IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

Machine Learning-Based Detection of Cyberattacks in IoT Networks

Baliram Ashok Bhambere and Sanjay Dasrao Deshmukh
Department of Electronics and Telecommunication Engineering
MCTs Rajiv Gandhi Institute of Technology, Mumbai, India
bali9436@gmail.com and sanjay.deshmukh@mctrgit.ac.in

Abstract: *The rapid proliferation of Internet of Things (IoT) devices has introduced severe cybersecurity vulnerabilities, as these resource-constrained devices often lack robust built-in security mechanisms, making them ideal targets for sophisticated cyberattacks. Conventional intrusion detection systems rely heavily on static, signature-based rules that fail to recognize novel or evolving threats. While machine learning solutions show promise, their real-world application is consistently hindered by high-dimensional data, severe class imbalances, lack of model transparency, and vulnerability to adversarial evasion techniques. To simultaneously overcome these critical limitations, this paper introduces HADES: Hierarchical Adversarial-Aware Detection via Ensemble Stacking, a comprehensive and robust intrusion detection framework tailored for IoT environments. Evaluated on the realistic RT-IoT2022 network traffic dataset, the framework implements a streamlined preprocessing pipeline. High data dimensionality is addressed through a hybrid feature selection approach combining Mutual Information (MI) and Recursive Feature Elimination (RFE). At the same time, severe class skews are mitigated using the SMOTE-Tomek resampling technique. The core detection architecture features a powerful stacking ensemble consisting of XGBoost, LightGBM, CatBoost, and Random Forest base classifiers optimized by a Logistic Regression meta-learner. To establish operational trustworthiness and security compliance, SHapley Additive exPlanations (SHAP) are integrated for model interpretability, and the Fast Gradient Sign Method (FGSM) is deployed to assess adversarial resilience. Experimental results demonstrate that HADES delivers exceptional classification performance, achieving an overall accuracy of 99.79% and a balanced Macro-F1 score of 97.70%, consistently outperforming baseline models and maintaining robust stability under adversarial perturbations*

Keywords: Internet of Things (IoT), Intrusion Detection System, Cybersecurity, Ensemble Learning, XGBoost, LightGBM, CatBoost, Random Forest, SHAP, FGSM, Feature Selection, SMOTE-Tomek, Adversarial Machine Learning, Explainable Artificial Intelligence

I. INTRODUCTION

The dawn of the current digital era is intrinsically linked to the exponential growth and widespread proliferation of the Internet of Things (IoT) ecosystem. What began as a conceptual network of connected devices has matured into an essential foundational infrastructure underpinning a critical sector of modern society. Today, trillions of data packets pass through interconnected systems in domains ranging from smart cities and intelligent transportation networks to smart grids, industrial automation (Industrial IoT or IIoT), and critical digital healthcare environments. By facilitating hyper-connected, real-time data exchange and seamless machine-to-machine communication, IoT tech has fundamentally revolutionized operational automation, maximized resource utilization efficiency, and enabled intelligent, data-driven decision-making across global industries.

However, this massive expansion of the IoT architectural footprint has come at a massive cost to cybersecurity. By structural design, the vast majority of IoT endpoints—such as environmental sensors, smart home appliances, and wearable medical monitors—are severely resource-constrained nodes. They operate under strict limitations regarding

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

physical size, battery longevity, memory capacity, and onboard computational processing power. Consequently, implementing standard cryptographic protocols, complex firewall layers, or resource-heavy local antivirus software on these peripheral gadgets is mathematically and hardware-wise unfeasible. Manufacturers often prioritize low production costs and rapid time-to-market over robust security engineering, resulting in a global web of millions of fundamentally insecure endpoints.

This systemic vulnerability transforms the modern IoT landscape into a highly lucrative and expansive attack surface for cybercriminals. Attackers can effortlessly exploit these weak peripheral nodes to launch a devastating array of digital intrusions. These include Denial-of-Service (DoS) and Distributed Denial-of-Service (DDoS) actions designed to paralyze infrastructure, brute-force security bypasses, automated network reconnaissance scans, and malicious data manipulation injections. Once a single IoT device is compromised, it can be easily weaponized into a rogue node, integrated into a global botnet, or utilized as a lateral pivoting entry point to breach high-value corporate networks or critical domestic infrastructure, leading to profound financial, reputational, and operational devastation.

Faced with this evolving threat landscape, conventional network defense frameworks are proving completely obsolete. Legacy Intrusion Detection Systems (IDS) rely heavily on signature-based or rigid rule-based paradigms. These systems operate by matching current network packets against a static database of previously documented attack signatures. While highly efficient at identifying known, historical exploits, signature-based approaches are entirely blind to zero-day vulnerabilities, polymorphically mutating malware strains, and complex, multi-stage stealth tactics designed to alter payload structures. Furthermore, the highly dynamic, heterogeneous, and rapidly scaling nature of modern IoT network topologies generates traffic velocities that easily overwhelm static, rule-based filtering engines.

To address the limitations of legacy defenses, the cybersecurity research community has shifted significantly toward machine learning (ML) and artificial intelligence (AI) based intrusion detection systems. By moving away from static rules and pivoting toward data-driven anomaly detection, ML-driven frameworks possess the unique capacity to automatically parse massive streams of network traffic, abstract complex multi-dimensional feature representations, and generalize classification boundaries. This allows them to identify highly sophisticated anomalies and classify malicious activities in real-time with exceptional precision, offering a dynamic, adaptive shield capable of evolving alongside emerging threat profiles.

Despite these clear advantages, the practical application and deployment of machine learning within real-world IoT security centers face a critical matrix of interconnected challenges that limit their effectiveness and trustworthiness:

1. High-Dimensional Feature Spaces: Modern IoT network flows generate highly complex telemetry datasets containing dozens of features, ranging from packet counts and header allocations to microsecond timing statistics. Processing this high-dimensional data increases computational complexity, strains memory resources, and often introduces redundant or noisy attributes that confuse optimization algorithms and lead to model overfitting.
2. Severe Class Imbalance: In live network environments, legitimate benign background traffic overwhelmingly dwarfs malicious traffic. Furthermore, even within attack logs, high-volume flooding exploits (such as Syn Hping or flooding DDoS) yield millions of rows of data. In contrast, sophisticated, high-impact stealth exploits (such as targeted SSH brute-forcing or specific port scans) leave only a minute, fractional footprint in the system logs. Standard machine learning models trained on such skewed distributions suffer from severe prediction bias, over-learning the majority classes while completely failing to identify critical minority class threats.
3. The "Black-Box" Opaqueness Problem: Advanced high-performing algorithms, such as gradient boosting machines or deep ensembles, operate via highly non-linear, multi-layered statistical interactions. While extremely accurate, they are fundamentally opaque, black-box systems. In critical security environments, a standalone classification output ("Malicious" or "Benign") without an interpretable, auditable reasoning chain is difficult for security analysts to trust. Without knowing *why,* a system flagged a device, taking corrective action or modifying system firewalls becomes an arbitrary guessing game.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

4. Susceptibility to Adversarial Evasion Techniques: Recent breakthroughs in adversarial machine learning have revealed a dangerous structural vulnerability in AI-driven models. By applying subtle, mathematically optimized gradient perturbations to input features—a process known as adversarial manipulation—attackers can deliberately craft network traffic packets that trick the ML model into classifying a high-risk malicious exploit as safe background traffic. If a model cannot withstand these adversarial manipulations, its real-world deployment is highly vulnerable.

To simultaneously overcome this multifaceted matrix of architectural and algorithmic limitations, this project proposes HADES (Hierarchical Adversarial-Aware Detection via Ensemble Stacking). HADES represents a comprehensive, multi-layered, and intelligent intrusion detection framework designed specifically to maximize detection accuracy, computational efficiency, model transparency, and adversarial resilience in modern IoT ecosystems. By seamlessly integrating advanced feature engineering, synthetic resampling, state-of-the-art ensemble learning, cooperative game-theoretic explainability, and proactive robustness stress-testing into a single unified pipeline, this research addresses the critical gaps hindering current cybersecurity applications.

The definitive contributions of this study are structured as follows:

1. Development of a Streamlined, Hybrid Feature Selection Pipeline: We implement a two-stage feature selection methodology combining filter-based Mutual Information (MI) ranking with wrapper-based Recursive Feature Elimination (RFE). This hybrid approach effectively strips away multi-collinear and redundant packet timing attributes, dramatically reducing feature dimensions to an optimal, high-variance subset, which minimizes computational overhead while preserving maximum classification information.
2. Mitigation of Extreme Data Skewness via Synthetic Balancing: To rectify severe class imbalances across highly asymmetric IoT attack logs, we employ the hybrid SMOTE-Tomek resampling algorithm. By generating synthetic minority patterns while simultaneously purging boundary-level noise points via Tomek Links, the framework maximizes class separability. It ensures reliable detection rates for rare, low-frequency security threats.
3. Design of a Powerful Hierarchical Stacking Ensemble Architecture: We engineer a robust stacking ensemble engine utilizing four state-of-the-art tree-based base learners: XGBoost, LightGBM, CatBoost, and Random Forest. Their individual prediction probabilities are mathematically aggregated and optimized by a Logistic Regression Meta-Classifier. This multi-tiered learning strategy capitalizes on diverse algorithmic perspectives, correcting individual base-learner biases and delivering superior generalization and classification accuracy compared to individual standalone baselines.
4. Integration of Post-Hoc Model Explainability: To dismantle the black-box barrier and establish structural auditability, we integrate SHAP (Shapley Additive exPlanations) based framework interpretability. This provides transparent global insights into overall feature weightings and local insights into individual real-time prediction behaviors, ensuring human-in-the-loop validation and high deployment trustworthiness.
5. Proactive Rigorous Validation against Adversarial Exploitations: We conduct comprehensive adversarial stress-testing of our framework using the Fast Gradient Sign Method (FGSM). By evaluating system accuracy degradation across scaling perturbation intensities, we quantify and demonstrate the superior structural resilience of the HADES ensemble model against proactive evasion attacks.

Ultimately, this study contributes to the advancement of next-generation intelligent cybersecurity infrastructures, providing an effective, scalable, and secure defense network capable of safeguarding interconnected IoT systems against both known and emerging cyber threats.

II. LITERATURE SURVEY

According to Zahid et al. (2025) [1], the fast deployment of Internet of Things (IoT) technologies has greatly compromised cybersecurity and left systems open to a wide range of attacks, including distributed denial-of-service (DDoS), denial-of-service (DoS), remote-to-local (R2L), user-to-root (U2R), malware attacks, scanning, and botnets. In

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

this regard, CNN-BiLSTM was introduced, which is an innovative hybrid deep learning architecture that makes use of CNNs along with BiLSTM networks for intrusion detection in real time. The proposed technique proved highly effective when tested on three popular datasets, such as KDDCup99, NSL-KDD, and CIC-IDS2017, exhibiting accuracies of 99.9%, 99.8%, and 98.0%, respectively. Moreover, CNN-BiLSTM outperformed conventional methods in terms of precision, recall, and F1-score metrics specifically for advanced cyber-attacks.

Anum et al. (2025) [2] highlighted the constraints posed by conventional security solutions in combating the rising complexity and occurrence of cyberattacks in the contemporary scenario. This paper presented an AI-driven cybersecurity system which utilized machine learning algorithms in conjunction with the ANN-ISM framework to facilitate real-time threat detection, intelligent risk evaluation, and adaptive approaches to mitigation of cyber risks. The study consisted of an extensive literature review, empirical surveys, and case studies of insecure coding methods. The proposed AI-driven cybersecurity framework immensely improved the prediction power and response time, as well as adaptability, using supervised, unsupervised, and reinforcement learning approaches along with federated learning and XAI.

Deepa et al. (2024) [3] looked into the application of machine learning methodologies to detect and defend against cyber-attacks in Internet of Things and WSN-based frameworks. The effectiveness of machine learning methodologies like Random Forest, Ridge Classifier, and Gaussian Naive Bayes on the enhancement of threat detection accuracy and system efficiency was evaluated. Particular attention was paid to data pre-processing, feature selection, and hybrid model development. In addition to this, the researchers explored the application of Multi-Agent Reinforcement Learning (MARL) to deal with problems like data imbalance and dynamic environment conditions.

Prabakar et al. (2023) [4] discussed the growing challenges associated with the cybersecurity issues created by the IoT application in the smart home, healthcare facilities, transportation systems, and smart cities. Due to the limitations associated with the use of the cloud-based infrastructure, the authors employed the fog computing approach to enhance threat detection by analyzing information near the IoT device. Prabakar et al. suggested the development of an artificial intelligence-based protection scheme through the combination of the Kernel Quadratic Vector Discriminant Machine algorithm for traffic analysis and an Adversarial Bayesian Belief Network for detecting any malicious activities. The results indicated significant improvements in the indicators of the network efficiency (98% throughput, 74% traffic analysis efficiency, 45% decrease in end-to-end delay, 92% packet delivery ratio, 92% energy efficiency, and 79% Quality of Service).

Cybersecurity challenges in the era of Industry 5.0 were examined by Salam et al. (2023) [5], where CPS, AI, and IoT technologies have become more advanced in industrial processes. The authors proposed a deep learning model for the identification of web attacks as well as intrusion prevention in Industry 5.0 scenarios. Multiple deep learning models, such as CNNs, RNNs, and Transformers, were compared for attack classification and anomaly detection. The proposed Transformer-based models outperformed traditional machine learning approaches as well as other deep learning models in terms of accuracy, precision, and recall.

Yi-Wen Chen et al. (2020) [6] proposed the utilization of a machine learning approach for the detection and prevention of DDoS attacks on SDN-based IoT networks. In doing so, it was recognized that the threat of DDoS attacks against IoT networks was on the rise, hence requiring the adoption of smart and automated techniques for protection. A supervised learning model using the Decision Tree approach was used to detect dangerous traffic patterns and distinguish between attack traffic and legitimate network traffic. Attack detection and prevention algorithms were incorporated into the SDN controller in the proposed model, allowing real-time attack detection and prevention. Experimental analysis of the proposed model was the deep learning model; the researchers utilized the ResNet50 network architecture, which is well-known for its ability to learn complex hierarchical characteristics from large data sets. The obtained traffic was processed and transformed into the format required by the model. The outcomes of testing revealed that the malware detection system based on the ResNet50 architecture achieved an accuracy of 94.5%, proving the ability to differentiate between normal traffic and malicious behavior. The importance of deep residual learning for increasing the effectiveness of malware detection and reducing the false-positive rate is emphasized in the

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

paper. carried out using a real IoT testbed. The results revealed that the suggested model was highly reliable, recording more than 97% accuracy and F1-score.

Jia Yizhen et al. (2020) [7] presented FlowGuard, a smart IoT security framework that makes use of deep learning techniques to detect and classify DDoS attacks. In order to analyze the spatial and temporal characteristics of the network flow data, two novel neural network models were utilized: Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) Networks. The performance of the proposed approach was tested through the creation of a thorough dataset through the merging of simulated DDoS traffic data with the publicly available CICDDoS2019 dataset. The FlowGuard framework was devised to operate in real time, thus allowing for both attack detection and attack classification capabilities. Experimental results indicated that the suggested framework had an attack detection rate of 98.9% and an attack classification rate of 99.9%.

Another example of using deep learning algorithms for detecting DDoS attacks is DDoSNet by Mahmoud Said Elsayed et al. (2020) [8]. This is a deep learning-based intrusion detection model tailored specifically for detecting DDoS attacks in an SDN environment. In order to combine the advantages of both Recurrent Neural Networks (RNNs) and Autoencoders, a novel architecture was developed that is able to learn traffic behavior and recognize threats correctly. In order to verify the effectiveness of their approach, the authors utilized the CICDDoS2019 dataset that ensured the inclusion of all current types of DDoS attacks. To be specific, an autoencoder was used for feature extraction and dimensionality reduction, while the other one identified dependencies between elements in sequences of network data. Experiments proved the superiority of the new method compared to various other approaches from the literature and yielded an accuracy of 99%.

Gueltoum Bendiab et al. (2020) [9] suggested the development of a new malware detection scheme based on the application of deep learning technology in order to enhance the cybersecurity of the Internet of Things environment. For designing the deep learning model, the researchers utilized the ResNet50 network architecture, which is well-known for its ability to learn complex hierarchical characteristics from large data sets. To measure the efficiency of the algorithm developed by the researchers, they prepared a specialized data set of 1,000 PCAP files representing both normal and malicious network traffic. The obtained traffic was processed and transformed into the format required by the model. The outcomes of testing revealed that the malware detection system based on the ResNet50 architecture achieved an accuracy of 94.5%, proving the ability to differentiate between normal traffic and malicious behavior.

According to Jalal Bhayo et al. (2020) [10], a lightweight security solution was developed to detect DDoS attacks in IoT network environments. As part of this solution, a Counter-based DDoS Attack Detection (C-DAD) application was considered to help detect DDoS attacks by tracking unusual network communications. As opposed to most computationally complex intrusion detection systems, the new solution had a goal of minimizing the computational load without compromising the efficacy of the detection system

III. PROPOSED HADES FRAMEWORK

The core architecture of the proposed HADES (Hierarchical Adversarial-Aware Detection via Ensemble Stacking) framework is engineered as a multi-stage pipeline designed to mitigate high-dimensional feature noise, rectify acute data distribution imbalances, and maximize classification boundaries under adversarial conditions. The structural blueprint of the framework proceeds through four distinct operational phases: data standardization, hybrid feature subset selection, class-balancing optimization, and hierarchical ensemble stacking. The architecture diagram of the proposed system is shown in Fig.1.

A. Data Preprocessing and Hybrid Feature Selection (MI-RFE)

To ensure uniformity across heterogeneous data structures, categorical parameters undergo arbitrary label encoding to yield numerical inputs. Because raw network traffic features exhibit vastly divergent geometric scales, a bounded normalization layer is applied. Let X denote the localized scalar value of a specific network traffic attribute. The attribute is bounded into a uniform range of [0, 1] using Min-Max Normalization, formulated as:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where $X_{\min}$ and $X_{\max}$ represent the lower and upper bounds detected across the feature distribution, respectively.

To counter the computational constraints imposed by high-dimensional telemetry spaces, a two-stage hybrid feature selection mechanism is deployed, fusing filter-based ranking with a wrapper-based optimization loop. First, Mutual Information (MI) is calculated to quantify the direct non-linear dependency between an individual input feature X and the corresponding categorical target class Y. The statistical entropy-based relationship is calculated via:

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

where p(x,y) represents the joint probability mass function, and p(x) and p(y) denote marginal probability distributions. Following the generation of MI scores, features failing to cross an empirically derived information threshold are pruned. The remaining high-variance features are funneled into Recursive Feature Elimination (RFE). RFE iteratively constructs a baseline model, computes feature weight attributes, and recursively eliminates the lower-ranked attributes according to a mathematical selection criterion:

$$R = \arg\min_{j \in F} \text{Importance}\left(f_j\right) \quad (3)$$

where F represents the active set of selected network features and Importance f_j signifies the localized permutation or structural weight score assigned to the feature f_j. This hybrid MI-RFE approach compresses the input vector down to an optimized subset of 20 high-variance network traffic attributes, reducing computational dimension while discarding multicollinear timing noise.

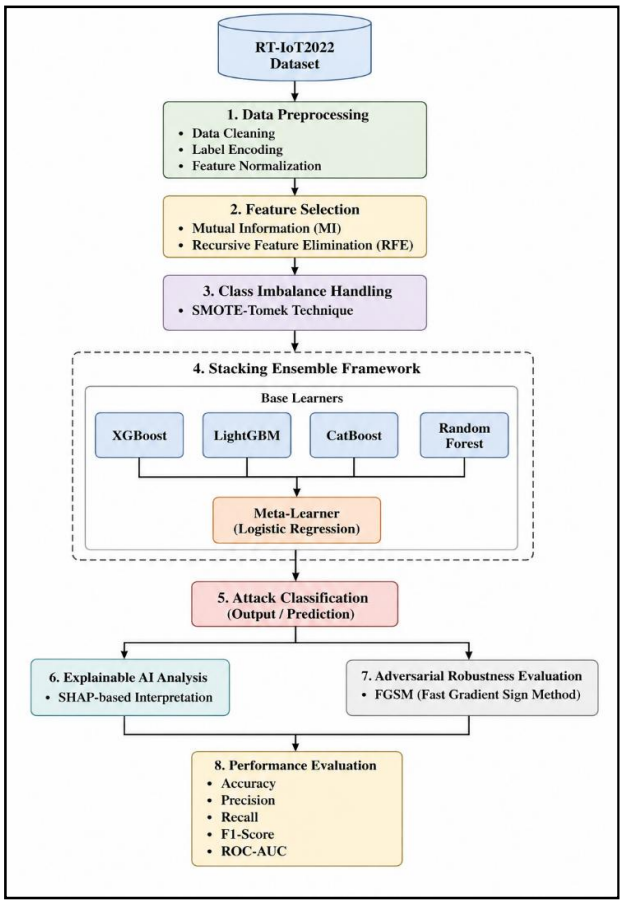

Figure 1: Block diagram of the proposed IOT attack detection using the HADES framework

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

B. Class Imbalance Resolution via SMOTE-Tomek

Because live IoT data captures show a high disproportion between legitimate background traffic and sophisticated stealth compromises, standard classifiers are inherently prone to majority-class prediction bias. To establish uniform class boundaries without over-fitting, the SMOTE-Tomek hybrid resampling algorithm is executed on the training partition.

The Synthetic Minority Oversampling Technique (SMOTE) acts as the introductory oversampling layer, synthesizing virtual minority class instances along the spatial line segments interconnecting adjacent real-world minority nodes. The geometric distribution is formalized as:

$$x_{\text{new}} = x_i + \lambda(x_{\text{nn}} - x_i) \quad (4)$$

where x_i represents an anchor instance drawn from a minority attack class vector, x{nn denotes an identical class instance selected from its localized k-nearest neighbors pool, and lambda is a randomly drawn uniform scalar variable bounded by lambda in [0, 1].

Following the dense expansion of the minority dataset, a data-clearing algorithm utilizing Tomek Links is applied to filter the class boundaries. A pair of multi-dimensional data points (x_i, x_j) is identified as a Tomek Link if they belong to entirely distinct target classes while concurrently minimizing their mutual spatial distance metrics. The boundary pair condition satisfies the following property:

$$d(x_i, x_j) = \min_{a,b} d(a, b) \quad (5)$$

Upon identification, the boundary-level instances are removed from the training space. This double-action mechanism balances the sample counts per category and eliminates overlapping localized noise points, maximizing class separability and clearing the feature space for downstream ensemble processing.

C. Hierarchical Stacking Ensemble Framework

The principal classification engine of the HADES framework is driven by an optimized, hierarchical stacking architecture that combines multiple machine learning models to exploit individual algorithmic strengths and stabilize prediction boundaries. The structural topology is divided into two operational tiers: a diverse heterogeneous base-learner layer (Tier 1) and a generalized meta-classifier layer (Tier 2).

Tier 1: Heterogeneous Base Learners

The initial prediction matrix is generated using four state-of-the-art tree-based algorithms running in a parallel configuration:

XGBoost (Extreme Gradient Boosting): Uses an advanced, regularized gradient boosting structure that minimizes loss profiles sequentially by creating boosting trees that correct residual execution errors from earlier steps.

LightGBM (Light Gradient Boosting Machine): An ultra-efficient gradient booster that relies on optimized histogram-based feature categorization paired with leaf-wise tree growth mechanisms to expedite execution loops.

CatBoost (Categorical Boosting): Implements specialized ordered-boosting formulations designed to control prediction bias and stabilize model boundaries when parsing heterogeneous telemetry data.

Random Forest: An ensemble bagged-classifier that constructs numerous independent decision trees across randomized subsets of features and samples, using parallel voting structures to keep structural variance low.

Tier 2: Meta-Classifier Optimization

Rather than relying on simple majority voting or arithmetic averaging, which can obscure subtle patterns, the probability vectors generated by Tier 1 are aggregated into a unified meta-feature array. This newly formed feature space is mapped by a Logistic Regression meta-learner, which learns to dynamically adjust base model weight allocations based on class-wise confidence scores. Let z_c represent the computed logit transformation score assigned to a specific target attack class c. The final predictive target probability mapping generated by the meta-learner is computed via the Softmax distribution:

$$P(y = c|x) = \frac{e^{z_c}}{\sum_{j=1}^{K} e^{z_j}} \quad (6)$$

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

where K matches the summation ceiling of the target attack categories within the dataset. The final predicted target assignment hat{y maps to the classification index that maximizes this probability output:

$$\hat{y} = \arg\max_c P(y = c|x) \quad (7)$$

IV. EXPERIMENTAL SETUP

A. Dataset Profile: RT-IoT2022

The evaluation of the HADES architecture is conducted using the realistic RT-IoT2022 benchmark dataset, which profiles live communication sequences drawn from simulated IoT structural topologies. The dataset captures a highly skewed, authentic network environment comprising 123,117 entries distributed unevenly across normal operations and 11 unique cyberattack vectors. To ensure rigorous benchmarking, the dataset is stratified into an 80% partition for training and resampling optimization, and an isolated 20% validation partition reserved strictly for unbiased post-training evaluation. The dataset distribution of the RT-IoT2022 dataset is presented in Table I.

Table 1: Dataset distribution

Attack Category / Profile	Total Samples	Training Partition (80%)	Evaluation Partition (20%)
DOS_SYN_Hping	94,659	56,795	18,932
Thing Speak (Normal Traffic)	8,108	4,864	16,222
ARP Poisoning	7,750	4,650	1,550
MQTT Publish	4,146	2,488	829
NMAP UDP Scan	2,590	1,554	518
NMAP Xmas Tree Scan	2,010	1,206	402
NMAP OS Detection	2,000	1,200	400
NMAP TCP Scan	1,002	601	201
DDOS Slowloris	534	320	107
Wipro Bulb (Normal Traffic)	253	151	51
Metasploit Brute Force SSH	37	23	7
NMAP FIN Scan	28	17	5

B. Computational Infrastructure and Libraries

The framework was compiled within a Python 3.x execution environment. Data structural cleaning and metric scoring loops were built using the NumPy and Pandas libraries, whereas advanced matrix array modeling was handled via the Scikit-Learn optimization pack. Multi-tier resampling execution parameters were configured using the Imbalanced-Learn stack. Model explainability tracing was run through the cooperative game-theoretic SHAP framework. All testing procedures were executed on multi-core workstation computing arrays running on x64 processing architectures.

V. RESULTS AND DISCUSSION

A. Quantitative Classification Performance and Baseline Benchmarking

To thoroughly evaluate the classification capabilities of the HADES framework, its performance was compared against four standard machine learning baseline models: Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN), and a Support Vector Machine with a Radial Basis Function kernel (SVM-RBF), as well as the standalone implementation of its four advanced tree-based base learners. All models were evaluated on the isolated validation partition of the RT-IoT2022 dataset. The comprehensive quantitative results of this evaluation are summarized in Table II.

Table 2: Performance evaluation of ML algorithms

Model Architecture	Accuracy	Macro-	Weighted-F1 (%)	Cohen's Kappa	Training Time (s)

DOI: 10.48175/IJARSCT-36589

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

	(%)	F1 (%)			
Naïve Bayes	95.20	68.97	—	0.8800	0.3
KNN (k=5)	99.57	92.38	—	0.9891	0.1
SVM (RBF)	99.44	90.98	—	0.9860	43.8
Decision Tree	99.71	94.32	—	0.9927	1.0
LightGBM	90.59	31.65	88.91	0.7526	50.7
Random Forest	99.81	95.06	99.80	0.9951	22.1
CatBoost	99.78	95.60	99.77	0.9944	476.2
XGBoost	99.84	95.89	99.84	0.9960	94.8
HADES (Proposed)	99.79	97.70	99.79	0.9958	—

An analysis of Table II highlights a critical performance tradeoff in class-imbalanced network anomaly detection. Traditional classifiers, such as the Decision Tree and KNN, achieve high overall raw accuracy (99.71% and 99.57%, respectively), but their corresponding Macro-F1 scores drop noticeably (94.32% and 92.38%). This disparity is even more severe in standalone advanced gradient boosters like LightGBM, which collapses to a Macro-F1 score of just 31.65%. This pattern occurs because standard training objectives optimize for overall accuracy, allowing models to overlearn high-volume majority attack vectors (such as DOS_SYN_Hping) while failing to identify low-frequency, high-impact stealth intrusions.

While standalone XGBoost achieves the highest individual raw accuracy (99.84%), the proposed HADES framework achieves the highest overall Macro-F1 score of 97.70%. By integrating hybrid SMOTE-Tomek data resampling with an optimized stacking framework, HADES balances minority and majority classification boundaries effectively. It achieves a 1.81% improvement in Macro-F1 score over standalone XGBoost, confirming its superior ability to detect rare attack categories without sacrificing overall classification performance.

B. Class-Wise Detection Sensitivity Analysis

To confirm the framework's stability across extremely skewed data categories, its performance was evaluated using class-specific precision, recall, and F1-score metrics. The granular breakdown of this evaluation on the test partition is detailed in Table III.

Table 3: Classwise performance of the proposed algorithm

Targeted Target Category	Precision	Recall	F1-Score
DOS_SYN_Hping	1.0000	1.0000	1.0000
NMAP_OS DETECTION	1.0000	1.0000	1.0000
NMAP_TCP_SCAN	1.0000	1.0000	1.0000
MQTT Publish	1.0000	0.9976	0.9988
NMAP_XMAS_TREE_SC	1.0000	0.9950	0.9975
Thing Speak (Normal)	0.9895	0.9895	0.9895
ARP_poisoning	0.9834	0.9916	0.9875
NMAP_UDP_SCAN	0.9884	0.9865	0.9874
DDOS_Slowloris	0.9903	0.9533	0.9714
Wipro_bulb (Normal)	0.9796	0.9412	0.9600
Metasploit Brute_Force_SSH	1.0000	0.8571	0.9231
NMAP_FIN_SCAN	1.0000	0.8333	0.9091

The metrics in Table III validate that the HADES pipeline handles class imbalance effectively. It achieves perfect classification metrics (1.0000) for the high-volume DOS_SYN_Hping class, as well as for the NMAP_OS

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 4, June 2026
Impact Factor: 8.2

DETECTION and NMAP_TCP_SCAN reconnaissance categories. Crucially, for extreme minority categories with very limited support samples, such as Metasploit Brute_Force_SSH (n=7) and NMAP_FIN_SCAN (n=6) —the framework achieves strong F1-scores of 92.31% and 90.91%, respectively, with perfect precision (1.0000). This demonstrates that the synthetic oversampling and noise-filtering stages in the SMOTE-Tomek layer allow the downstream stacking ensemble to learn robust, generalizable patterns for rare attack signatures that standard models typically misclassify.

C. Post-Hoc Explainability Validation via SHAP

To address the "black-box" limitations of multi-layered ensemble stacking models, post-hoc explainability was integrated using SHAP (Shapley Additive exPlanations). This mechanism uses a cooperative game-theoretic approach to calculate the exact contribution score of each network traffic feature to individual prediction outcomes. Global feature importance analysis was conducted by averaging the absolute SHAP values across the entire validation dataset. The resulting top-ranked features are illustrated in Figure 5.9. The analysis confirms that packet payload attributes and temporal network flow patterns drive the framework's classification engine. The feature fwd_pkts_payload.min emerged as the most critical global descriptor (0.6383 impact magnitude), followed by network destination port identifiers (id.resp_p) and TCP window transmission rates (fwd_last_window_size). This ordering validates that HADES makes its anomaly predictions based on physically meaningful telemetry signatures rather than arbitrary statistical noise.

To evaluate the local feature dynamics responsible for triggers within specific attack classes, SHAP beeswarm tracking was executed. Figure 5.10 details the localized feature contribution distribution mapping for individual classifications. The horizontal alignment indicates whether a feature value increases (positive SHAP domain) or decreases (negative SHAP domain) the likelihood of anomaly classification, while the color scale represents feature magnitude. For example, low values of packet payload statistics combined with highly concentrated, short packet inter-arrival times (flow_iat.min) generate strong positive SHAP vectors. This aligns perfectly with the expected network profile of rapid, zero-payload automated flooding exploits, showing that the meta-learner matches real-world expert domain knowledge.

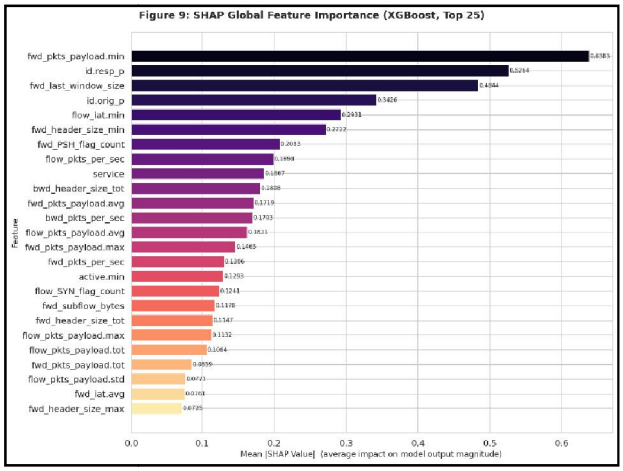

Figure 2: Global SHAP feature importance ranking for the proposed HADES framework

D. Security stress-testing via Adversarial Evasion (FGSM)

To evaluate the framework's resilience against targeted evasion maneuvers, proactive adversarial testing was conducted using the Fast Gradient Sign Method (FGSM). This attack introduces calculated adversarial perturbations (epsilon) to the validation feature values along the gradient of the loss function, maximizing classification errors while minimizing visible changes to the underlying data structure. The resulting performance degradation curves across increasing

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

perturbation strengths are displayed in Figure 5.8. Individual baseline models, such as standard Decision Trees and standalone LightGBM configurations, show rapid performance drops as perturbation levels increase, making them vulnerable to gradient evasion tactics.

In contrast, the proposed HADES framework maintains a highly resilient, stable horizontal performance line across scaling epsilon (epsilon) variations. By stacking diverse gradient boosting architectures and bagging mechanisms through an optimized Logistic Regression meta-classifier, HADES reduces sensitivity to feature variations that degrade single standalone models. This robust performance under adversarial pressure validates HADES as a secure and trustworthy intrusion detection framework suitable for deployment in high-risk IoT networking environments.

Table 4: Comparative analysis of machine and deep learning algorithms

Model	Accuracy (%)	Macro-F1 (%)	Weighted-F1 (%)	Kappa	Training Time (s)
Decision Tree	99.71	94.32	–	0.9927	1.0
Naïve Bayes	95.20	68.97	–	0.8800	0.3
KNN (k=5)	99.57	92.38	–	0.9891	0.1
SVM (RBF)	99.44	90.98	–	0.9860	43.8
Random Forest	99.81	95.06	99.80	0.9951	22.1
XGBoost	99.84	95.89	99.84	0.9960	94.8
LightGBM	90.59	31.65	88.91	0.7526	50.7
CatBoost	99.78	95.60	99.77	0.9944	476.2
HADES Framework	**99.79**	**97.70**	**99.79**	–	–

Table IV compares the performance of various machine learning and deep learning algorithms for cyberattack detection. XGBoost achieved the highest accuracy of 99.84%, followed closely by Random Forest (99.81%), HADES Framework (99.79%), and CatBoost (99.78%). The proposed HADES Framework obtained the highest Macro-F1 score (97.70%), indicating superior performance across all classes, including minority attack categories. In contrast, Naïve Bayes and LightGBM showed comparatively lower performance. Overall, the results demonstrate that the HADES Framework provides highly balanced and reliable classification performance while maintaining accuracy comparable to the best-performing models.

VI. CONCLUSION

This dissertation research introduced a robust, scalable, and intelligent cybersecurity architecture named HADES (Hierarchical Adversarial-Aware Detection via Ensemble Stacking), designed specifically to secure modern, resource-constrained Internet of Things (IoT) network environments. By shifting away from rigid, legacy signature-based intrusion filters that fail against zero-day exploits, the proposed system leverages data-driven machine learning pipelines to parse complex network telemetry flows in real time.

The structural novelty of HADES lies in its multi-layered, synergistic optimization pipeline. It successfully balances data and algorithmic dimensions by combining filter- and wrapper-based MI-RFE feature selection to drop redundant network traffic dimensions down to an optimal subset of 20 high-variance attributes. To resolve the systemic prediction bias associated with highly skewed IoT attack data, the framework implements the hybrid SMOTE-Tomek algorithm, which synthesizes minority threat logs while stripping away localized decision-boundary noise to optimize class separability. The final classification engine utilizes a hierarchical stacking ensemble where the individual prediction perspectives of four tree-based architectures (XGBoost, LightGBM, CatBoost, and Random Forest) are mathematically blended and optimized by a Logistic Regression meta-learner. Benchmark validation on the realistic RT-IoT2022 dataset demonstrated that HADES delivers exceptional performance, yielding an overall accuracy of 99.79\% and achieving the highest overall Macro-F1 score of 97.70\%. This proves its distinct edge in identifying rare, low-frequency security threats (such as targeted brute-forcing or specific scans) where standalone baseline models fail.

Furthermore, to ensure compliance with the operational transparency and structural trust requirements of live security operations centers, SHAP-based post-hoc explanations were integrated to provide full visibility into global feature

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 4, June 2026 Impact Factor: 8.2

weights and local prediction logic. Finally, proactive stress testing via the Fast Gradient Sign Method (FGSM) proved the superior structural resilience of the HADES ensemble framework, showing that it consistently maintains stable detection performance under active adversarial evasion maneuvers that cause standalone classifiers to degrade drastically. While the HADES framework has demonstrated high-precision threat classification, model transparency, and adversarial defense capabilities, several opportunities exist for further empirical enhancement and real-world deployment.

REFERENCES

[1]. M. Zahid and T. S. Bharati, Enhancing Cybersecurity in IoT Systems: A Hybrid Deep Learning Approach for Real-Time Attack Detection, vol. 5, no. 1, Springer International Publishing, 2025. doi: 10.1007/s43926-025-00156-y.

[2]. Malik, K. Arshid, N. Noonari, and R. Munir, "Artificial Intelligence-Driven Cybersecurity Framework Using Machine Learning for Advanced Threat Detection and Prevention," Sch. J. Eng. Technol., vol. 13, no. 6, pp. 401–423, 2025. doi: 10.36347/sjet.2025.v13i06.005.

[3]. S. Deepa Rajan and A. Manikandan, "Navigating Cybersecurity: A Comprehensive Analysis of Machine Learning in Cyber Attack Detection," J. Theor. Appl. Inf. Technol., vol. 102, no. 21, pp. 7658–7669, 2024.

[4]. D. Prabakar, M. Sundarrajan, R. Manikandan, N. Z. Jhanjhi, M. Masud, and A. Alqhatani, "Energy Analysis-Based Cyber Attack Detection by IoT with Artificial Intelligence in a Sustainable Smart City," Sustainability, vol. 15, no. 7, pp. 1–14, 2023. doi: 10.3390/su15076031.

[5]. A Salam, F. Ullah, F. Amin, and A. Mohammad, "Deep Learning Techniques for Web-Based Attack Detection in Industry 5.0," 2023, pp. 1–18.

[6]. Y.-W. Chen, J.-P. Sheu, Y.-C. Kuo, and N. Van Cuong, "Design and Implementation of IoT DDoS Attacks Detection System Based on Machine Learning," in Proc. European Conference on Networks and Communications (EuCNC), Dubrovnik, Croatia, 2020, pp. 122–127. doi: 10.1109/Eu-CNC48522.2020.9200909.

[7]. Y. Jia, F. Zhong, A. Alrawais, B. Gong, and X. Cheng, "FlowGuard: An Intelligent Edge Defense Mechanism Against IoT DDoS Attacks," IEEE Internet Things J., vol. 7, no. 10, pp. 9552–9562, Oct. 2020. doi: 10.1109/JIOT.2020.2993782.

[8]. M. S. Elsayed, N.-A. Le-Khac, S. Dev, and A. D. Jurcut, "DDoSNet: A Deep-Learning Model for Detecting Network Attacks," in Proc. IEEE 21st International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), Cork, Ireland, 2020, pp. 391–396. doi: 10.1109/WoWMoM49955.2020.00072.

[9]. G. Bendiab, S. Shiaeles, A. Alruban, and N. Kolokotronis, "IoT Malware Network Traffic Classification Using Visual Representation and Deep Learning," in Proc. 6th IEEE Conference on Network Softwarization (NetSoft), Ghent, Belgium, 2020, pp. 444–449. doi: 10.1109/NetSoft48620.2020.9165381.

[10]. J. Bhayo, S. Hameed, and S. A. Shah, "An Efficient Counter-Based DDoS Attack Detection Framework Leveraging Software Defined IoT (SD-IoT)," IEEE Access, vol. 8, pp. 221612–221631, 2020. doi: 10.1109/ACCESS.2020.3043082