x_i \in N_k(x)] \quad (12)$$

Where N_k (x) denotes the set of k nearest training samples to x.

Explainable AI Techniques

Tree-based feature importance: Tree-based models inherently provide feature importance scores by measuring the total reduction in impurity (e.g., Gini impurity or entropy) contributed by each feature across all trees (Breiman, 2001):

$$\text{Importance}(X_j) = \frac{1}{T}\sum_{\{t=1\}}^{T}\sum_{n \in N_t} \Delta i(n, X_j) \quad (13)$$

Where N_t is the set of nodes in tree t that split on feature X_j, and Δi is the impurity reduction at node n.

SHAP Values

SHAP assigns each feature a contribution value based on game theory (Lundberg, 2017).

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \quad (14)$$

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} K$$

$$K = \left[f_{\{S \cup \{i\}\}}(x_{\{S \cup \{i\}\}}) - f_S(x_S)\right] \left[f_{\{S \cup \{i\}\}}(x_{\{S \cup \{i\}\}}) - f_S(x_S)\right] \quad (15)$$

Where F is the set of all features, S is a subset excluding feature i, and f_s represents the model prediction with only features in S. SHAP values satisfy properties of local accuracy, missingness, and consistency.

LIME: LIME explains individual predictions by approximating the model locally with an interpretable linear model (Ribeiro, 2016). For instance x, LIME minimizes:

$$\xi(x) = \text{argmin}_{g \in G}\ \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (16)$$

Where g is an interpretable model, L measures the approximation fidelity, π_x defines the local neighborhood, and Ω(g) penalizes model complexity. LIME generates perturbations around x and fits a sparse linear model to explain local behavior.

Experimental Results

Data preprocessing: The preprocessing process of the N-BaIoT dataset consists of a series of systematic steps aimed at ensuring data quality in botnet detection and optimizing

the representation of attributes. The raw dataset comprises 7,062,819 records with 123 distinct features, collected from 92 separate CSV files reflecting the behavior of various IoT devices in both benign and attack scenarios. During the initial data cleaning, 213 metadata entries were removed. As a result of this process, 7,062,606 valid data records remained. These records were divided into a total of 11 classes: five were Mirai attack types (ack, scan, syn, udp, udpplain); five were Gafgyt attack types (combo, junk, scan, tcp, udp); and the last one was normal (benign) network traffic.

Six metadata columns (file name, data count, device ID, device name, feature name, feature description) were removed because they were completely empty. These columns did not contain any useful information for classification. Next, a feature whose value never changed across all data was identified and removed from the list. This reduced the total number of features from 117 to 116. Features with fixed values do not contribute to machine learning models, creating unnecessary overhead and remaining useless. Therefore, eliminating them was important for both speed and efficiency.

The most significant reduction in the dataset occurred through the removal of duplicate records. This process resulted in the elimination of 4,579,930 redundant records (64.8% of the dataset) from the system. This left 2,482,676 unique data records. This intensive deduplication step eliminates repetitive traffic patterns, a common problem in IoT datasets. It also aims to preserve the fundamental behavioral signatures necessary for accurate classification.

Feature Selection Outcomes

During the feature selection phase, a correlation-based redundancy removal method was used to address the issue of multicollinearity and reduce the risk of overfitting in the model. Pearson correlation analysis revealed 93 features that were highly correlated with each other (correlation coefficient >0.95). This corresponded to 80.9% of the original feature set. The vast majority of these redundant features came from the MI direction statistics group (e.g., MI_dir_L3_weight, MI_dir_L3_mean, MI_dir_L3_variance, MI_dir_L1_weight). Their nearly perfect linear relationship with each other indicated that they carried overlapping information about network traffic characteristics.

The number of features was reduced from 115 to 22 (by 80.9%). This radical step was a deliberate choice. This resulted in four key gains: first, in a task where it was difficult to distinguish between seven different types of attacks, the "curse of dimensionality" effect was minimized. Second, processing speed increased approximately fivefold. Third, the comprehensibility of the models was improved by focusing only on the most effective traffic features. Finally, the risk of overfitting was reduced by removing unnecessary data that created false relationships. The remaining 22 features constitute the cleanest dataset needed to capture the fundamental patterns that distinguish normal traffic from botnet attacks. This number retains sufficient information to accurately identify all seven categories.

Classification Results

In tests conducted on the N-BaIoT dataset, seven different machine learning algorithms were compared to detect botnet attacks. As shown in **Table 2**, ensemble learning methods, particularly gradient boosting algorithms, were more successful than others in this problem. These algorithms both distinguished attack types more clearly and produced outputs suitable for practical use by keeping the processing load at reasonable levels.

Table 2. Models performance

Model	Test accuracy	CV mean	CV SD
LightGBM	0.998153	0.999146	0.00058
XGBoost	0.996991	0.996998	0.000071
Random Forest	0.996874	0.996889	0.000092
Gradient Boosting	0.996866	0.996752	0.000274
Decision Tree	0.994834	0.921168	0.044626
Extra Trees	0.994544	0.993838	0.000763
KNN	0.832777	0.830706	0.000566

LightGBM: Light gradient boosting machine, XGBoost: Extreme gradient boosting, KNN: K-nearest neighbors, CV: Cross-validation, SD: Standard deviation

The LightGBM algorithm demonstrated the best classification performance. This algorithm achieved high success rates of 99.81% accuracy on the test data and 99.91% on the cross-validation average. LightGBM was followed by XGBoost 99.70%, Random Forest 99.69% and Gradient Boosting 99.68%. It is particularly noteworthy that all four algorithms achieved accuracy rates above 99%. Furthermore, the cross-validation standard deviations, ranging from 0.00058 to 0.000274, were quite low. This low deviation clearly indicates that the models have high generalization ability and are not affected by overfitting.

On the other hand, although the decision tree algorithm achieved a relatively high accuracy rate of 99.48% on the test data, the cross-validation average dropped to 92.12%. Moreover, the standard deviation is also quite high at 0.044626. These results indicate that the model is weak in terms of stability and generalization ability. Such behavior indicates that a single decision tree is sensitive to noise and variance in the dataset, while confirming that ensemble methods successfully overcome this problem. The KNN algorithm, on the other hand, showed the lowest performance with a test accuracy of 83.27%. This low performance supports the idea that distance-based algorithms have limited effectiveness on high-dimensional and large-scale datasets.

XAI Insights

Model interpretability is a critical requirement in cybersecurity applications. In this section, feature importance analysis was conducted to understand the decision-making mechanisms of the AI model and to identify which network features are the most decisive in detecting botnet attacks.

Feature importance consensus: As shown in **Figure 1** the top 20 network features identified by the LightGBM algorithm are presented. The feature importances were computed based on each feature's cumulative contribution to class discrimination during model training, using the split-based importance criterion.

The feature importance analysis of the LightGBM model reveals the most distinctive network traffic attributes for botnet attack detection. As illustrated in **Figure 2**, the visualization and output data indicate that MI based features contribute the most to the classification performance.

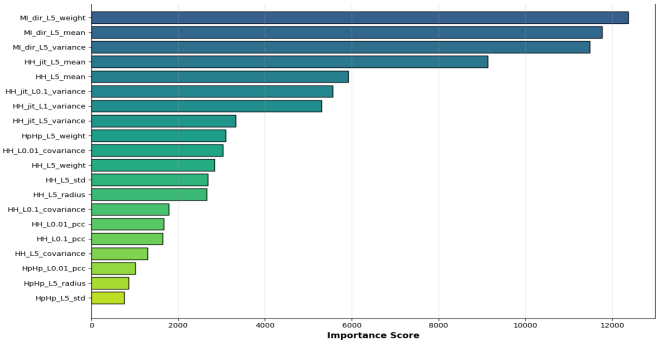


Figure 2. LIME-based feature importance for the LightGBM classifier

MI: MAC-IP, dir: Direction, HH: Host-to-host, jit: Jitter, HpHp: Host-port to host-port, L5: 1-minute time window, L0.1: 500-millisecond time window, L0.01: 100-millisecond time window, pcc: Pearson correlation coefficient, LIME: Local interpretable model-agnostic explanations, LightGBM: Light gradient boosting machine

In the analyses of the N-BaIoT dataset, the MI_dir_L5_weight feature showed the highest impact (importance score: 12.370). It was followed by MI_dir_L5_mean (11.772) and MI_dir_L5_variance (11.490). What is noteworthy is that all three features belong to the MI category and examine the last 5 packets (L5). This highlights how decisive short-term network traffic behavior and inter-packet information flow are in detecting botnet attacks. MI measurements successfully capture botnets' typical command-and-control mechanisms and coordinated attack patterns because they monitor communication between source and destination IP addresses. In other words, these features reflect the digital traces that show how attackers manage their "bot armies".

The second group of important features consists of jitter and statistical descriptors of the host-host (HH) and host-port-host-port (Hp-Hp) types. For example, metrics such as HH_jit_L5_mean (9.129) and HH_L5_mean (5.926) highlight how important indicators fluctuations in packet arrival times (jitter) are in characterizing botnet traffic. The temporal behavioral differences between normal user traffic and botnet traffic have been successfully captured, particularly through second-order statistical moments such as variance, standard deviation, and covariance. The multiple window sizes used (L0.01, L0.1, L1, L5) enabled the model to learn both micro- and macro-scale temporal patterns in network activity. The uneven distribution of feature importance scores, and particularly the clear dominance of the first three features, suggests that these metrics should be prioritized in the dimension reduction and feature selection processes. In conclusion, the high classification performance achieved by the LightGBM model can be attributed to the effective use of Mutual Information and temporal variation metrics. Together, these metrics enabled the model to capture the intricate dynamics of botnet communication behavior.

The LIME analysis in **Figure 3** makes the LightGBM model's decision on a single example understandable. Specifically, in black-box models, green-colored features support the prediction (increasing the prediction probability), while red ones work against the decision. For example, the HpHp_L5_weight > 0.00 feature alone increases the prediction probability by 30%. This shows that communication patterns at the host machine port level function as a "digital fingerprint" for botnet attacks. Conversely, the HH_jit_L5_mean feature being nearly zero (-0.12 effect) provides an important clue: Regular time intervals between packets indicate normal network behavior, not botnet traffic. Other red features similarly show how staying below or above certain threshold values affects the model. Another noteworthy point is that features highlighted by LIME, such as MI_dir_L0.01_weight and MI_dir_L5_mean, confirm the list of "most critical features" from previous analyses. In other words, the model not only makes accurate predictions but also provides intelligent justifications for each decision. This transparency is crucial in cybersecurity as it enhances reliability and auditability.

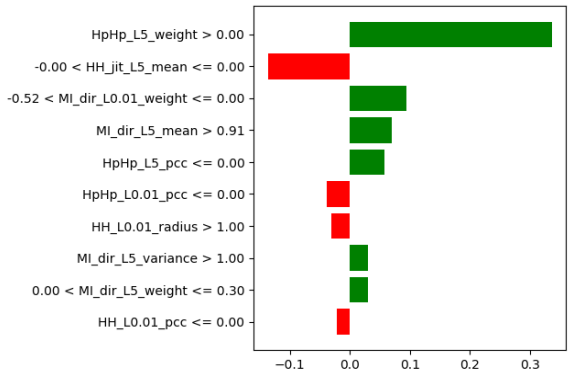


Figure 3. LIME-based local explanation of feature contributions for a representative Mirai test instance (ID: 121958). The visualization highlights both positive (green) and negative (red) influences on the LightGBM classifier's decision

MI: MAC-IP, dir: Direction, HH: Host-to-host, jit: Jitter, HpHp: Host-port to host-port, L5: Lambda (λ)=5 decay window, L0.01: Lambda (λ)=0.01 decay window, pcc: Pearson correlation coefficient, LIME: Local interpretable model-agnostic explanations, LightGBM: Light gradient boosting machine, ID: Identifier

Overall, the local interpretability provided by the LIME analysis shows that the model not only achieves high accuracy but also offers a logical and traceable rationale for each individual decision an aspect of great significance for reliability and auditability in cybersecurity applications.

SHAP analysis: SHAP values were used to examine how the LightGBM-based botnet detection model explains its decisions. This method is based on a solid mathematical foundation derived from cooperative game theory and clearly shows which features are most influential in the model. SHAP stands out as an ideal tool for making model decisions transparent and reliable in critical areas such as cybersecurity.

The SHAP analysis in **Figure 4** shows the ten most influential features used by the LightGBM model when predicting the "benign" traffic class and the average effect sizes of these features. Unlike LIME, which provides local-level interpretation, SHAP analysis allows us to understand the model's behavior on a global scale. It does this by reporting the mean absolute SHAP values across the entire dataset.

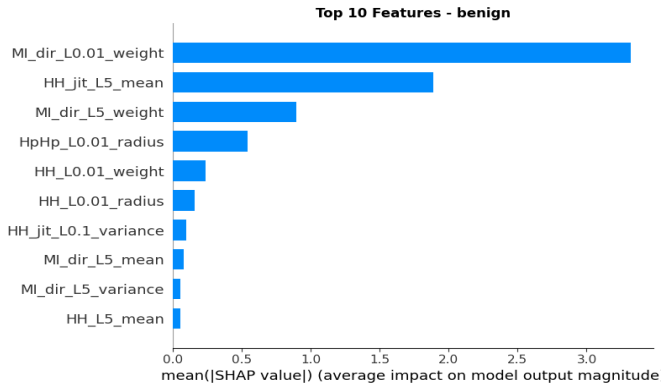


Figure 4. SHAP-based feature importance ranking for benign traffic classification

MI: MAC-IP, dir: Directional, HH: Host-to-host, jit: Jitter, HpHp: Host-port to host-port, L5: λ=5 decay factor, L0.1: λ=0.1 decay factor, L0.01: λ=0.01 decay factor, SHAP: SHapley additive exPlanations

The visualization shows that MI_dir_L0.01_weight is the most critical discriminative feature in detecting benign traffic, with a SHAP value of approximately 3.0. This metric represents the weighted average of directional mutual information within a very short time window (L0.01) and characterizes the regular and predictable communication patterns of normal user traffic.

The second most effective feature, HH_jit_L5_mean (average SHAP value approximately 2.0), shows that the average host-level delay (jitter) duration is a very important indicator in determining normal traffic. The relatively consistent arrival times and low delay values of packets in normal network traffic explain why this feature has such strong separation power. Although MI_dir_L5_weight (average SHAP value approximately 1.0) and other Mutual Information-based metrics have lower SHAP scores, this indicates that while these features are important across all classes, they are less decisive in specifically characterizing the benign class. Features with moderate SHAP values, such as HpHp_L0.01_radius, HH_L0.01_weight, and HH_L0.01_radius, indicate that spatial and statistical metrics at the host-port level play a supporting role in identifying benign traffic. SHAP's global perspective systematically reveals which features the model relies on for each class. This provides invaluable insights for feature engineering and model optimization. Furthermore, the mathematical consistency and aggregability of SHAP values ensure that the model's decisions can be explained in a transparent and auditable manner. This is a fundamental requirement in cybersecurity applications.

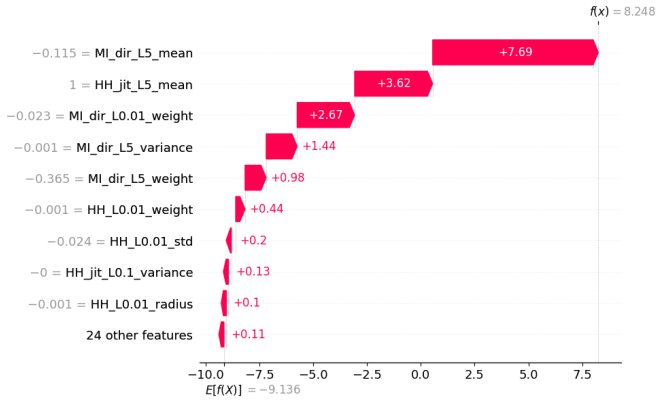


Figure 5. SHAP-based waterfall plot illustrating feature-level contributions for a representative prediction

MI: MAC–IP, dir: Directional, HH: Host-to-host, jit: Jitter, L5: 1-minute time window, L0.1: 500-millisecond time window, L0.01: 100-millisecond time window, SHAP: SHapley additive exPlanations

The SHAP waterfall diagram in **Figure 5** shows step-by-step how the LightGBM model makes a decision for a single example. Starting from the baseline value (E [f (x)]=-9.136), the diagram reaches the final result [f (x)=8.248] by summing the positive/negative effects of each feature.

In **Figure 5**, red-colored features support the prediction. MI_dir_L5_mean has the largest effect (+7.69). This feature strongly influences the model even though its actual value on the example (-0.115) is low. HH_jit_L5_mean (value=1, +3.62) ranks second. Inter-packet timing deviation (jitter) plays a critical role in identifying attack traffic. Other MI-based features (MI_dir_L0.01_weight, MI_dir_L5_variance, etc.) also confirm the global analysis results with positive contributions. Host-level statistics (HH_L0.01_weight, HH_jit_L0.1_variance, etc.) have small but significant effects, while the total contribution of the remaining twenty four features is only +0.11. This reveals that the model bases its decision on the top ten features.

6G Integration and Applications

6G wireless networks, expected to be commercialized around 2030 with standardization efforts beginning in 2025, are poised to bring significant change to the IoT ecosystem. 6G is designed as "Intelligent Networks of Everything." These networks will offer high speed and capacity, utilizing integrated Artificial Intelligence/machine learning (AI/ML) from core to edge. Thus, they will serve as a universal IoE, offering advanced functions in applications such as computing, sensing, positioning, and energy management (Khan Waqar, 2025). While this major transformation creates unique opportunities for IoT environments, it also brings serious security challenges. Therefore, the attack detection methods developed in this study are particularly important for this future ecosystem.

Our enhanced LightGBM-based botnet detection system can serve as a digital proxy in the dense IoT ecosystems of 6G networks. The terahertz frequencies offered by 6G, AI-driven management, and the simultaneous connection of billions of devices necessitate a shift in our security approaches.

Our model is ideal for real-time attack detection in these environments, with 99.82% accuracy and high stability (cross-validation standard deviation: 0.00058). SHAP and LIME analyses show how traffic metrics such as MI_dir_L5_weight and HH_jit_L5_mean distinguish botnets like Mirai-Gafgyt or DDoS-scanning attacks. Thanks to 6G's edge computing architecture, this model can be distributed to every corner of the network. Thus, dependency on central servers is eliminated, response time reduction is targeted, by working in conjunction with network slicing and software-defined networking (SDN), security rules specific to each service layer (eMBB, URLLC, mMTC) can be automatically applied. This allows the logic behind every "attack" alert to be traced. 6G operators can monitor decisions and trust the system. Because cyber defense now relies on evidence-based transparency, not speculation.

DISCUSSION

Security Imperatives in 6G-IoT Ecosystems

Protecting 6G IoT networks from threats, especially next-generation attacks, is one of the most significant challenges this infrastructure must overcome. This situation has necessitated the development of innovative architectures and paradigms enhanced by intelligence, software-defined approaches, and infrastructure virtualization (Asma Alotaibi, 2023). The N-BaIoT dataset used in this research includes real-world botnet attacks such as Mirai and Bashlite on commercial IoT devices. This reveals critical information about attack patterns that will persist and evolve in 6G-enabled environments. As Wireless Sensor Networks scale up within 6G frameworks, their exposure to cyber threats will rise sharply due to diverse communication protocols, dynamic node topologies, AI-enabled breaches, adversarial attacks targeting AI models, and resource-constrained device architectures.

The attack detection model developed in this study provides a foundation for the fundamental security requirements of future 6G networks.

In the IoT world of 6G, which is extremely dense and diverse, traditional security methods will no longer suffice. In this environment, active and AI-powered defense systems have become essential. The LightGBM-based detection system we present in our study has the capacity to meet the layered security needs of future 6G networks. This is because 6G networks will carry data at terabit speeds and are expected to have low response times. This will lead to botnet attacks becoming both more complex and faster. Existing systems will not be able to respond at this speed; defenses will have to be as fast and intelligent as the attacks.

Feature importance analysis results indicate that temporal mutual information and jitter metrics must be monitored in real time. In particular, metrics such as MI_dir_L5_weight (importance score: 12.370) and HH_jit_L5_mean (importance score: 9.129) have become a security requirement in this context. And also, as seen with SHAP analysis, the MI_dir_L0.01_weight feature (SHAP value approx. 3.0) plays a dominant role in detecting normal traffic. This role forms the basis for accurately characterizing the normal behavior profiles of 6G-IoT devices and enabling sensitive anomaly detection. Furthermore, eliminating highly correlated features (correlation > 85%) using the correlation filtering approach ensures that the model can be efficiently deployed on edge computing infrastructures.

Future Perspectives: 6G Security and XAI Integration

The critical role of explainable AI in 6G security: The integration of XAI techniques in this study represents a fundamental requirement for 6G network security rather than merely an enhancement. As 6G manages a wide range of mission-critical services, increased explainability to enable trust is critical, particularly as the migration from traditional mathematical model-dependent optimization to data-dependent deep learning decreases the insight and trust in optimization modules. The loss of model explainability creates vulnerabilities to adversarial attacks and regulatory compliance challenges.

Using XAI techniques in this study is one of the new research topics in 6G security. 6G will manage critical systems such as healthcare, energy, or transportation. The transition from traditional mathematical models to deep learning-based systems turns decision-making processes into a black-box. This creates security vulnerabilities and complicates legal oversight. For example, techniques such as LIME and SHAP were used for similar purposes in a 6G-IoT security study published in 2025 (Navneet Kaur, 2025). In our N-BaIoT-based analysis, XAI also allows experts to see "how each attack was detected." This capability is critical, especially for rapidly changing threats such as Mirai or DDoS. The biggest problem with black-box models was that they broke down the trust barrier between humans and machines. For it to be reliable, the logic behind every "attack" alert must be traceable. This is where our model comes into play: SHAP and LIME clarify the reasoning behind each prediction. This allows security teams to use automated systems not by questioning them, but by understanding them. This ensures control in high-risk 6G networks and allows humans to trust machines.

Application scenarios in 6G-enabled smart cities: The attack detection capabilities demonstrated in this study are applicable in emerging 6G use cases. 6G technology, combined with Wireless Sensor Networks, is poised to fundamentally transform smart city infrastructure (Khan Waqar, 2025). With ultra-fast communication, efficient real-time data transfers, and high reliability, 6G will be the preferred choice, particularly in areas such as smart transportation systems, environmental monitoring, and public safety. The botnet attacks analyzed in this study using the N-BaIoT dataset (particularly those targeting security cameras, thermostats, and other IoT devices) represent potential threat vectors that could spread to 6G-enabled smart city infrastructures.

Addressing 6G-specific challenges: The ability to detect zero-day attacks is a truly important area of research for 6G networks. Signature-based systems cannot detect these attacks; due to the complexity of 6G, experts prefer anomaly-based solutions (Iftikhar Rasheed, 2025). Our model learns "normal traffic" from N-BaIoT data and can detect deviations. This lays the foundation for detecting new types of attacks that will emerge as 6G technology matures. The distributed and heterogeneous nature of 6G requires security solutions that are both flexible and scalable. As proven in tests, our model works seamlessly even on IoT devices with limited resources. This makes it fully compatible with devices in 6G edge networks. Moreover, thanks to XAI integration, security teams in different regions can understand threats in real time and coordinate their response. This approach, which learns threat behavior rather than their source, shapes security in future smart networks through human-machine collaboration.

Future Works

The integration of the methodologies developed in this research with next-generation 6G technologies presents several promising avenues for future investigation. Adapting the proposed models trained on the N-BaIoT dataset to accommodate the substantially increased data throughput and ultra-low latency characteristics of 6G networks will necessitate the optimization of both detection algorithms and XAI interpretation mechanisms. Future research will focus on developing lightweight inference pipelines capable of maintaining detection accuracy while meeting the stringent real-time processing demands of 6G architectures.

Future studies will investigate the generalization capability of the proposed framework across diverse IoT botnet datasets, including BoT-IoT, IoT-23, and MedBIoT, to validate its robustness against heterogeneous attack patterns and network configurations.

Exploring federated learning approaches for distributed model training across edge devices without centralizing sensitive network traffic data constitutes another significant research direction. This approach would address privacy concerns while enabling collaborative model improvement across geographically distributed 6G IoT deployments.

Assessing the resilience of the proposed detection framework against adversarial evasion attacks and developing countermeasures to enhance model robustness in adversarial environments will be addressed in subsequent research.

CONCLUSION

The integration of the methods developed in this research with next-generation 6G technologies opens up many promising avenues for further research. First, adapting our models trained on N-BaIoT to manage the increased data speeds and reduced latency of 6G networks will require the

optimization of both detection algorithms and Explainable Artificial Intelligence (XAI) interpretation mechanisms. Second, integrating our XAI-enhanced detection capabilities with blockchain-based security frameworks could enhance data integrity and trust in 6G smart city applications.

The experimental results obtained using the N-BaIoT dataset in this study, when combined with the latest XAI techniques, form a foundation for ensuring the security of next-generation wireless networks. As 6G networks transition from theoretical designs to real-world deployment, rigorously tested attack detection and analysis methods will become indispensable. These techniques are essential for preserving security, safety, and reliability in network environments that are rapidly growing more complex and autonomous.

Emerging AI governance frameworks and regulatory standards increasingly mandate algorithmic transparency and explainability for automated security systems. The integration of SHAP and LIME methodologies in this study ensures compliance with these evolving requirements, facilitating the adoption of AI-driven security solutions in regulated industries such as healthcare, critical infrastructure, and financial services.

ETHICAL DECLARATIONS

Ethics Committee Approval

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study.

Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

Author Contributions

Concept: MG, EA, SB, NB; Design: MG, EA, SB, NB; Control: MG, EA, SB, NB; Resources: MG, EA, SB, NB; Materials: MG, EA, SB, NB; Data Collection and/or Processing: MG, EA, SB, NB; Analysis and/or Interpretation: MG, EA, SB, NB; Literature Review: MG, EA, SB, NB; Writing the Article: MG, EA, SB, NB; Critical Review: MG, EA, SB, NB.

Editor Note

Necaattin Barışçı is a member of the Editorial Board of this journal. He was not involved in the editor assignment, reviewer selection, peer review process, or editorial decision for this manuscript. The peer review process was managed independently by another editor.

REFERENCES

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

Abbasi, F., Naderan, M., & Alavi, S. E. (2025). Dimensionality reduction with deep learning classification for botnet detection in the internet of things. *Expert Systems with Applications, 267*, 126149. https://doi.org/10.1016/j.eswa.2024.126149

Alani, M. M., & Damiani, E. (2023). XRecon: an explainbale IoT reconnaissance attack detection system based on ensemble learning. *Sensors, 23*(11), 5298. https://doi.org/10.3390/s23115298

Ali, S., Ghazal, R., Qadeer, N., Saidani, O., Alhayan, F., Masood, A., ..., & Gupta, D. (2024). A novel approach of botnet detection using hybrid deep learning for enhancing security in IoT networks. *Alexandria Engineering Journal, 103*, 88-97. https://doi.org/10.1016/j.aej.2024.05.113

Alkahtani, H., & Aldhyani, T. H. H. (2021). Botnet attack detection by using CNN-LSTM model for internet of things applications. *Security and Communication Networks, 2021*, 3806459. https://doi.org/10.1155/2021/3806459

Alotaibi, A., & Barnawi, A. (2023). Securing massive IoT in 6G: Recent solutions, architectures, future directions. *Internet of Things, 22*, 100715. https://doi.org/10.1016/j.iot.2023.100715

Al-Othman, Z., Alkasassbeh, M., & Al-Haj Baddar, S. W. (2020). An efficient approach to detect IoT botnet attacks using machine learning. *Journal of High Speed Networks, 26*(3), 241-254. https://doi.org/10.3233/JHS-200641

Alsamiri, J., & Alsubhi, K. (2019). Internet of things cyber attacks detection using machine learning. *International Journal of Advanced Computer Science and Applications, 10*(12), 627-634. https://doi.org/10.14569/IJACSA.2019.0101280

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

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). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Do, P. H., Le, D. T., Vishnevsky, V. M., Berezkin, A., & Kirichek, R. (2024). A horizontal federated learning approach to IoT malware traffic detection: an empirical evaluation with N-BaIoT dataset. In 2024 26th International Conference on Advanced Communications Technology (ICACT) (pp. 1494-1506). IEEE. https://doi.org/10.23919/ICACT60172.2024.10471929

Emmanuel, E. J. (2025). Systematic review of 6G-IoT privacy risks, emerging threats, mitigation strategies, and cybersecurity. *Asian Journal of Advanced Research and Reports, 19*(9), 180-190. https://doi.org/10.9734/ajarr/2025/v19i91151

Garcia, S., Parmisano, A., & Erquiaga, M. J. (2020). IoT-23: a labeled dataset with malicious and benign IoT network traffic (Version 1.0.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4743746

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*, 3-42. https://doi.org/10.1007/s10994-006-6226-1

Guerra-Manzanares, A., Medina-Galindo, J., Bahsi, H., & Nõmm, S. (2022). Using MedBIoT dataset to build effective machine learning-based IoT botnet detection systems. In S. Furnell, P. Mori, E. Weippl, & O. Camp (Eds.), Information systems security and privacy: 6th International Conference, ICISSP 2020, revised selected papers (pp. 222-243). Springer. https://doi.org/10.1007/978-3-030-94900-6_11

Gummadi, A. N., Napier, J. C., & Abdallah, M. (2024). XAI-IoT: an explainable AI framework for enhancing anomaly detection in IoT systems. *IEEE Access, 12*, 71024-71054. https://doi.org/10.1109/ACCESS.2024.3402446

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.

Karanja, E. M., Masupe, S., & Jeffrey, M. G. (2020). Analysis of internet of things malware using image texture features and machine learning techniques. *Internet of Things, 9*, 100153. https://doi.org/10.1016/j.iot.2019.100153

Kaur, N., & Gupta, L. (2025). Securing the 6G-IoT environment: a framework for enhancing transparency in Artificial Intelligence decision-making through explainable Artificial Intelligence. *Sensors, 25*(3), 854. https://doi.org/10.3390/s25030854

Khan, W., Usama, M., Khan, M. S., Saidani, O., Al Hamadi, H., Alnazzawi, N., ..., & Ahmad, J. (2025). Enhancing security in 6G-enabled wireless sensor networks for smart cities: a multi-deep learning intrusion detection approach. *Frontiers in Sustainable Cities, 7*, 1580006. https://doi.org/10.3389/frsc.2025.1580006

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. https://doi.org/10.1016/j.future.2019.05.041

Krishna, E. S. P., & Thangavelu, A. (2021). Attack detection in IoT devices using hybrid metaheuristic lion optimization algorithm and firefly optimization algorithm. *International Journal of System Assurance Engineering and Management*, 1-14. https://doi.org/10.1007/s13198-021-01150-7

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765-4774.

Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Shabtai, A., ..., & Elovici, Y. (2018). N-BaIoT-Network-based detection of IoT botnet attacks using deep autoencoders. *IEEE Pervasive Computing, 17*(3), 12-22. https://doi.org/10.1109/MPRV.2018.03367731

Montgomery, D. C. (2017). Design and analysis of experiments (9th ed.). John Wiley & Sons.

Nazir, A., He, J., Zhu, N., Qureshi, S. S., Qureshi, S. U., Ullah, F., ..., & Pathan, M. S. (2024). A deep learning-based novel hybrid CNN-LSTM architecture for efficient detection of threats in the IoT ecosystem. *Ain Shams Engineering Journal, 15*(7), 102777. https://doi.org/10.1016/j.asej.2024.102777

Rallapalli, S., Tiwari, V. K., Sagar, S. D. V., & Pareek, P. K. (2023). Detecting the anomaly attacks in industrial IoT using hybrid feature selection based modified convolutional neural network. In 2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT) (pp. 1-7). IEEE. https://doi.org/10.1109/EASCT59475.2023.10393443

Rasheed, I., & Mostafa, H. (2025). Federated learning-based anomaly detection for zero-day attack prevention in 6G network slices. *Computer Networks, 268*, 111376. https://doi.org/10.1016/j.comnet.2025.111376

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Saied, M., & Guirguis, S. (2025). Explainable Artificial Intelligence for botnet detection in internet of things. *Scientific Reports, 15*, 7632. https://doi.org/10.1038/s41598-025-90420-6

Saied, M., Guirguis, S., & Madbouly, M. (2023). A comparative analysis of using ensemble trees for botnet detection and classification in IoT. *Scientific Reports, 13*, 21632. https://doi.org/10.1038/s41598-023-48681-6

Santhadevi, D., & Janet, B. (2023). Stacked deep learning framework for edge-based intelligent threat detection in IoT network. *The Journal of Supercomputing, 79*(11), 12622-12655. https://doi.org/10.1007/s11227-023-05153-y

Savant, R. V., M, S., Mohan, P. K., Bhaskaran, S., & Rajagopal, S. M. (2024). A comparative analysis of black-box and white-box models for IoT botnet detection. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). IEEE. https://doi.org/10.1109/I2CT61223.2024.10544270

Serhane, A., Ibrahimi, K., Hamzaoui, E.-M., Jouhari, M., & Ben-Othman, J. (2025). IoT intrusion detection using machine learning classifiers and PCA dimensionality reduction for N-BaIoT dataset. In ICC 2025-IEEE International Conference on Communications (pp. 5809-5814). IEEE. https://doi.org/10.1109/ICC52391.2025.11161305

Son, B. D., Hoa, N. T., Chien, T. V., Khalid, W., Ferrag, M. A., ..., & Debbah, M. (2024). Adversarial attacks and defenses in 6G network-assisted IoT systems. *IEEE Internet of Things Journal, 11*(11), 19168-19187. https://doi.org/10.1109/JIOT.2024.3373808

Steinbach, M., & Tan, P.-N. (2009). kNN: K-nearest neighbors. In X. Wu & V. Kumar (Eds.), The top ten algorithms in data mining (pp. 151-161). Chapman & Hall/CRC. https://doi.org/10.1201/9781420089653-15

Trajanovski, T., & Zhang, N. (2021). An automated and comprehensive framework for IoT botnet detection and analysis (IoT-BDA). *IEEE Access, 9*, 124360-124383. https://doi.org/10.1109/ACCESS.2021.3110188

Ullah, S., Wu, J., Lin, Z., Kamal, M. M., Mostafa, H., ..., & Chuah, T. C. (2025). Comparative analysis of deep learning and traditional methods for IoT botnet detection using a multi-model framework across diverse datasets. *Scientific Reports, 15*, 31072. https://doi.org/10.1038/s41598-025-16553-w

Wang, D., Zhang, Y., & Zhao, Y. (2017). LightGBM: an effective miRNA classification method in breast cancer patients. In Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics (pp. 7-11). Association for Computing Machinery. https://doi.org/10.1145/3155077.3155079