An Intelligent and Explainable Intrusion Detection Framework for Internet of Sensor Things using Generalizable Optimized Active Machine Learning

Muhammad Hasnain[a], Nadeem Javaid[a,*,1], Abdul Khader Jilani Saudagar[b,*] and Neeraj Kumar[c,d,e]

[a]ComSens Lab, International Graduate School of Artificial Intelligence, National Yunlin University of Science and Technology, Douliu, Yunlin 64002, Taiwan

[b]Information Systems Department College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

[c]Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala 147001, India

[d]School of Computer Science, University of Petroleum and Energy Studies, Dehradun 248001, India

[e]Department of Computer Science and Information Engineering, Asia University, Taichung 41354, Taiwan

ARTICLE INFO

Keywords:
Intrusion Detection
Internet of Sensor Things
Logistic Regression
Sandpiper Optimization
Entropy-based
Uncertainty-based Active Learning
10-Fold Cross Validation
Paired *t*-test
SHapley Additive exPlanations
Local Interpretable Model-agnostic Explanations

ABSTRACT

Intrusion Detection (ID) in the Internet of Secure Things (IoST) has become increasingly critical due to the rising frequency and sophistication of cyber-attacks, which can lead to severe consequences such as data breaches, financial losses, and service disruptions. These risks are further intensified in computationally limited environments, where limited computational capacity and rapidly evolving threats make accurate and efficient detection challenging. In this study, a data-efficient ID framework tailored for resource-constrained environments is proposed by leveraging active learning and meta-heuristic optimization techniques. The proposed framework systematically addresses three critical limitations commonly observed in traditional models: data imbalance, inefficient hyperparameter tuning, and dependency on large labeled datasets. Initially, to mitigate mitigate class imbalance, adaptive synthetic sampling generates synthetic instances for minority classes, thereby enhancing learning in complex regions of the feature space. Next, for hyperparameter optimization, the Sandpiper Optimization (SO) algorithm fine-tunes the regularization parameter of Logistic Regression (LR), yielding significant improvements in model generalization. Finally, the challenge of limited labeled data is addressed through two active learning strategies: Active Learning Uncertainty-based (ALU) and Active Learning Entropy-based (ALE). These strategies selectively query the most informative samples from the unlabeled pool, ensuring maximum learning with minimal annotation effort. The performance of the proposed models is evaluated on two benchmark datasets: the wireless sensor networks and network intrusion detection datasets. Simulation results demonstrate that proposed models outperform base model LR. LRALE achieves improvements of 10.48% and 3.16% in accuracy, 19.48% and 3.16% in recall, and 7.23% and 1.04% in F1-score on WSN-DS and CIC-IDS-DS datasets, respectively. LRALU shows improvements of 18.18% and 2.11% in accuracy, 18.18% and 2.11% in recall, and 14.63% and 2.08% in Receiver Operating Characteristic – Area Under the Curve (ROC-AUC). Similarly, LRSO achieves improvements of 9.09% and 2.11% in accuracy, 9.09% and 1.05% in recall, and 9.76% and 3.12% in ROC-AUC on WSN-DS and CIC-IDS-DS datasets, respectively. To ensure model generalization and stability across different data partitions, a rigorous 10-fold cross-validation is conducted. Model interpretability is further enhanced using eXplainable artificial intelligence techniques, including Local interpretable model-agnostic explanations and Shapley additive explanations, to elucidate feature contributions and improve transparency. Additionally, statistical significance testing through paired *t*-tests confirms the robustness and reliability of the proposed models. Overall, this framework introduces a comprehensive, annotation-efficient, and transparent ID solution that significantly advances the domain, making it well-suited for practical deployment in IoSTs environments.

1. Introduction

Internet of Sensors Things (IoSTs) are now essential to many vital applications, such as disaster management, smart healthcare, industrial automation, battlefield surveillance, and environmental monitoring [1]. These networks are made up of many sensor nodes that wirelessly gather and send data to a base station or central sink [2]. Due to their distributed architecture and deployment in often hostile or unattended environments, IoSTs are exposed to numerous security threats, making security a primary concern in their design and operation . Among the various security mechanisms, Intrusion Detection Systems (IDS) are pivotal in identifying and mitigating malicious behavior within the network [3]. The IoSTs refers to a global infrastructure of linked devices, each with a unique identifier, operating through standardized communication protocols to enable seamless interaction between people and objects anywhere, anytime, using any available network or service [4].
Reliability and trustworthiness of sensor data play a critical role in ensuring the effectiveness of IoST IDS. In resource-constrained IoST environments, faulty, degraded, or compromised sensors can produce inaccurate or misleading

*Corresponding Author: javaidn@yuntech.edu.tw, aksaudagar@imamu.edu.sa
ORCID(s):
[1] https://orcid.org/0000-0003-3777-8249

data, which often lead to incorrect classification decisions and, consequently, serious operational failures. Therefore, the robustness of the proposed framework is inherently dependent on the quality of the sensor data it receives. As highlighted in [5], incorporating mechanisms for sensor reliability assessment, fault detection, and data validation can significantly enhance trust in the sensing infrastructure. Traditional security measures such as encryption and authentication are often insufficient, especially against internal attacks where a node within the network becomes compromised. IDS provides an additional security layer by constantly monitoring node activity, detecting anomalies, and addressing potential attacks in real time. Designing an effective IDS is tough owing to the inherent energy, memory, and processor limitations of IoSTs nodes [6]. The dynamic characteristics of IoSTs [7], stemming from node failures, activity fluctuations, or energy depletion, significantly hamper Intrusion Detection (ID). This study is an extension of [8]. Attacks in IoSTs often manifest in several forms, such as packet dropping, data transformation, identity impersonation, or Denial-of-Service (DoS), each targeting specific network layers and components. An effective IDS must be adaptive, lightweight, scalable, and adept at handling diverse and evolving attack strategies while minimizing false positives and maintaining network sustainability. To address these challenges, recent research has explored intelligent ID mechanisms powered by Machine Learning (ML) [9] and Deep Learning (DL) techniques. These data-driven approaches enable the IDS to learn from historical and real-time data, detect abnormal patterns, and improve its performance over time without explicit reprogramming. By capturing hidden relationships in network behavior and identifying subtle deviations that might go unnoticed by rule-based systems, ML and DL offer a promising direction for robust and adaptive ID. However, their integration into IoSTs must be carefully designed to ensure computational efficiency and suitability for resource-constrained environments.

Network ID has been extensively explored to safeguard IoSTs infrastructures against evolving threats. However, traditional systems often face significant challenges such as class imbalance, inefficient hyperparameter tuning, and high computational demands. To overcome these issues, recent research focuses on leveraging data balancing technique like ADAptive SYNthetic Sampling (ADASYN), integrating Active Learning (AL) strategies based on uncertainty and entropy, and employing nature-inspired optimization methods like Sandpiper Optimization (SO). These approaches aim to enhance detection performance, improve operational efficiency, and reduce annotation efforts, thereby offering more robust, scalable, and annotation-efficient solutions for real-world IoSTs deployments. Table 1 provides a comprehensive list of all abbreviations and mathematical symbols used throughout this study, along with their corresponding descriptions.

Table 1
List of Abbreviations and Symbols

Abbreviation/ Symbol	Description
10-FCV	10-Fold Cross Validation
$[lb, ub]$	Lower and upper bounds of search space
ADASYN	Adaptive Synthetic Sampling
AL	Active Learning
ALE	Active Learning Entropy-based
ALU	Active Learning Uncertainty-based
α	Random number $\sim U(-1, 1)$
C	Regularization hyperparameter of Logistic Regression
DL	Deep Learning
DoS	Denial-of-Service
f_i	Fitness value (F1-score) of sandpiper i
IDS	Intrusion Detection System
IoST	Internet of Sensor Things
K	Total number of classes
LR	Logistic Regression
LRALE	Logistic Regression with Active Learning Entropy-based
LRALU	Logistic Regression with Active Learning Uncertainty-based
LRSO	Logistic Regression with Sandpiper Optimization
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
$\mathbf{x}$	Input feature vector
$\mathbf{w}_k$	Weight vector of class k
$P(y = k\|\mathbf{x})$	Conditional probability of class k
ROC-AUC	Receiver Operating Characteristic - Area Under Curve
r_1, r_2	Random numbers $\sim U(0, 1)$
S_i	Position of sandpiper i in search space
S_{best}	Best solution found (optimal C value)
SHAP	SHapley Additive exPlanations
SO	Sandpiper Optimization
t-test	Paired t-test (statistical significance testing)
T	Maximum number of iterations (in SO algorithm)
v_i	Velocity of sandpiper i
X_{train}	Training feature dataset
X_{test}	Testing feature dataset
XAI	Explainable Artificial Intelligence
y_{train}	Training label dataset
$\hat{y}_{\text{test}}$	Predicted class labels on test data
$\mathbb{R}^d$	d-dimensional real space

2. Related Work

Several studies have proposed solutions to address the inherent challenges in Network IDS (NIDS). Study [10] focused on mitigating issues related to data imbalance and computational inefficiency. The authors employ ADASYN to enhance minority class representation and adopt Light Gradient Boosting Machines (LightGBM) to accelerate

model training and inference without compromising accuracy. Evaluations on benchmark datasets, including NSL-KDD, UNSW-NB15, and CIC-IDS-DS, demonstrate that the integration of ADASYN and LightGBM consistently improves detection performance and reduces computational overhead, outperforming conventional deep learning-based detection methods. However, a key limitation of this approach lies in its reliance on shallow learning models, which often struggle to capture highly complex attack patterns compared to deeper architectures, potentially limiting scalability to evolving threat landscapes.

Authors in [11] proposed an ID framework called Tabular Auxiliary Classifier Generative Adversarial Networks (TACGAN) to tackle the challenges posed by imbalanced ID datasets. The framework integrates feature extraction, K-Nearest Neighbor (KNN)-based undersampling, and a novel data generation approach using an enhanced generative adversarial network. Specifically, the TACGAN model extends the traditional Auxiliary Classifier GAN (ACGAN) by incorporating information loss constraints into the generator, producing more realistic minority class samples. KNN-based undersampling is applied to normal data, while noise filtering enhances the quality of generated samples, effectively balancing the dataset and improving detection performance across multiple benchmarks. Nonetheless, a notable limitation of TACGAN-based IDS is the potential for mode collapse and overfitting within the GAN component, which can compromise the diversity and representativeness of synthetic samples in highly heterogeneous network traffic environments.

In [12], the authors propose a load balancing framework for Software Defined Networks (SDNs) to efficiently distribute vehicular sensor loads using network and application-level information. The model enhances resource utilization, reduces latency, and optimizes system performance by incorporating a convergence-based mechanism for dynamic traffic management. To improve ID, an entropy-based AL strategy is integrated, selectively augmenting training instances by leveraging uncertainty measures. This approach refines decision boundaries and enhances classification performance in both sparse and dense data environments. However, a limitation of this framework is its dependency on accurate and timely vehicular sensor data, which often not always be feasible in dynamic or large-scale IoSTs environments, potentially degrading detection reliability under real-time constraints.

Authors in [13] address the challenge of effectively prioritizing high-importance alerts generated by NIDS, which often overwhelm analysts with numerous low-relevance notifications. The study highlights the underexplored potential of AL to enhance NIDS alert classification by incorporating expert feedback during the model lifecycle. The proposed AL-based alert classification framework selectively identifies critical alerts, improving decision-making efficiency and minimizing analyst workload in large-scale alert environments. Despite its strengths, the primary limitation of this approach lies in the requirement for continuous human-in-the-loop interaction, which can introduce subjectivity, increase operational costs, and reduce scalability when deployed in resource-constrained settings.

Study [14] introduces a novel Sandpiper Optimizer with Hybrid DL-based (SPOHDL) ID framework designed for blockchain-assisted Internet of Things (IoT) environments. The model integrates the SO for feature selection, a hybrid Convolutional Neural Network-Stacked Autoencoder (CNN-SAE) model for ID, and the Beetle Search Optimization Algorithm (BSOA) for fine-tuning hyperparameters. SO is employed to select relevant features from complex IoT network traffic data, addressing challenges posed by high-dimensional and dynamic datasets. Experimental results indicate superior accuracy and detection performance on benchmark datasets. However, a key limitation of the SPO-HDL framework is its computational complexity due to the integration of multiple optimization and DL components, potentially hindering real-time applicability and scalability in resource-constrained IoT infrastructures.

Lastly, for ID and classification, the authors in [15] proposed a Gravitational Search Algorithm-based Feature Selection with Optimal Quantum Neural Network (GSAFS-OQNN) model. The approach utilizes Z-score normalization for preprocessing, applies GSAFS to identify optimal features, and employs a Quantum Neural Network (QNN) for detection. Additionally, the SO is used to fine-tune the QNN parameters, improving model accuracy. Experimental analysis on benchmark IDS datasets demonstrates the superior performance of GSAFS-OQNN compared to recent methods. Despite its promising results, the limitation of this method stems from the inherent complexity and interpretability challenges associated with quantum-inspired models, which can impede practical deployment and explainability in real-world network security applications. A concise comparative analysis of the reviewed studies, their limitations, and the ways in which the proposed models address these gaps is presented in Table 2.

2.1. Knowledge Gaps

The major research gaps identified in the existing studies are mentioned below.

- Existing studies on ID often fail to address the class imbalance problem effectively, which leads to biased predictions toward majority classes and degrades the detection of minority attack categories.
- Many approaches lack efficient mechanisms for selecting the most informative samples, resulting in higher labeling costs and reduced learning efficiency.
- Current models often rely on fixed or manually chosen hyperparameters, limiting their adaptability and generalization capability across diverse network environments.
- There is insufficient consideration of rigorous validation strategies, which raises concerns about the robustness and reproducibility of the reported results.

Table 2
Summary of Related Work

Method	Limitations	Solutions Provided by Proposed Framework (LRALE, LRALU, & LRSO)
ADASYN + LightGBM [10]	Relies on shallow learning (LightGBM) which struggles to capture complex nonlinear attack patterns; limited scalability to evolving threats.	Proposed LRALE/LRALU integrate LR with AL to iteratively select highly-informative samples (entropy/uncertainty-based), enabling improved boundary refinement and adaptability to evolving intrusion patterns while retaining computational efficiency.
TACGAN-IDS [11]	GAN-based generation prone to mode collapse and overfitting, reducing diversity of synthetic samples in heterogeneous traffic.	LRALE/LRALU avoid generative instability by leveraging ADASYN for controlled minority oversampling, combined with targeted AL queries to select hard-to-classify real samples, improving diversity without adversarial collapse.
Entropy-based AL in SDN load balancing [12]	Performance depends heavily on accurate/timely vehicular sensor data; less reliable in large-scale or dynamic IoST scenarios.	Our AL strategies (LRALE/LRALU) operate on static IoSTs-derived features with uncertainty-based selection, minimizing dependence on rapidly-changing sensor contexts, improving robustness in dynamic environments.
AL-based NIDS alert prioritization [13]	Requires continuous human-in-the-loop annotation, increasing subjectivity, operational cost, and reducing scalability.	LRALE/LRALU reduce annotation requirements by selecting only the most uncertain/entropy-maximized samples per iteration, thus minimizing human labeling while retaining high information gain.
SPOHDL (SO + CNN-SAE + BSOA) [14]	High computational complexity from deep architectures and multiple optimizers, limiting real-time deployment in resource-constrained IoT.	LRALE/LRALU maintain linear model complexity (LR core) and lightweight AL loops, ensuring low-latency training and inference suitable for constrained environments while still improving detection performance.
GSAFS-OQNN [15]	Quantum-inspired models have interpretability issues and high complexity, hindering practical deployment.	Proposed models remain interpretable via LR, with feature contribution explainability (SHAP & LIME) integrated, while AL-driven sampling ensures efficient use of labeled data without model complexity overhead.

- Statistical significance of improvements is frequently overlooked, leaving the reliability of performance gains unverified.
- Interpretability of ID models remains limited, restricting their adoption in practical network security applications where transparency is critical.
- Few studies evaluate the generalizability of their models across multiple datasets, which raises questions about the robustness and applicability of the proposed solutions in real-world scenarios.

2.2. Contributions

This study makes pivotal technical contributions through the development of the Logistic Regression (LR) with Active Learning Uncertainty-based (LRALU), LR with AL Entropy-based (LRALE), and LR with Sandpiper Optimization (LRSO) models for enhanced ID in IoSTs.

- First, to mitigate the class imbalance issue, ADASYN is employed to generate synthetic samples for minority classes. LR is then used as the base classifier for ID, ensuring a balanced and effective learning process.
- Second, two new proposed models, namely LRALU and LRALE, integrate AL strategies based on uncertainty and entropy, respectively. These models selectively query the most informative unlabeled instances, thereby improving detection accuracy while reducing annotation effort.
- Third, the proposed LRSO model incorporates the SO algorithm for hyperparameter tuning of LR. This enhances the generalization ability of the model and significantly improves classification performance in ID tasks.
- Forth, we perform a rigorous 10-Fold Cross-Validation (10-FCV) to ensure robust generalization across diverse data partitions.
- Fifth, paired t-tests are conducted to statistically validate the significance of improvements, confirming the superiority of the proposed models.
- Sixth, to enhance model interpretability and explanability, Local Interpretable Model-agnostic Explanations (LIME) and explainability for SHapley Additive exPlanations (SHAP) are utilized for transparent feature contribution analysis.

Table 3
Methodologies used in this Study for Intrusion Detection in Internet of Sensor Things

Method	Description	Application
Label Encoding	Converts categorical labels into numerical values without implying any ordinal relationship.	Transforms non-numeric attack labels into numeric format for model compatibility.
Min-Max Scaling	Normalizes features to a fixed range, typically [0,1].	Ensures feature uniformity and accelerates model convergence.
ADASYN	An adaptive oversampling technique that generates synthetic samples for minority classes according to their learning difficulty.	Addresses class imbalance and enhances learning in sparse or complex regions.
LR	A simple yet effective linear classification model using the sigmoid function.	Serves as the base classifier for ID in multiclass IoSTs settings.
LRALU	Selectively queries samples where the model is most uncertain.	Reduces annotation cost by focusing on the most ambiguous unlabeled instances.
LRALE	Uses entropy to quantify uncertainty and selects samples with highest entropy.	Improves boundary refinement and model generalization with fewer labels.
LRSO	Integrates a nature-inspired meta-heuristic to fine-tune model hyperparameters.	Optimizes LR regularization parameter to boost detection performance and reduce overfitting.
Statistical *t*-Test	Hypothesis testing method to compare means between paired models.	Validates the statistical significance of observed performance gains.
10-FCV	Repeatedly splits data into 10 subsets for training and testing.	Ensures robustness and consistency of the model across diverse partitions.
LIME	Model-agnostic interpretability method that explains local predictions.	Enhances transparency by identifying influential features in individual predictions.
SHAP	Uses game-theoretic approach to assign importance values to each feature.	Provides global and local feature contribution insights for models explainability.

- Seventh, to verify the generalizability and robustness of the proposed models, they are evaluated on different datasets, including the Wireless Sensor Network (WSNs) dataset and the Network Intrusion dataset (CIC-IDS-DS).

3. Proposed Data-Driven Framework for Intrusion Detection in Internet of Sensor Things

This study introduces a comprehensive data-driven framework for IoSTs ID that integrates several advanced techniques. First, the dataset is preprocessed using label encoding and min-max normalization to ensure feature uniformity and model stability. Second, class imbalance is addressed through ADASYN, which generates synthetic minority class samples to improve classifier performance. Third, uncertainty-based (ALU) and entropy-based (ALE) AL strategies are applied to minimize annotation cost by selecting the most informative instances. Finally, SO is used for hyperparameter tuning of the LR classifier to maximize detection accuracy and efficiency. The complete framework is shown in Figure 1. The methodologies used in this study described in Table 3.

3.1. Dataset description

This study utilizes two publicly available benchmark datasets to assess the effectiveness of ML algorithms in identifying and categorizing DoS attacks. The first is the Wireless Sensor Networks (WSNs) dataset [16], acquired within an IoSTs architecture, and the second is the Canadian Institute for Cybersecurity–IDS Dataset (CIC-IDS-DS) [17]. WSN dataset, designed for cybersecurity experimentation in IoSTs, facilitates the deployment of advanced detection algorithms. Integrating ML-based decision-making at the sensor node level enables the system to efficiently identify aberrant patterns indicating hostile activity, thereby enhancing real-time threat response. The dataset comprises 374,661 instances with 19 intended features that encapsulate the behavioral and energy characteristics of IoSTs nodes [18]. It encompasses a multi-class target, Attack type, which distinguishes between normal activities and other forms of the intrusion. The dataset provides different Features, such as *id*, *Time*, *is_CH*, *who_CH*, *Dist_To_CH*, *ADV_CH*, *ADV_S*, *ADV_R*, *JOIN_S*, *JOIN_R*, *SCH_S*, *SCH_R*, *Rank*, *DATA_R*, *Data_Sent_To_BS*, *dist_CH_To_BS*, *send_code*, *Expanded Energy*, which are suitable for training and assessing sophisticated ML and DL ID models, as shown in Table 4.

Similarly CIC-IDS-DS dataset employed in this study is publicly available on Kaggle. It contains five different csv files. From this dataset, we utilized the *Thursday-WorkingHours-Morning-WebAttacks.pcap_ISCX* file, which consists of 170,366 instances and 79 features. For computational efficiency and to ensure manageable processing within the proposed framework, a subset of 50,000 instances is utilized. The target column *label* contains two classes: *BENIGN* (48,913 instances) and *Web Attack – Brute Force* (1,082 instances). The dataset exhibits a highly imbalanced class distribution, which is addressed using the ADASYN

Table 4
IoSTs Dataset Features Description

Feature	Details	Description
id	Integer	Node unique identifier
Time	Integer	Timestamp of the observation
Is_CH	Binary (0/1)	Whether the node is a Cluster Head (1) or not (0)
who CH	Integer	ID of the associated Cluster Head
Dist_To_CH	Float	Distance of node to its Cluster Head
ADV_S	Integer	Number of advertisement messages sent
ADV_R	Integer	Number of advertisement messages received
JOIN_S	Integer	Number of join requests sent
JOIN_R	Integer	Number of join requests received
SCH_S	Integer	Number of schedule messages sent
SCH_R	Integer	Number of schedule messages received
Rank	Integer	Node rank in the cluster
DATA_S	Integer	Number of data messages sent
DATA_R	Integer	Number of data messages received
Data_Sent_To_BS	Integer	Number of data packets sent directly to Base Station
dist_CH_To_BS	Float	Distance from Cluster Head to Base Station
send_code	Integer	Code representing the sending action
Expaned Energy	Float	Expanded energy used by the node
Attack type	Categorical	Type of attack (Normal, Attack, etc.)

oversampling technique to generate additional minority class samples, thereby balancing the dataset and improving model training stability. The important features for CIC-IDS-DS features are shown in Table 5.

Table 5
CIC-IDS-DS Dataset Features Description

Feature	Details	Description
Flow Duration	Float	Total duration of a network flow in microseconds
Fwd IAT Mean	Float	Average inter-arrival time between two packets in the forward direction
Idle Mean	Float	Average idle time duration between flows
Flow Bytes/s	Float	Number of bytes transmitted per second in the flow
Fwd IAT Max	Float	Maximum inter-arrival time between forward packets
Active Mean	Float	Average active time of a flow before becoming idle
Idle Max	Float	Maximum idle time observed in a flow
Flow IAT Mean	Float	Average inter-arrival time between all packets in a flow
Bwd IAT Total	Float	Total inter-arrival time of all backward packets
Fwd URG Flags	Integer	Number of packets with URG flag set in forward direction
Flow IAT Min	Float	Minimum inter-arrival time between packets in a flow
URG Flag Count	Integer	Total count of URG flags in a flow
Idle Std	Float	Standard deviation of idle time durations
Bwd IAT Max	Float	Maximum inter-arrival time between backward packets
Packet Length Mean	Float	Average length of packets in the flow
Down/Up Ratio	Float	Ratio of download to upload packets
Fwd Avg Bulk Rate	Float	Average bulk transfer rate in forward direction
Fwd Packets/s	Float	Number of packets transmitted per second in forward direction
Subflow Bwd Bytes	Float	Number of bytes in backward subflow

3.2. Data Preprocessing

Efficient data preparation is crucial for effective ID using ML and DL models. To this end, we employ two fundamental preprocessing techniques: label encoding and min-max scaling as shown in Figure 1 [19] [20]. Initially, categorical attack types are converted into numerical form using *label encoding*, wherein each unique category is assigned a distinct non-negative integer [21] [22]. This transformation allows categorical features to be utilized in algorithms that require numerical inputs, although it does not encode any ordinal relationships.

$$f : C \to \mathbb{N}, \quad f(c_i) = n_i \quad \forall\, c_i \in C \tag{1}$$

where $C = \{c_1, c_2, \ldots, c_k\}$ represents the set of unique categorical values, and $n_i \in \{0, 1, \ldots, k-1\}$ is the integer assigned to category c_i.

Subsequently, min-max scaling is applied to normalize the input features into a standard range, typically $[0, 1]$, as

defined by the following transformation as:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2}$$

where, X is the original feature value, while $X_{\min}$ and $X_{\max}$ represent the minimum and maximum values of the feature, respectively. X_{scaled} denotes the normalized value of X, linearly rescaled to the [0, 1] range. This normalization maintains the relative distribution of the original data while ensuring uniform feature contribution during model training [23]. By scaling features to a common range, min-max scaling improves convergence behavior and enhances the performance of gradient-based optimization algorithms.

3.3. Data Balancing with ADAptive SYNthetic Sampling

ADASYN is an advanced oversampling technique designed to handle class imbalance by adaptively generating synthetic samples for the minority class, focusing on difficult-to-learn regions near decision boundaries. For binary classification on CIS-IDS-DS, the number of synthetic samples to generate for a minority instance x_i is determined as

$$G_i = \frac{r_i}{\sum_{j=1}^{n_{min}} r_j} \times G, \tag{3}$$

where r_i represents the ratio of majority samples in the k-nearest neighbors of x_i, n_{min} is the number of minority samples, and G is the total number of synthetic samples required.
For a multi-class on WSNs dataset:

$$D = \{(x_i, y_i)\}_{i=1}^{n}, \quad y_i \in \{1, 2, \ldots, C\}, \tag{4}$$

minority classes c_m are processed using a one-vs-all scheme. For each $x_i \in c_m$, the difficulty ratio is:

$$r_i = \frac{\Delta_i}{k}, \tag{5}$$

where Δ_i is the count of majority-class neighbors among the k-nearest neighbors. The total synthetic samples for c_m is G, with per-instance allocation:

$$G_i = \frac{r_i}{\sum r_j} \cdot G. \tag{6}$$

Synthetic samples are created as:

$$\hat{x} = x_i + \lambda \cdot (x_{zi} - x_i), \tag{7}$$

where x_{zi} is a random minority neighbor and $\lambda \sim U(0, 1)$. By focusing on difficult regions, ADASYN improves generalization and robustness in binary and multi-class imbalanced learning.

One key advantage of ADASYN in ID for IoSTs and CIC-IDS-DS is its strategic handling of class imbalance by concentrating synthetic sample generation in regions with higher learning difficulty. Unlike conventional oversampling, ADASYN adaptively targets minority instances near decision boundaries or in sparsely populated feature-space regions, where classification uncertainty is greatest. This targeted synthesis refines decision boundaries and enhances classifier discriminative power while mitigating overfitting to easily classifiable samples, as shown in Figure 1.
Empirical evidence demonstrates that models trained with ADASYN-balanced datasets consistently exhibit superior performance in the key of various evaluation metrics, ultimately producing more confident and reliable predictions. This improvement in predictive performance is especially critical in IoSTs environments, where accurately detecting sophisticated intrusion patterns in highly imbalanced data is paramount for safeguarding network integrity and ensuring operational resilience. Therefore, ADASYN not only ensures a more representative class distribution but also substantially enhances classifier robustness and generalization, making it an indispensable technique for ID within IoSTs frameworks.

3.4. Working Mechanism of Logistic Regression in Internet of Sensor Things-based Intrusion Detection

Logistic Regression (LR) is a widely used supervised learning method, primarily for binary classification, and extended to multi-class problems via the One-vs-Rest (OvR) approach [26, 27]. In OvR, K binary classifiers are trained, each distinguishing one class from all others, with the predicted class being the one with the highest probability.
For a binary dataset, Binary LR [28] models the probability of class 1 as:

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b), \tag{8}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the *sigmoid* activation, $\mathbf{w}$ is the weight vector, b the bias term, and $\mathbf{w}^\top \mathbf{x}$ the dot product with the input $\mathbf{x}$. The parameters $(\mathbf{w}, b)$ are learned by minimizing the binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left[y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right], \tag{9}$$

where $\hat{y}^{(i)} = \sigma(\mathbf{w}^\top \mathbf{x}^{(i)} + b)$. For a multi-class dataset, OvR trains K LR models:

$$P(y = k \mid \mathbf{x}) = \sigma(\mathbf{w}_k^\top \mathbf{x} + b_k), \tag{10}$$

assigning $\mathbf{x}$ to the class k with the highest $P(y = k \mid \mathbf{x})$. LR is extensively used in academic research and industry because of its simplicity, interpretability, and effectiveness in classification tasks. This is particularly advantageous when the correlation between input features and the target variable is approximately linear, and when model transparency is critical, such as in medical diagnosis, economics, or social sciences. LR provides probabilistic outputs that not only enable class prediction but also facilitate uncertainty

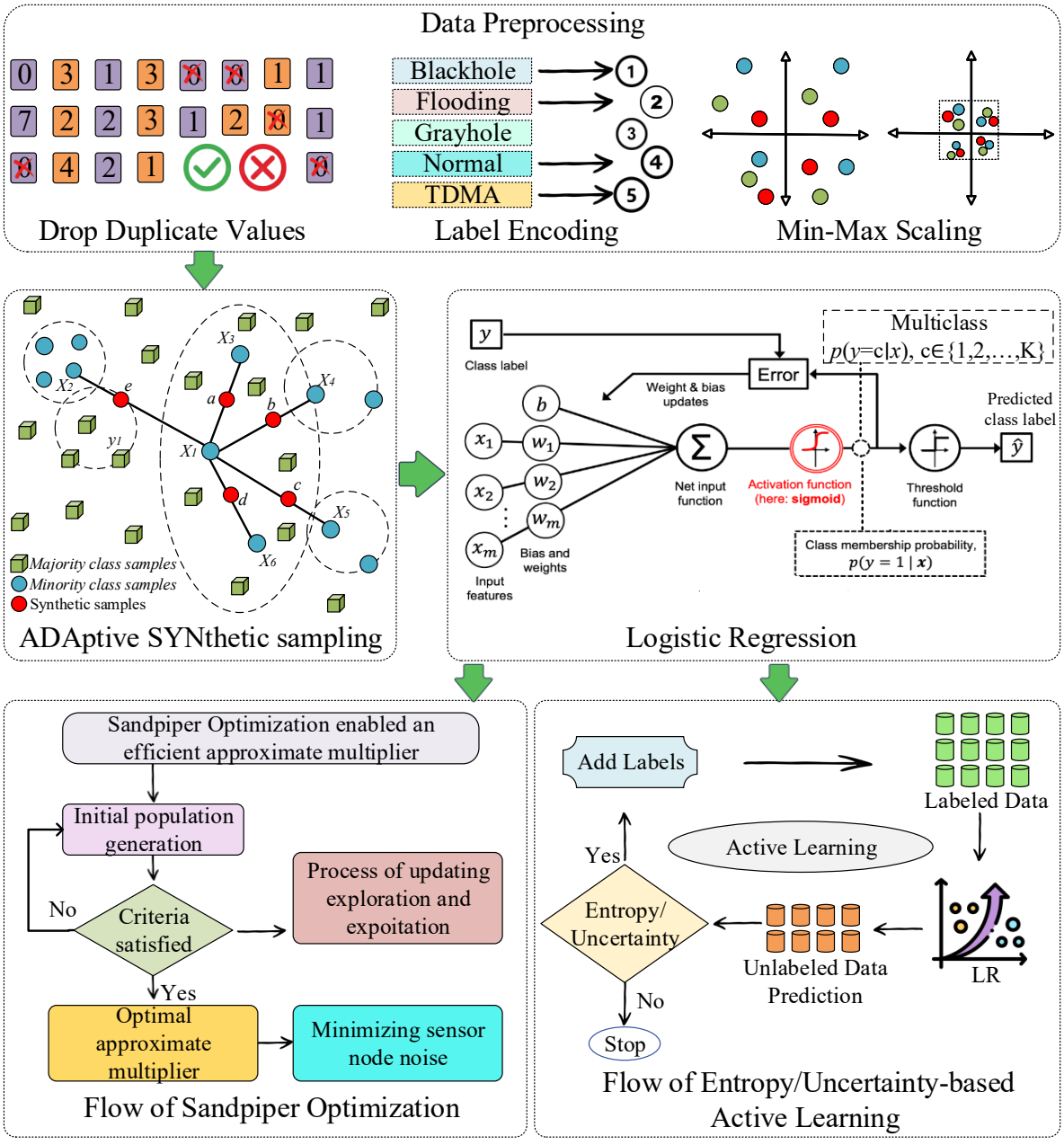


Figure 1: Proposed Data-Efficient Intrusion Detection Framework for Internet of Sensor Things

estimation, which is essential for decision-making in risk-sensitive contexts [29].

From a technical standpoint, LR is chosen as the base classifier because its convex optimization ensures stable convergence to a global minimum, making it highly compatible with AL strategies and metaheuristic optimization techniques. The model's linear structure allows feature coefficients to directly reflect the impact of each input variable, offering clear interpretability and enabling insights into feature importance, which is crucial in domains where understanding model reasoning is required. Furthermore, LR functions effectively on small datasets, can adapt to multi-class problems via OvR, and serves as a robust baseline for benchmarking more complex algorithms. A limitation of LR remains its assumption of linear decision boundaries, which often reduce performance on datasets with strong non-linear patterns.

3.5. Leveraging Active Learning for Intrusion Detection in Internet of Sensor Things

AL is an ML technique whereby the model selectively requests labels for the most informative samples to enhance performance with limited labeled data [30]. It is particularly beneficial in situations when labeling is costly or labor-intensive. *Uncertainty-based* sampling is a fundamental method in which the model emphasizes instances for which it has the least confidence in its predictions. An established

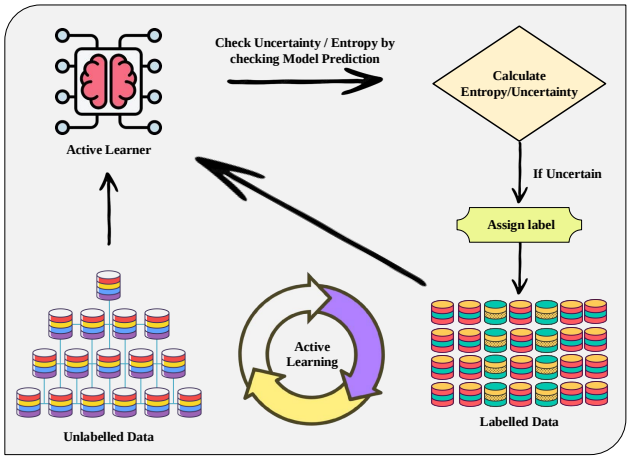


Figure 2: Working Flow of Active Learning Strategy

method in this context is *entropy-based* sampling, which measures prediction uncertainty via the entropy of the output probability distribution; higher entropy signifies more uncertainty. These strategies ensure the model concentrates on ambiguous samples, thereby enhancing learning efficiency and minimizing annotation effort [31]. The Figure 2 illustrates the working flow of the AL process.

3.5.1. Proposed LRALE: Logistic Regression with Entropy-based Active Learning

Entropy-based Active Learning (AL) is a sampling technique that selects unlabeled samples with the highest prediction entropy for labeling, with the objective of enhancing model performance while minimizing annotation cost [32]. The proposed approach leverages the LRALE, as detailed in Algorithm 1, to efficiently select informative samples based on entropy-driven uncertainty measures as shown in Figure 2. It utilizes entropy quantification to selectively annotate the most informative samples [33]. For a sample x_i with anticipated class probabilities derived from a trained LR model in a C-class classification task:

$$\mathbf{p}_i = [p_{i1}, p_{i2}, \ldots, p_{iC}], \quad \text{where} \quad \sum_{c=1}^{C} p_{ic} = 1, \quad (11)$$

where, $\mathbf{p}_i$ represents the predicted probability distribution over C possible classes for the i-th instance. Each element p_{ic} denotes the probability that the i-th instance belongs to class c, where $c \in \{1, 2, \ldots, C\}$. The condition $\sum_{c=1}^{C} p_{ic} = 1$ ensures that the probabilities across all classes sum to one, forming a valid probability distribution. Uncertainty is measured by the Shannon entropy. For the multi-class case:

$$H(x_i) = -\sum_{c=1}^{C} p_{ic} \log(p_{ic}), \quad (12)$$

while for the binary case ($C = 2$), this reduces to:

$$H(x_i) = -\Big[p_{i1} \log(p_{i1}) + (1 - p_{i1}) \log(1 - p_{i1})\Big]. \quad (13)$$

Algorithm 1 Proposed Logistic Regression with Active Learning Entropy-based (LRALE) for Intrusion Detection in Internet of Sensor Things

Require: Unlabeled dataset $\mathcal{D}_U$ with features {*id*, *Time*, *is_CH*, *who_CH*, *Dist_To_CH*, *ADV_CH*, *ADV_S*, *ADV_R*, *JOIN_S*, *JOIN_R*, *SCH_S*, *SCH_R*, *Rank*, *DATA_R*, *Data_Sent_To_BS*, *dist_CH_To_BS*, *send_code*, *Expanded Energy*}, initial labeled dataset $\mathcal{D}_L$, learning rate η, number of iterations T, number of queries n

Ensure: Trained Logistic Regression Model with Active Learning

1: Initialize Logistic Regression model $\mathcal{M}$
2: Train $\mathcal{M}$ on labeled data $\mathcal{D}_L$
3: **while** Stopping criterion is not met **do**
4: **for** $i = 1$ to T **do**
5: Train Logistic Regression model $\mathcal{M}$ on $\mathcal{D}_L$ using features such as *is_CH*, *Dist_To_CH*, *Rank*, *Data_Sent_To_BS*, *Expanded Energy*, etc.
6: Calculate class probabilities for each sample $\mathbf{x}_i \in \mathcal{D}_U$:

$$\mathbf{p}_i = [p_{i1}, p_{i2}, \ldots, p_{iC}], \quad \sum_{c=1}^{C} p_{ic} = 1$$

7: Compute entropy for each sample x_i:

$$H(x_i) = -\sum_{c=1}^{C} p_{ic} \log(p_{ic})$$

8: **end for**
9: Select top n high-entropy samples:

$$\mathcal{Q} = \underset{x_i \in \mathcal{D}_U}{\arg\text{ top-}n}\ H(x_i)$$

10: Query labels for samples in $\mathcal{Q}$
11: Update labeled dataset:

$$\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \mathcal{Q}$$

12: Remove selected samples from unlabeled pool:

$$\mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{Q}$$

13: Retrain model $\mathcal{M}$ on updated $\mathcal{D}_L$
14: **end while**
15: **return** Trained LRALE $\mathcal{M}$

Here, $H(x_i)$ denotes the predictive entropy of the model for the i-th instance, which quantifies the uncertainty associated with the predicted class probabilities. A higher value of $H(x_i)$ indicates greater uncertainty in the model's prediction for the given instance.

The entropy value reaches its peak when the model's predictions are uniformly distributed across classes, signifying maximum uncertainty [34]. This property makes entropy

a reliable criterion to identify ambiguous samples that can potentially enhance the model's decision boundary upon labeling. Entropy-based querying strategically prioritizes samples near class overlaps or regions with sparse training examples, thereby improving class separability and model robustness. Additionally, because entropy is agnostic to class distribution, it inherently addresses class imbalance by focusing on uncertain instances irrespective of class frequency.

In an AL loop, we first train an LR model on a preliminary labeled dataset $\mathcal{D}_L$. Subsequently, entropy scores are calculated for all samples in the unlabeled pool $\mathcal{D}_U$, and the top n cases exhibiting the highest uncertainty are chosen:

$$\mathcal{Q} = \underset{x_i \in \mathcal{D}_U}{\arg\,\text{top-}n}\ H(x_i), \tag{14}$$

where $\mathcal{Q}$ denotes the set of queried instances selected from the unlabeled pool $\mathcal{D}_U$. The queried samples $\mathcal{Q}$ are labeled and incorporated into the training set:

$$\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \mathcal{Q}, \quad \mathcal{D}_U \leftarrow \mathcal{D}_U \setminus \mathcal{Q}. \tag{15}$$

This iterative update mechanism ensures that the labeled dataset progressively expands with the most informative samples, while the unlabeled pool diminishes accordingly, thereby facilitating a more efficient and focused model retraining process in each AL cycle. Retraining the model on the updated labeled dataset enables it to refine decision boundaries in areas of highest uncertainty. This is particularly valuable in both binary and multi-class situations characterized by class overlap or imbalance.

By integrating entropy-based AL with LR, several substantial benefits are realized. Firstly, the selective annotation of high-entropy instances maximizes information gain per labeled sample, thereby significantly reducing the overall labeling burden. Secondly, by systematically resolving regions of high uncertainty, the model progressively improves its decision boundary precision and generalization performance, especially in challenging overlap scenarios. Moreover, this strategy mitigates the adverse effects of class imbalance by implicitly prioritizing informative minority-class samples. As a result, applying entropy-based AL in conjunction with LR not only enhances classification accuracy and robustness, but also leads to a more data-efficient learning process, achieving superior performance with fewer labeled samples and better resource utilization.

3.5.2. Proposed LRALU: Logistic Regression with Uncertainty-based Active learning

Uncertainty-based Active Learning (AL) is an iterative ML methodology that strategically identifies the most valuable unlabeled samples for labeling, based on the model's prediction uncertainty [35]. Rather than arbitrarily labeling data, it focuses on situations where the model exhibits the least confidence, with the objective of enhancing performance using fewer labeled samples, as illustrated in Figure 2. We integrate this approach with Logistic Regression (LR), a linear probabilistic classifier applicable to both binary and multi-class classification, as depicted in Figure 1. The AL process begins with a small subset of labeled data and incrementally requests additional samples from an unlabeled pool using uncertainty sampling, thereby improving the model in a data-efficient manner [36]. This combination exploits the statistical interpretability of LR with the targeted sampling strategy of AL, maximizing both predictive clarity and annotation efficiency. The proposed framework leverages LRALU, as detailed in Algorithm 2, to efficiently select informative samples based on uncertainty measures.

Algorithm 2 Proposed Logistic Regression with Active Learning Uncertainty-based (LRALU) for Intrusion Detection in Internet of Sensor Things

Require: Unlabeled dataset $\mathcal{D}_U$ with features {*id*, *Time*, *is_CH*, *who_CH*, *Dist_To_CH*, *ADV_CH*, *ADV_S*, *ADV_R*, *JOIN_S*, *JOIN_R*, *SCH_S*, *SCH_R*, *Rank*, *DATA_R*, *Data_Sent_To_BS*, *dist_CH_To_BS*, *send_code*, *Expanded Energy*}; initial labeled dataset $\mathcal{D}_L$, learning rate η, iterations T, queries n
Ensure: Trained Logistic Regression Model $\mathcal{M}$

```
Initialize Logistic Regression model M
Train M on D_L using selected features
while stopping criterion not met do
    for i = 1 to T do
        Train M on D_L
        Compute class probabilities P(y = c | x_i) for x_i ∈ D_U
        Compute uncertainty U(x_i) = 1 − max_c P(y = c | x_i)
    end for
    Select top n samples with highest U(x_i): Q
    Query oracle for labels of Q
    Update D_L ← D_L ∪ Q
    Update D_U ← D_U \ Q
end while
return Trained model M
```

In the mathematical setting, LR estimates the conditional class probability differently depending on the task. For binary classification, the probability of assigning a sample $\mathbf{x} \in \mathbb{R}^d$ to the positive class is given by:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w}^\top \mathbf{x} + b))} \tag{16}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector, $b \in \mathbb{R}$ is the bias, and $\mathbf{x}$ is the feature vector.

For multi-class classification, adopting the OvR or softmax formulation, the conditional probability of class $c \in \{1, 2, \ldots, C\}$ is:

$$P(y = c \mid \mathbf{x}) = \frac{e^{\mathbf{w}_c^\top \mathbf{x} + b_c}}{\sum_{j=1}^{C} e^{\mathbf{w}_j^\top \mathbf{x} + b_j}} \tag{17}$$

where $\mathbf{w}_c \in \mathbb{R}^d$ and $b_c \in \mathbb{R}$ denote the parameters for class c, and C is the total number of classes.

During each AL iteration, the model estimates class probabilities $P(y = c \mid \mathbf{x}_i)$ for every unlabeled sample $\mathbf{x}_i \in \mathcal{U}$, where $\mathcal{U}$ is the unlabeled pool. The uncertainty of each prediction is quantified using Shannon entropy:

$$H(\mathbf{x}_i) = -\sum_{c=1}^{C} P(y = c \mid \mathbf{x}_i) \log P(y = c \mid \mathbf{x}_i) \tag{18}$$

For the binary case ($C = 2$), the summation reduces to:

$$H(\mathbf{x}_i) = -\Big[P(y = 1 \mid \mathbf{x}_i) \log P(y = 1 \mid \mathbf{x}_i) + P(y = 0 \mid \mathbf{x}_i) \log P(y = 0 \mid \mathbf{x}_i) \Big] \tag{19}$$

thus providing a unified entropy-based measure of uncertainty across both binary and multi-class settings.
The uncertainty-based acquisition function is defined as:

$$\mathbf{x}^* = \arg\max_{\mathbf{x}_i \in \mathcal{U}} H(\mathbf{x}_i) \tag{20}$$

where $\mathbf{x}^*$ is the unlabeled sample with maximum entropy, selected for annotation at the current iteration. This sample is incorporated into the labeled set $\mathcal{L}$, and the model is retrained on the updated $\mathcal{L}$, thereby improving classification iteratively.
By integrating LR with uncertainty-based AL, the framework capitalizes on LR's calibrated probability estimates to quantify prediction uncertainty accurately, ensuring that only the most informative samples are annotated. This synergy reduces the annotation burden while systematically refining decision boundaries. Moreover, LR's convex optimization guarantees stable retraining at each iteration, ensuring robust convergence under evolving labeled datasets. Consequently, the combined LRALU approach enhances sample efficiency, model generalization, and decision boundary refinement across both binary and multi-class tasks, particularly in data-constrained environments.

3.6. Proposed LRSO: Logistic Regression with Sandpiper Optimization

The SO algorithm is a recent population-based meta-heuristic optimization method inspired by the intelligent foraging and adaptive survival strategies of sandpiper birds [37]. In nature, sandpipers exhibit an optimal balance between *exploration* (searching for new food sources) and *exploitation* (intensively feeding in high-reward areas), adapting their movements dynamically according to environmental cues. This biological behavior is mathematically modeled in SO to navigate a high-dimensional search space effectively, avoiding premature convergence while maintaining convergence speed. Compared to other meta-heuristic algorithms such as Reptile Optimization (RO) and Dipper Throated Optimization (DTO), SO incorporates stochastic perturbations and adaptive step-size control that enhance its ability to escape local minima and maintain search diversity, making it particularly suitable for complex, non-convex optimization landscapes such as hyperparameter tuning in

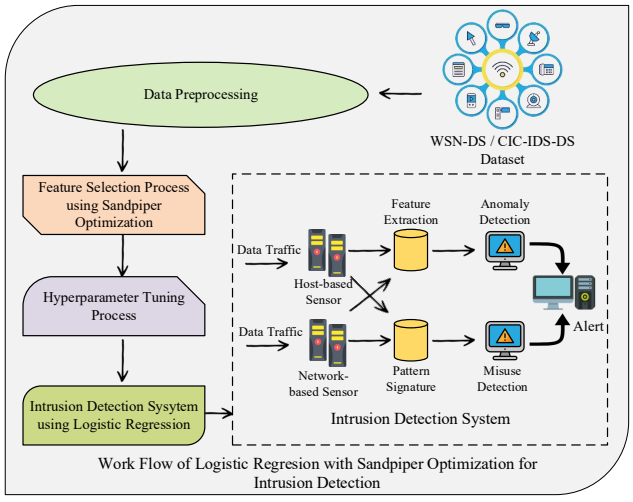


Figure 3: Work Flow of Logistic Regression with Sandpiper Optimization (LRSO) for Intrusion Detection in Internet of Sensor Things

ML models. Specifically, for LR hyperparameter tuning, SO's adaptive exploration–exploitation mechanism enables efficient searching of the regularization parameter space, balancing bias–variance trade-offs, and ensuring robust convergence without getting trapped in suboptimal regions, thereby improving classification accuracy and generalization.
The LRSO, outlined in Algorithm 3, optimizes the hyperparameters of LR using the SO technique. In SO, each sandpiper (agent) signifies a potential solution inside the search space, and its location is progressively refined using the global optimal solution [38]. The flow of LRSO is captured as shown in Figure 3. The updating velocity and location are provided as follows:

$$V_i^{(t+1)} = r_1 \cdot \left(X_{\text{best}}^{(t)} - X_i^{(t)} \right) + r_2 \cdot \left(X_{\text{best}}^{(t)} - X_i^{(t)} \right), \tag{21}$$

$$X_i^{(t+1)} = X_i^{(t)} + \alpha \cdot V_i^{(t+1)}, \tag{22}$$

where $V_i^{(t+1)}$ is the updated velocity of agent i at iteration $t+1$, r_1 and r_2 are uniformly distributed random scalars sampled from $\mathcal{U}(0,1)$ that introduce stochasticity for balancing exploration and exploitation, and $\alpha \in [-1, 1]$ represents a control factor that fine-tunes the step size towards equilibrium. The variable $X_i^{(t)}$ denotes the current position of the i^{th} agent in the solution space at iteration t, and $X_{\text{best}}^{(t)}$ signifies the global best solution found so far. To ensure the feasibility of solutions within the defined search space, boundary limitations are implemented:

$$X_i^{(t+1)} = \text{clip}(X_i^{(t+1)}, \text{lb}, \text{ub}) \tag{23}$$

where lb and ub represent the lower and upper bounds of the search space, respectively, and the clip$(\cdot)$ function ensures that $X_i^{(t+1)}$ remains within these boundaries. We employ the SO algorithm to optimize the regularization hyperparameter

C of LR for both binary and multi-class classification. In binary classification, the conditional probability of assigning a sample $\mathbf{x} \in \mathbb{R}^d$ to the positive class is computed using the *sigmoid* activation function:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}, \tag{24}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector and $\mathbf{x}$ is the input feature vector of dimensionality d. For multi-class classification, we adopt the OvR strategy, which decomposes a K-class problem into K binary subproblems. For each class k, a binary classifier distinguishes class k from all others. The conditional probability of assigning a sample $\mathbf{x}$ to class k is:

$$P(y = k \mid \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}_k^\top \mathbf{x})}, \quad \forall k \in \{1, 2, \ldots, K\}. \tag{25}$$

The optimized regularization parameter C governs the trade-off between maximizing the likelihood and penalizing model complexity in LR, directly influencing decision boundaries in both binary and multi-class tasks. By integrating LR with the SO technique, the framework intelligently tunes the regularization hyperparameter C to achieve optimal generalization performance. The stochastic yet convergent dynamics of SO enable an extensive exploration of the hyperparameter space, avoiding local minima that often hinder conventional grid or random search techniques. This adaptively controls model flexibility, minimizing both underfitting and overfitting across diverse data distributions. Moreover, SO's inherent balance of exploitation and exploration ensures that the selected C value maximizes classification accuracy while enhancing robustness in both binary and multi-class scenarios. Consequently, the optimized LR model exhibits sharper decision boundaries, improved class separability, and superior generalization on unseen data, all while maintaining computational efficiency due to SO's meta-heuristic-driven search. This fusion of LRSO thus provides a statistically rigorous yet computationally tractable framework for both binary and multi-class classification tasks constrained by hyperparameter sensitivity.

By systematically navigating the parameter space in Table 6, SO facilitates efficient convergence toward optimal solutions, improving both model generalization and classification robustness. This adaptive tuning mechanism enables LRSO to deliver highly calibrated probabilistic outputs and superior ID performance, particularly in complex and imbalanced IoSTs.

Algorithm 3 Proposed Logistic Regression with Sandpiper Optimization (LRSO) for Intrusion Detection in Internet of Sensor Things

Require: Training data $(X_{\text{train}}, y_{\text{train}})$ and Testing data $(X_{\text{test}}, y_{\text{test}})$ using features {*id*, *Time*, *is_CH*, *who_CH*, *Dist_To_CH*, *ADV_CH*, *ADV_S*, *ADV_R*, *JOIN_S*, *JOIN_R*, *SCH_S*, *SCH_R*, *Rank*, *DATA_R*, *Data_Sent_To_BS*, *dist_CH_To_BS*, *send_code*, *Expanded Energy*}, lower and upper bounds $[lb, ub]$, number of agents N, maximum iterations T

Ensure: Optimal hyperparameter C^*, Trained logistic regression model, Evaluation metrics

Initialize N sandpipers $S_i \in [lb, ub]$, $i = 1, 2, ..., N$
for each sandpiper S_i **do**
    Set logistic regression hyperparameter $C = S_i$
    Train logistic regression model on X_{train} using selected features: *is_CH*, *Dist_To_CH*, *Rank*, *Data_Sent_To_BS*, *Expanded Energy*
    Evaluate fitness $f_i \leftarrow$ F1-score on validation data
end for
$S_{\text{best}} \leftarrow \arg\max f_i$
$f_{\text{best}} \leftarrow \max f_i$
for $t = 1$ to T **do**
    for each sandpiper S_i **do**
        Generate $r_1, r_2 \sim U(0, 1)$ and $\alpha \sim U(-1, 1)$
        $v_i \leftarrow r_1 \cdot (S_{\text{best}} - S_i) + r_2 \cdot (S_{\text{best}} - S_i)$
        $S_i \leftarrow S_i + \alpha \cdot v_i$
        Clip S_i to bounds $[lb, ub]$
        Set $C = S_i$ and retrain logistic regression model using features *is_CH*, *Dist_To_CH*, *Rank*, *Data_Sent_To_BS*, *Expanded Energy*
        Evaluate new fitness $f_i \leftarrow$ F1-score on validation data
    end for
    if $\max f_i > f_{\text{best}}$ **then**
        $S_{\text{best}} \leftarrow \arg\max f_i$
        $f_{\text{best}} \leftarrow \max f_i$
    end if
end for
$C^* \leftarrow S_{\text{best}}$
Train final logistic regression model with $C = C^*$ using selected features: *is_CH*, *Dist_To_CH*, *Rank*, *Data_Sent_To_BS*, *Expanded Energy* on $(X_{\text{train}}, y_{\text{train}})$
Predict class labels $\hat{y}_{\text{test}}$ and predicted probabilities on X_{test} using same features
Compute performance metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC
Record execution time **return** C^*, trained LRSO, evaluation metrics

4. Experimental Results

To rigorously assess the effectiveness, resilience, and practicality of the proposed LRSO, LRALU, and LRALE models for ID in IoSTs and CIC-IDS-DS, a comprehensive series of experiments is conducted. This section presents an in-depth evaluation covering multiple aspects, including simulation outcomes, comparative analyses against baseline and state-of-the-art models, optimization strategy assessments, AL effectiveness, data balancing impacts, statistical significance tests, and generalization capabilities through

Table 6
Optimized Hyperparameters

Hyperparameter	Values Range	Optimal Value
C	{ 0.001, 0.01, 0.1, 1, 10, 100 }	0.1
solver	{ 'liblinear', 'saga', 'newton-cg', 'lbfgs' }	liblinear
multi_class	{ 'ovr', 'multinomial' }	ovr
max_iter	{ 10, 100, 1000 }	1000

Table 7
Performance Metrics and their Formulas

Measure	Formula
Accuracy	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\text{Precision} = \frac{TP}{TP+FP}$
Recall	$\text{Recall} = \frac{TP}{TP+FN}$
F1-Score	$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
ROC-AUC	$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$
Cohen's Kappa	$\kappa = \frac{p_o - p_e}{1 - p_e}$

10-FCV. Additionally, XAI techniques are utilized to interpret and explain model decisions. Each subsection systematically addresses a critical component of performance validation, ensuring a holistic understanding of the proposed framework's strengths and its suitability for real-world deployment. For a rigorous comparison, we evaluate the performance of each model across multiple metrics, as presented in Table 7 [39].

4.1. Simulation and Discussion

To evaluate the efficacy of the proposed and baseline models, extensive simulations are conducted on the IoSTs datasets using various configurations and optimization methods. The models are evaluated using common evaluation measures, such as log-loss, Receiver Operating Characteristic – Area Under the Curve (ROC-AUC), F1-score, accuracy, precision, and recall. These metrics offer a comprehensive understanding of each model's capability to detect anomalies within imbalanced binary and multi-class intrusion scenarios.

4.2. Justification of Benchmark Models

To evaluate the effectiveness of the proposed models, LRSO, LRALU, and LRALE, a set of widely used benchmark models from both ML and DL domains are selected. These models provide a comprehensive baseline for comparison in terms of performance, interpretability, and computational complexity.

4.2.1. Machine Learning Models

The ML benchmarks include LR, Random Forest (RF), and Extra Tree (ET). LR is a simple, interpretable, and widely adopted linear classifier that serves as a strong baseline for tabular classification tasks. Its probabilistic output is particularly suitable for integration with AL strategies, making it a natural choice for comparing with LRSO, LRALU, and LRALE. Next, RF is an ensemble-based non-linear model that aggregates multiple decision trees to improve generalization and robustness [40]. It is widely used in tabular data classification due to its ability to handle feature interactions and strong out-of-the-box performance. Comparing the proposed models with RF helps assess whether AL and optimization strategies offer advantages over traditional ensemble approaches. Finally, ET is similar to RF but introduces more randomness in the tree construction process, often leading to faster training and reduced variance. Including ET provides insight into whether the proposed models outperform other ensemble variants on tabular datasets [41].

4.2.2. Deep Learning Models

The DL benchmarks include Tabular Network (TabNet), FT-Transformer, AlexNet, and Long Short Term Memory (LSTM). TabNet is a specialized DL architecture for tabular data that uses sequential attention to select salient features [42]. Its interpretability and strong performance make it an ideal benchmark to compare the proposed models against modern DL approaches. Similarly, FT-Transformer leverages self-attention mechanisms for tabular data, enabling it to capture complex inter-feature dependencies, providing a state-of-the-art DL models for evaluating the capability of the proposed LR-based models in handling high-dimensional tabular data [43]. Now, AlexNet, although originally designed for image data, is included to compare performance in highly parameterized networks and to demonstrate the effectiveness of simpler, tabular approaches like LRSO, LRALU, and LRALE [44]. Finally, LSTM networks, capable of capturing sequential dependencies, are suitable for temporal tabular datasets. Including LSTM allows assessment of whether the proposed models can achieve competitive performance against models designed for sequential patterns.

4.3. Performance Analysis of Proposed Models and State-of-the-Art Models on Internet of Sensor Things and CIC-IDS-DS Datasets

The comparative analysis of results presented in Table 8, and Figures 4, 6, 5, 7, and 8 demonstrates the technical superiority of the proposed LRALE, LRALU, and LRSO models over state-of-the-art and baseline ML and DL models in the domain of ID in IoSTs. Specifically, Table 8 reveals that LRALE achieves the highest accuracy, recall, and ROC-AUC of 0.92, highlighting its strength in correctly identifying both attack and normal classes. This performance stems from the entropy-based AL loop that systematically selects highly informative samples based

Table 8
Proposed Models Comparison with Existing Machine and Deep Learning Techniques on IoSTs and CIC-IDS-DS Datasets

Datasets	Models	Accuracy	Precision	Recall	F1-score	ROC-AUC	Log-Loss	MCC	Cohen's Kappa	Execution Time (s)
IoSTs Dataset	TabNet	0.83	0.85	0.84	0.83	0.97	0.33	0.80	0.80	2237.88
	FT-Transformer	0.81	0.85	0.83	0.83	0.98	0.37	0.82	0.82	195.59
	AlexNet	0.76	0.86	0.76	0.78	0.91	0.31	0.61	0.54	1818.77
	LSTM	0.61	0.66	0.61	0.52	0.89	1.25	0.48	0.34	2.59
	RF	0.48	0.91	0.48	0.62	0.82	0.63	0.59	0.49	132.10
	ET	0.75	0.55	0.77	0.60	0.90	0.48	0.43	0.40	38.23
	LR	0.77	0.78	0.77	0.82	0.82	0.55	0.46	0.36	60.41
	Proposed-LRSO	**0.84**	**0.94**	**0.84**	**0.88**	**0.90**	**0.20**	**0.69**	**0.63**	**560.22**
	Proposed-LRALU	**0.91**	**0.94**	**0.91**	**0.92**	**0.94**	**0.29**	**0.59**	**0.57**	**11.29**
	Proposed-LRALE	**0.92**	**0.87**	**0.92**	**0.89**	**0.92**	**0.29**	**0.53**	**0.53**	**0.08**
CIC-IDS-DS	TabNet	0.59	0.59	0.59	0.58	0.60	0.71	0.18	0.18	114.07
	FT-Transformer	0.85	0.87	0.85	0.85	0.96	0.42	0.73	0.71	6.28
	AlexNet	0.88	0.97	0.88	0.92	0.89	0.23	0.33	0.22	186.88
	LSTM	0.68	0.80	0.68	0.65	0.52	0.66	0.47	0.37	9.82
	RF	0.87	0.90	0.87	0.87	0.87	0.69	0.78	0.75	161.03
	ET	0.81	0.54	0.86	0.53	0.92	0.67	0.26	0.14	0.11
	LR	0.95	0.98	0.95	0.96	0.96	0.21	0.56	0.47	10.24
	Proposed-LRSO	**0.97**	**0.98**	**0.96**	**0.96**	**0.99**	**0.20**	**0.58**	**0.50**	**13.46**
	Proposed-LRALU	**0.97**	**0.98**	**0.97**	**0.98**	**0.98**	**0.19**	**0.74**	**0.72**	**0.72**
	Proposed-LRALE	**0.98**	**0.98**	**0.98**	**0.97**	**0.98**	**0.26**	**0.73**	**0.72**	**0.03**

on prediction uncertainty, enabling the model to refine its decision boundaries even with 40% labeled data. The LRALU model, employing uncertainty sampling, shows competitive results with an accuracy of 0.91, F1-score of 0.92, and ROC-AUC of 0.94, indicating its robust capability to generalize across diverse attack types by focusing training on samples where the classifier is most unsure. Meanwhile, LRSO achieves a reduced log-loss of 0.20 and a high precision of 0.94, showcasing the efficacy of SO in fine-tuning hyperparameters that control regularization, ultimately leading to better calibrated probabilistic outputs.

In Figure 4a, the metric-wise performance comparison on the IoSTs dataset clearly shows that the proposed models significantly outperform traditional ones such as TabNet, FT-Transformer, AlexNet, LSTM [45], RF [46], ET, and even baseline LR. The proposed LRSO, LRALU, and LRALE consistently achieve superior scores across all metrics, with F1-scores of 0.88, 0.92, and 0.89, respectively. The weaker performance of the baseline models is evident from the reported metrics. LSTM, designed primarily for sequential data, achieves only 0.61 accuracy and 0.52 F1-score, indicating its inability to capture the relationships in tabular data. RF, although achieving a high precision of 0.91, suffers from poor recall of 0.48 and an F1-score of just 0.62, reflecting its bias toward the majority class in imbalanced datasets. ET shows similar instability with a precision of 0.55 and MCC of only 0.43, suggesting weak discriminative power. AlexNet, intended for image data, delivers moderate results with accuracy of 0.76 and F1-score of 0.78 but is computationally expensive with execution time of 1818.77s without performance gains. Even more advanced models like TabNet with accuracy and F1-score of 0.83 and FT-Transformer with accuracy of 0.81 and F1-score of 0.83, do not incorporate AL or adaptive optimization, leading to inefficiencies in handling scarce labeled data. Baseline LR achieves 0.77 accuracy and 0.82 F1-score, but lacks targeted sample selection and optimized hyperparameters.

In contrast, the proposed models integrate mechanisms to directly address these shortcomings. LRALE employs entropy-based AL to select the most informative samples, LRALU uses uncertainty-driven selection to focus on ambiguous cases, and LRSO leverages a biologically inspired swarm optimization algorithm for effective hyperparameter tuning. As a result, LRALU achieves accuracy of 0.91, F1-score of 0.92, and ROC-AUC of 0.94, with an execution time of just 11.29s. LRALE matches this efficiency with 0.92 recall and 0.89 F1-score, while LRSO maintains competitive results with accuracy of 0.84, F1-score of 0.88 despite more complex optimization. These targeted strategies enable the proposed models to manage class imbalance, improve label efficiency, and refine decision boundaries more effectively, explaining the stark performance gap compared to conventional approaches. In Table 8 and Figure 4b, the comparative analysis of accuracy, precision, recall, and F1-score on the CIC-IDS-DS dataset clearly highlights the performance gap between state-of-the-art and baseline models and the proposed LRSO, LRALU, and LRALE. Among the State-of-the-Art models, TabNet records an accuracy, precision, recall of 0.59, and an F1-score of 0.58, reflecting its limited ability to generalize in the ID setting due to the absence of adaptive sample selection. FT-Transformer improves on these results with an accuracy, F1-score, and recall of 0.85, and precision of 0.87 benefiting from its capacity to model complex feature interactions, yet still constrained by its reliance on large amounts of labeled data without targeted query strategies. AlexNet achieves an accuracy and recall of 0.88, precision of 0.97, and F1-score of 0.92, but its convolutional architecture shows inefficiencies in effectively capturing the complex relationships within tabular ID data, which limits its performance compared to the proposed models.

LSTM, specialized for sequential inputs, underperforms with accuracy and recall of 0.68, precision of 0.80, and F1-score of 0.65, as the non-sequential nature of the dataset limits its temporal modeling advantage. RF and ET produce reasonable accuracy of 0.87 and 0.81, respectively, but show inconsistency across metrics, such as ET's precision of only 0.54 and F1-score of 0.53, suggesting sensitivity to imbalanced data and limited decision boundary refinement. Baseline LR performs better in stability, with accuracy of 0.95, precision of 0.98, recall of 0.95, and F1-score of 0.96, but still lacks advanced mechanisms to optimize learning from scarce and ambiguous samples.

The proposed models exhibit consistently superior results. LRALU achieves accuracy and recall of 0.97, precision and F1-score of 0.98, demonstrating that uncertainty-based AL effectively concentrates training on the most ambiguous and informative instances, refining decision boundaries while reducing annotation cost. LRALE attains the highest accuracy, precision, and recall of 0.98, and F1-score of 0.97, as entropy-based selection identifies samples with maximal prediction uncertainty, enhancing robustness against class imbalance. LRSO matches LRALU in accuracy at 0.97, with precision of 0.98, recall and F1-score of 0.96, benefiting from hyperparameter tuning through the SO algorithm, which adaptively balances bias–variance trade-offs and ensures well-calibrated decision boundaries.

The weaker performance of the state-of-the-art and baseline models can largely be attributed to their lack of targeted data sampling and adaptive optimization. Without AL, these models are forced to train uniformly across all available data, diluting the impact of rare but critical samples and reducing sensitivity to minority class patterns. Moreover, in models like LSTM and AlexNet, architectural designs intended for other data modalities hinder their capacity to fully exploit the statistical structure of tabular ID features. In contrast, the proposed LRSO, LRALU, and LRALE integrate ADASYN for data balancing, AL for selective annotation, and metaheuristic optimization for hyperparameter tuning, enabling them to efficiently focus on challenging samples, optimize learning resources, and improve classification performance across all four metrics. This synergy results in more accurate, precise, and balanced ID, with the added advantage of computational efficiency and better handling of imbalanced datasets.

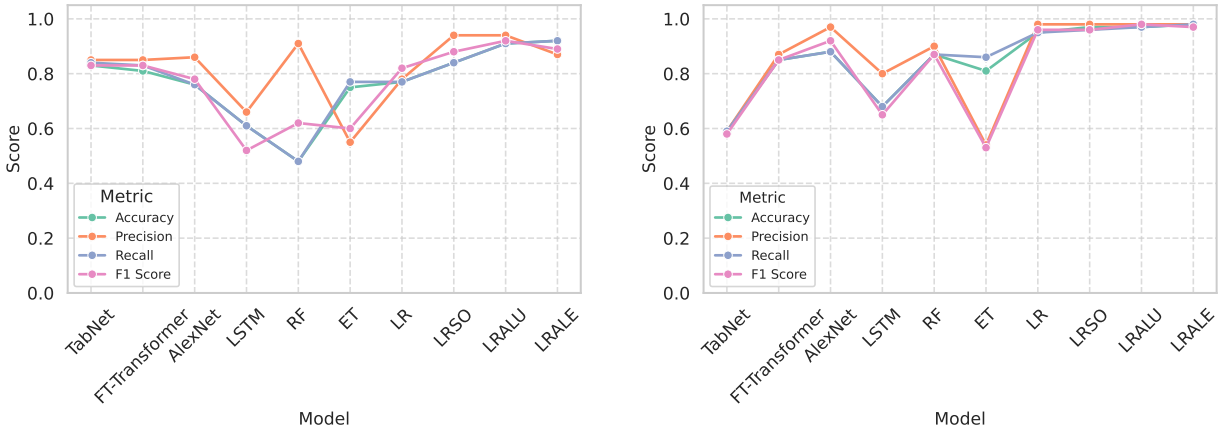
Figure 5a presents the log-loss values on the IoSTs dataset, clearly highlighting the differences in prediction calibration across models. LRSO achieves the lowest log-loss of 0.20, indicating highly reliable probability estimates and strong generalization even in the presence of class imbalance. This is a direct outcome of the swarm optimization process, which enables efficient hyperparameter tuning and robust decision boundary formation. LRALE and LRALU follow closely with log-loss values of 0.29 each, demonstrating that their AL strategies enable well-calibrated predictions by focusing on informative and uncertain samples.

Among the baseline models, FT-Transformer records a log-loss of 0.37, reflecting its ability to capture complex feature interactions, though without targeted sample selection its calibration remains slightly inferior to the proposed approaches. TabNet registers a log-loss of 0.33, suggesting less consistent probability outputs due to limited optimization and no adaptive learning component. AlexNet shows a log-loss of 0.31, indicating moderate calibration but limited suitability for tabular data. ET records a value of 0.48, revealing weaker probability confidence, while LR achieves 0.55, showing that simple linear modeling without targeted learning leads to suboptimal calibration. RF produces a log-loss of 0.63, highlighting poor probability reliability despite good precision, as it struggles with recall and balanced decision thresholds. The highest log-loss is seen in LSTM at 1.25, demonstrating severe calibration issues likely due to mismatches between its sequential architecture and the tabular dataset structure. Overall, the proposed models achieve superior calibration through a combination of AL and optimization mechanisms, ensuring that predicted probabilities align closely with true outcome distributions, a property essential for reliable decision-making in IoSTs applications.

In Figure 5b, the Log loss comparison between existing and proposed models on the CIC-IDS-DS dataset highlights the differences in prediction calibration and confidence. Among the baseline approaches, TabNet records a Log Loss of 0.71, indicating that its probability estimates deviate considerably from the true class labels. FT-Transformer performs better with a Log Loss of 0.42, benefiting from its ability to capture complex feature relationships, though without targeted sample selection it still produces less optimally calibrated probabilities. AlexNet achieves a relatively low Log Loss of 0.23, reflecting confident predictions, yet its convolutional architecture struggles to effectively capture the intricate feature interactions in tabular ID tasks, which constrains its overall suitability. LSTM records a higher Log Loss of 0.66, suggesting significant calibration issues caused by the mismatch between its sequential learning design and the non-temporal dataset. RF and ET obtain Log Loss values of 0.69 and 0.67, respectively, showing moderate calibration but reduced probability reliability in imbalanced scenarios. Baseline LR achieves a strong Log Loss of 0.21, showing stable probability outputs but lacking the refinement gained through adaptive optimization or AL.

The proposed models achieve superior calibration results, producing probability estimates that align closely with actual class outcomes. LRALU achieves a Log loss of 0.19, the lowest among all models, by applying uncertainty-based AL to iteratively focus on samples where the model is least confident, thereby improving probability alignment. LRALE records a Log Loss of 0.26, benefiting from entropy-based AL to target high-uncertainty instances, which enhances probability distribution accuracy even in challenging, imbalanced regions of the feature space. LRSO matches LRALU closely with a Log loss of 0.20, as the SO algorithm fine-tunes LR's regularization parameters, balancing bias and variance for well-calibrated predictions.

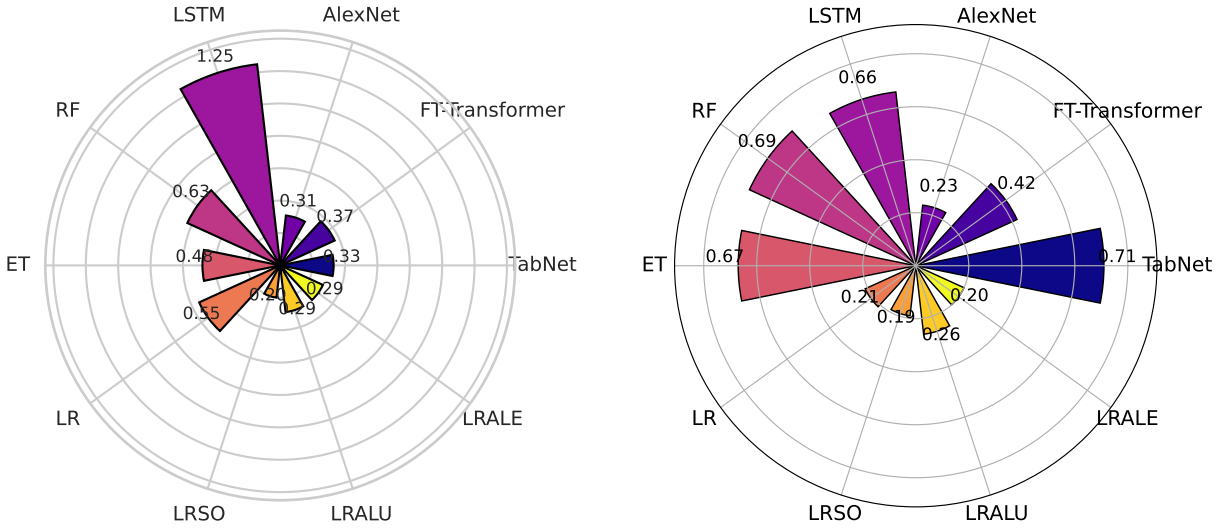
The ROC-AUC curves in Figure 6a on the IoSTs dataset



(a) Performance Comparison of Existing Models and Proposed LRSO, LRALU, and LRALE Models for Intrusion Detection on WSN-DS

(b) Performance Comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for Network Intrusion on CIC-IDS-DS

Figure 4: Performance Comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for WSN-DS and CIC-IDS-DS Datasets

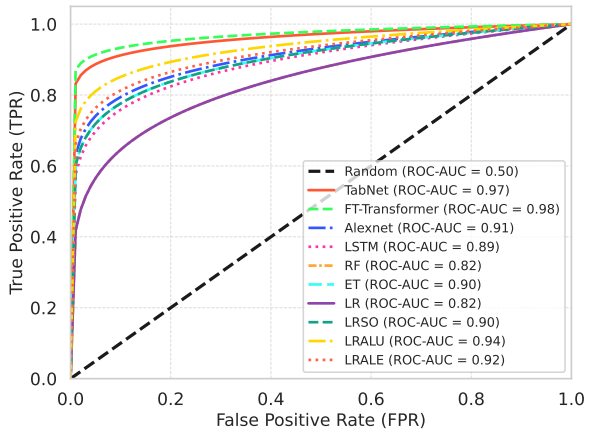


(a) Log-Loss Comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for Intrusion Detection on WSNs

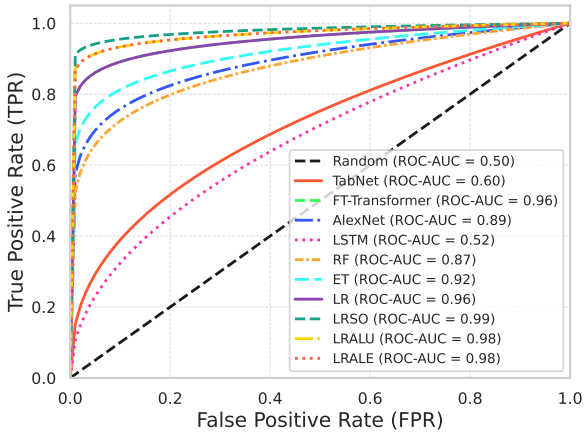
(b) Log Loss Comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for Network Intrusion on CIC-IDS-DS

Figure 5: Log-Loss comparison of Existing and proposed LRSO, LRALU, and LRALE Models for WSN-DS and CIC-IDS-DS Datasets

further reinforce each model's ability to distinguish between normal and attack classes. LRALU achieves a ROC-AUC of 0.94, followed by LRALE at 0.92 and LRSO at 0.90. Among the baselines, FT-Transformer achieves the highest score of 0.98 and TabNet records 0.97, both demonstrating strong discriminative capacity due to their ability to capture complex feature interactions. AlexNet achieves 0.91, showing reasonable class separation but limited adaptability to tabular features. LSTM records 0.89, reflecting moderate discrimination with weaker handling of class imbalance. RF and LR both achieve 0.82, indicating reduced sensitivity to minority class patterns, while ET achieves 0.90 but without consistent calibration across decision thresholds. Although the proposed models do not surpass FT-Transformer and TabNet in absolute ROC-AUC values, their strength lies in the integration of targeted learning mechanisms. In LRALU, uncertainty-based sampling intensifies training on the most ambiguous samples, refining sensitivity and specificity. LRALE's entropy-driven selection emphasizes boundary cases with evenly distributed class probabilities, thereby improving decision sharpness. LRSO optimally tunes regularization, balancing discrimination and preventing bias toward dominant classes. These mechanisms enable



(a) ROC-AUC Comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for Intrusion Detection on WSNs



(b) ROC-AUC Comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for Network Intrusion on CIC-IDS-DS

Figure 6: ROC-AUC comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for WSN-DS and CIC-IDS-DS Datasets

the proposed approaches to maintain competitive ROC-AUC scores while offering robustness in IoSTs scenarios characterized by limited labeled data, dynamic network topologies, and heightened vulnerability to attacks.

In Figure 6b, the ROC-AUC comparison between existing baseline models and the proposed LRSO, LRALU, and LRALE on the CIC-IDS-DS dataset clearly demonstrates the superior discriminative ability of the proposed approaches. Among the existing models, TabNet records a ROC-AUC of 0.60, indicating limited capability in distinguishing between attack and normal traffic. FT-Transformer achieves a much higher ROC-AUC of 0.96 by effectively modeling complex feature interactions, though it lacks targeted sample selection to further enhance performance. AlexNet reaches a ROC-AUC of 0.89, which, while respectable, indicates limited effectiveness in capturing the complex feature dependencies within tabular ID data. LSTM underperforms with a ROC-AUC of only 0.52, showing that its sequential modeling strengths are not well suited to this non-temporal dataset. RF and ET achieve ROC-AUC values of 0.87 and 0.92, respectively, but these results are constrained by their sensitivity to class imbalance. Baseline LR achieves a ROC-AUC of 0.96, demonstrating stable performance yet lacking advanced optimization and adaptive learning mechanisms.

The proposed models deliver the highest ROC-AUC scores, confirming their improved separability between classes. LRALU and LRALE achieves a ROC-AUC of 0.98 by employing uncertainty-based AL and entropy-based AL, which focuses on labeling the most ambiguous samples to refine the decision boundary. LRSO achieves the highest ROC-AUC of 0.99, benefiting from SO, which fine-tunes hyperparameters to optimize generalization and maximize separation between positive and negative classes.

Figure 7a provides deeper insight into model robustness on the IoSTs dataset by comparing the Matthews Correlation Coefficient (MCC) and Cohen's Kappa across all models. LRSO achieves MCC of 0.69 and a Cohen's Kappa of 0.63, indicating strong predictive agreement even in the presence of class imbalance. MCC is particularly valuable in this context as it accounts for all four elements of the confusion matrix, providing a balanced evaluation unaffected by skewed class distributions. The high MCC and Cohen's Kappa values for LRSO are the result of its swarm-based optimization, which fine-tunes the LR regularization coefficient to minimize both false positives and false negatives across classes. LRALU and LRALE also show competitive results, with MCCs of 0.59 and 0.53 and Cohen's Kappa values of 0.57 and 0.53, respectively. Their AL strategies uncertainty-based in LRALU and entropy-based in LRALE ensure that the models prioritize labeling the most informative samples, steadily enhancing class boundary refinement and improving agreement between predictions and actual labels. Among the baseline models, FT-Transformer achieves relatively strong robustness with an MCC of 0.82 and a Cohen's Kappa of 0.83, while TabNet records MCC and Cohen's Kappa of 0.80, reflecting their ability to capture complex feature interactions but without targeted sample acquisition to boost minority class performance. AlexNet records MCC 0.61 and Cohen's Kappa 0.54, demonstrating reasonable balance but higher susceptibility to errors in challenging decision regions. RF achieves MCC 0.59 and Cohen's Kappa 0.49, showing bias toward the majority class, and ET has the lowest MCC among the tree-based methods at 0.43 with a Cohen's Kappa of 0.40, highlighting weaker agreement. Baseline LR records MCC 0.46 and Cohen's Kappa 0.36, indicating limited capability to handle imbalanced classes. LSTM, with MCC 0.48 and Cohen's Kappa 0.34, reflects the lowest robustness among neural models due to its architecture being less suited for tabular and imbalanced data. These results confirm that the proposed models, through a combination

of adaptive sample selection and optimization strategies, achieve consistently higher robustness, making them more reliable for deployment in IoSTs environments where class imbalance and data scarcity are prevalent.

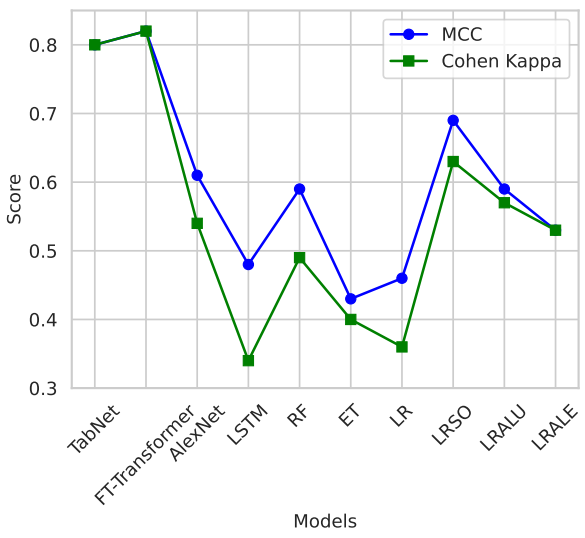
In Figure 7b, the comparison of MCC and Cohen's Kappa on the CIC-IDS-DS dataset reveals important insights into the models' overall reliability and agreement beyond chance. Among the baseline models, TabNet achieves low scores for both metrics, with MCC and Cohen's Kappa values of 0.18, indicating weak correlation between predicted and true classes and poor agreement once random chance is accounted for. FT-Transformer significantly improves these results, obtaining an MCC of 0.73 and a Cohen's Kappa of 0.71, which reflects its stronger classification consistency, though its lack of adaptive sample selection limits further gains. AlexNet shows moderate performance, with MCC of 0.33 and Cohen's Kappa of 0.22. LSTM, while achieving a reasonable MCC of 0.47 and Cohen's Kappa of 0.37, still suffers from the structural incompatibility of sequence modeling on non-temporal features. RF achieves strong MCC and Cohen's Kappa scores of 0.78 and 0.75, respectively, but is prone to bias toward the majority class in imbalanced settings. ET records much lower scores, with MCC of 0.26 and Kappa of 0.14, indicating unstable and inconsistent classification performance. Baseline LR produces an MCC of 0.56 and a Kappa of 0.47, showing good baseline reliability but lacking targeted optimization.

The proposed models demonstrate superior agreement and balanced classification capability. LRALU achieves an MCC of 0.74 and Kappa of 0.72, driven by uncertainty-based AL, which prioritizes difficult and ambiguous samples to enhance model decision boundaries. LRALE matches these results closely, with MCC of 0.73 and Kappa of 0.72, benefiting from entropy-based selection to improve classification consistency across all classes. LRSO records an MCC of 0.58 and Kappa of 0.50, which, while lower than LRALU and LRALE, still surpasses most baseline models due to the effective hyperparameter tuning provided by SO. In Figure 8a, the execution time on the IoSTs dataset provides an essential perspective on the operational feasibility of each model. LRALE achieves the fastest performance with just 0.08 seconds, a result of its lightweight linear regression core and minimal retraining cycles enabled by selective entropy-based querying. LRALU, while slightly slower at 11.29 seconds, remains highly efficient as its uncertainty-based AL loop limits redundant computations by targeting only the most ambiguous samples. The lower execution times of LRALU and LRALE compared to the base LR model are a direct consequence of the data splitting and selective processing strategy. The dataset is initially split into 80% training and 20% testing sets. The training portion is further divided into 40% labeled and 40% unlabeled samples for AL. As a result, LRALU and LRALE only retrain on the labeled subset during each iteration, rather than the full training data, significantly reducing computation. Moreover, LRALE employs a lightweight linear regression core with minimal retraining, while LRALU leverages selective entropy-based querying to focus on the most uncertain samples, avoiding redundant calculations. In contrast, the base LR model processes the entire training set in a single pass without any selective sampling, which explains its comparatively higher execution time despite the simpler model structure.

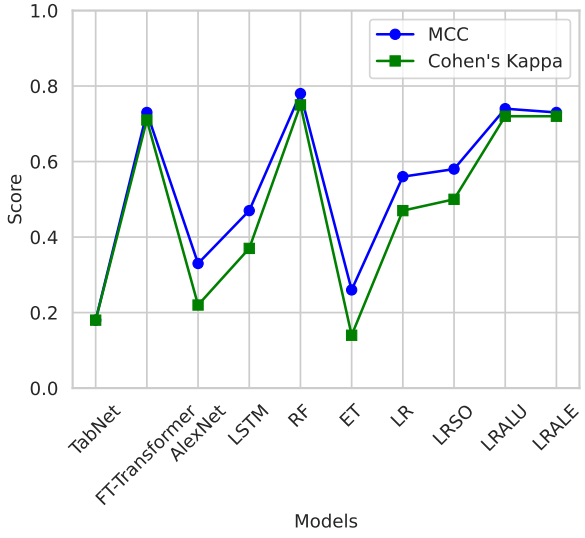
LRSO records a higher execution time of 560.22 seconds due to its iterative swarm optimization process, which explores a large hyperparameter space; however, this computational cost is offset by its strong precision and the lowest log-loss among all models, making it suitable for scenarios where predictive confidence is critical. Among the baselines, LSTM completes in 2.59 seconds but delivers weak performance with an F1-score of only 0.52, indicating that faster execution does not necessarily translate into better detection. RF and ETs require 132.10 seconds and 38.23 seconds, respectively, reflecting the overhead of building multiple decision trees while still achieving limited generalization. LR finishes in 60.41 seconds, but without AL or optimization, its accuracy remains at 0.77. TabNet and FT-Transformer, at 2237.88 seconds and 195.59 seconds, respectively, incur substantial computational costs due to their deep attention-based architectures, despite achieving strong ROC-AUC values. AlexNet is the most time-consuming baseline at 1818.77 seconds, and although it reaches a ROC-AUC of 0.91, it is poorly suited for tabular data and offers no efficiency advantage. These results demonstrate that the proposed models, particularly LRALE and LRALU, achieve a superior balance between predictive performance and computational cost, making them highly practical for real-time IoSTs deployments where both accuracy and speed are critical.

In Figure 8b, the execution time comparison between existing and proposed models on the CIC-IDS-DS dataset underscores the computational efficiency advantages of the proposed approaches. Among the baseline models, TabNet exhibits the execution time of 114.07 seconds, reflecting its complex attention-based architecture and iterative feature selection process. FT-Transformer significantly reduces this to 6.28 seconds by employing efficient transformer-based computation, though its performance is still limited by the absence of AL mechanisms. AlexNet requires 186.88 seconds, one of the slowest among all models, as its convolutional architecture introduces computational overhead and fails to efficiently process the structured feature space of tabular ID data. LSTM completes execution in 9.82 seconds, which is relatively efficient for a DL model, yet its sequential nature still introduces additional processing overhead. RF and ET achieve execution times of 161.03 and 0.11 seconds, respectively, with ET being the fastest baseline model due to its highly parallelizable structure, albeit with performance trade-offs. Baseline LR runs in 10.24 seconds, maintaining reasonable efficiency without specialized optimization.

The proposed models exhibit a clear edge in runtime performance while maintaining high classification accuracy. LRALU executes in just 0.72 seconds, benefiting from the



(a) MCC and Cohan Kappa Comparison of Existing and Proposed LRSO, LRALE, and LRALU Models for Intrusion Detection on WSNs



(b) MCC and Cohen's Kappa Comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for Network Intrusion on CIC-IDS-DS

Figure 7: Correlation metrics comparison of Existing and proposed LRSO, LRALU, and LRALE Models for WSN-DS and CIC-IDS-DS Datasets

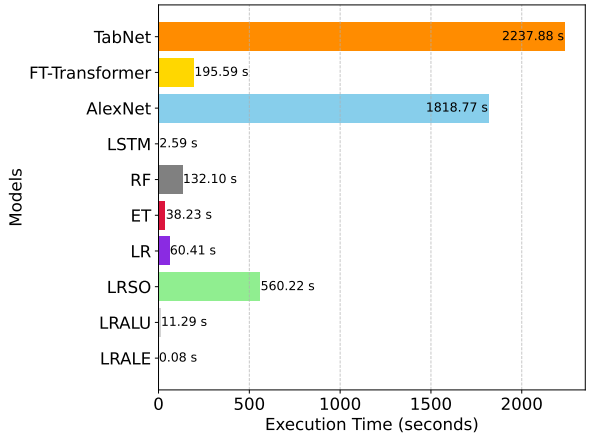
Table 9
Accuracy and Insights for LR, LRSO, LRALU, and LRALE for Intrusion Detection in Internet of Sensor Things Dataset

Model / Accuracy	**Insight**
LR 0.77	Linear decision boundaries provide interpretability but fail to capture complex nonlinear attack patterns; absence of hyperparameter tuning or selective querying limits accuracy.
LRSO 0.84 (**9.09% improvement over LR**)	Employs SO for hyperparameter tuning; strong probabilistic calibration but slightly lower accuracy due to absence of iterative AL in training loop.
LRALU 0.91 (**18.18% improvement over LR**)	Integrates uncertainty-based AL to focus on most ambiguous samples; substantially improves boundary precision and minority-class representation without full retraining overhead.
LRALE 0.92 (**19.48% improvement over LR**)	Entropy-driven sampling maximizes information gain per query; achieves best accuracy by systematically refining decision boundaries in regions of high class overlap.

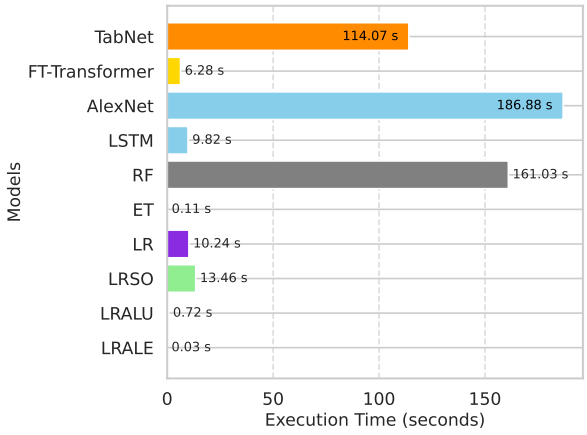
lightweight nature of LR combined with uncertainty-based AL that reduces the training load by selectively labeling only the most informative samples. LRALE achieves the fastest execution time of 0.03 seconds, as entropy-based selection further minimizes the number of required computations per iteration. LRSO records a slightly higher execution time of 13.46 seconds due to the additional overhead introduced by the SO process for hyperparameter tuning, yet this is still competitive when compared to more computationally intensive DL baselines. The insights underlying the accuracy achieved by both State-of-the- Art and proposed models are presented in Table 9.

4.4. Comparative Analysis of Optimization Techniques

The performance of the proposed LRSO is compared with two other optimization techniques, Reptile Optimization (RO) and Dipper Throated Optimization (DTO), in the context of intrusion detection within IoSTs. These evaluations are conducted using the WSNs dataset to assess the effectiveness of each optimization strategy in enhancing model performance under realistic sensor network conditions. This analysis is grounded in various performance metrics. The LRSO model significantly outperforms both the LR with RO (LR-RO) and LR with DTO (LR-DTO) approaches across all evaluated metrics as shown in Table 10. Specifically, the LRSO achieves an accuracy of 0.84, in contrast to the much lower values of 0.31 and 0.41



(a) Execution Time Comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for Intrusion Detection on WSNs



(b) Execution Time Comparison of Existing and Proposed LRSO, LRALU, and LRALE Models for Network Intrusion on CIC-IDS-DS

Figure 8: Execution time(s) comparison of state-of-the-art models with the proposed LRSO, LRALU, and LRALE for WSN-DS and CIC-IDS-DS Datasets

Table 10
Comparison of Different Optimization Techniques

Metric	Proposed LR with RO	Proposed LR with DTO
Accuracy	0.31	0.41
Precision	0.43	0.43
Recall	0.31	0.37
F1 Score	0.45	0.39
ROC-AUC	0.60	0.44
Log-Loss	0.41	0.40
MCC	0.69	0.59
Cohan Kappa	0.39	0.41
Execution Time(s)	22.78	67.43

for RO and DTO, respectively. This trend continues with precision, where LRSO scores 0.94 compared to only 0.43 for both other methods. In terms of recall, the LRSO attains a value of 0.84, which far surpasses the 0.31 and 0.37 achieved by RO and DTO. Furthermore, the F1-score for LRSO stands at 0.88, again considerably higher than 0.45 and 0.39, demonstrating its balanced performance in terms of precision and recall. The ROC-AUC, which reflects the model's discriminative ability, further highlights LRSO's superiority with a value of 0.90 compared to 0.60 and 0.44. Moreover, LRSO records the lowest log-loss of 0.20, indicating highly confident and well-calibrated predictions, whereas RO and DTO methods show higher losses of 0.41 and 0.40 respectively.

In terms of robustness measures, LRSO achieves an MCC of 0.69, a notable improvement over the 0.69 and 0.59 seen with RO and DTO. Although RO shows the same MCC, its overall poor performance in other metrics indicates that this often not be a balanced outcome. Cohen's Kappa also favors LRSO with a value of 0.63, compared to 0.39 and 0.41 for the alternative techniques, signifying stronger agreement between predicted and actual classes. While LRSO's execution time is relatively higher at 560.22 seconds, as compared to 22.78 seconds for RO and 67.43 seconds for DTO, this computational expense is justified by the substantial improvement in predictive performance. The iterative and exploration-intensive nature of SO contributes to its higher execution time, yet it results in superior generalization, accuracy, and calibration of the model. This makes LRSO is an ideal choice for risk-sensitive IoSTs environments where predictive confidence and accuracy are paramount.

The proposed SO-based method outperforms the RO and DTO techniques due to its superior balance between exploration and exploitation in the hyperparameter search space. Unlike the other two methods, SO leverages biologically inspired adaptive velocity updates that allow agents (solutions) to dynamically refine their positions toward the global optimum while avoiding local minima. This mechanism effectively tunes the regularization parameter of the LR model, resulting in improved generalization and classification capability. Furthermore, SO emphasizes global search early on and gradually intensifies local refinement, enabling the model adapts to calibrate its decision boundaries in complex, high-dimensional feature spaces typical of IoSTs intrusion data. As a result, the LRSO model consistently yields higher accuracy, precision, recall, and ROC-AUC scores, while minimizing log-loss, despite the additional computational time required for optimization.

4.5. Comparative Analysis of Active Learning Techniques

The proposed AL techniques, LRALE and LRALU, significantly outperform traditional methods such as Marginal-based AL and Monte-Carlo AL across all key evaluation metrics on the WSNs dataset. LRALE achieves an accuracy

Table 11
Comparison with Different Active Learning Techniques

Metric	Marginal-based AL	Monte-Carlo AL
Accuracy	0.82	0.78
Precision	0.19	0.20
Recall	0.20	0.78
F1 Score	0.16	0.77
ROC-AUC	0.50	0.67
Log-Loss	0.57	1.76
MCC	0.49	0.68
Cohan Kappa	0.57	0.64
Execution Time(s)	16.27	5.83

of 0.92 and LRALU scores 0.91, both substantially higher than the 0.82 of Marginal-based AL and 0.78 of Monte-Carlo AL. In terms of F1-score, which balances precision and recall, LRALE and LRALU record values of 0.89 and 0.92, respectively, sharply contrasting with the notably low 0.16 for Marginal-based AL and 0.77 for Monte-Carlo AL. Precision is highest in LRALU with 0.94, compared to only 0.19 and 0.20 in the two baseline methods, indicating the proposed models' superior ability to minimize False Positive (FPs). Recall also favors LRALU with 0.91 and LRALE with 0.92, while Marginal-based and Monte-Carlo AL lag behind with 0.20 and 0.78. Moreover, the ROC-AUC values 0.94 for LRALU and 0.92 for LRALE underscore their powerful discriminative capabilities, unlike the poor performance of Marginal-based AL at 0.50 and Monte-Carlo AL at 0.67.

The log-loss metric further differentiates the approaches: the proposed models maintain a low value of 0.29, signaling high-confidence, well-calibrated predictions, while the baselines suffer from much higher values, particularly 1.76 for Monte-Carlo AL as shown in Table 11. The proposed techniques also demonstrate robust generalization under class imbalance as reflected in their MCC (0.59 for LRALU and 0.53 for LRALE) and Cohen's Kappa values (0.57 and 0.53), clearly exceeding the baselines' values. Although LRALU has an execution time of 11.29 seconds compared to Marginal-based AL 16.27 seconds and Monte-Carlo AL 5.83 seconds, LRALE is the most computationally efficient at just 0.08 seconds.

The superior performance of the proposed methods can be attributed to their informed sample selection mechanisms: LRALE uses entropy to prioritize samples with maximum prediction uncertainty, effectively refining model decision boundaries, while LRALU focuses on instances where the model is least confident, thereby targeting the most informative points. In contrast, Marginal-based AL and Monte-Carlo AL lack robust strategies for capturing true uncertainty, often leading to sub-optimal sample acquisition and poor generalization. This fundamental limitation causes them to perform poorly in multi-class, imbalanced ID scenarios common in IoSTs, whereas the proposed models adaptively learn from minimal yet informative data, making them far more effective for real-world applications.

4.6. Comparative Analysis of Balancing Techniques

In this study, AdaSyn is selected as the primary data balancing technique using the WSNs dataset because of its adaptive ability to generate synthetic samples in the most challenging regions of the feature space, particularly near decision boundaries and within sparse minority clusters, thereby enhancing minority class representation without distorting feature distributions. Unlike LoRAS, which often fails to emphasize difficult regions and tends to yield weaker generalization, and ProWSyn, which, despite improving class balance, struggles to achieve competitive discriminative power across key metrics, AdaSyn ensures robust generalization and consistently higher predictive performance. This adaptive generation process makes AdaSyn superior for imbalanced scenarios in intrusion detection within IoSTs, where precise learning of complex minority patterns is critical. Accordingly, this study presents a comparative analysis of three data balancing techniques: AdaSyn (used in the proposed models), LoRAS, and ProWSyn. According to the results in Table 8, AdaSyn consistently demonstrates superior performance across all key metrics. When integrated with models such as LRSO, LRALU, and LRALE, AdaSyn yields outstanding results. For instance, LRALU with AdaSyn achieves a 0.91 accuracy, 0.94 precision, 0.91 recall, 0.92 F1-score, 0.94 ROC-AUC, and a 0.29 log-loss. In comparison, the best performance under LoRAS balancing appears with LRALE, recording an accuracy of 0.63, precision of 0.65, recall of 0.75, F1-score of 0.54, and ROC-AUC of 0.69. Similarly, ProWSyn with LRALE obtains a relatively higher accuracy of 0.86, but its F1-score 0.46 and ROC-AUC 0.85 remain lower than those achieved with AdaSyn, as shown in Table 12.

In examining individual metrics, AdaSyn-based models consistently outperform the alternatives. The F1-score, a balance between precision and recall, reaches 0.92 with LRALU, whereas LoRAS and ProWSyn achieve only 0.65 and 0.59, respectively. ROC-AUC, reflecting the model's ability to discriminate between classes, remains highest for AdaSyn-based models, with value of 0.92 for LRALE, while LoRAS and ProWSyn show 0.69 and 0.85 for LRALE at best. Additionally, AdaSyn models exhibit lower log-loss of LRALE of 0.29 compared to LoRAS of 0.70 and ProWSyn of 0.78, indicating more confident and reliable predictions. The MCC and Cohen's Kappa, both crucial for imbalanced datasets, also validate AdaSyn's effectiveness. With AdaSyn, LRSO reaches an MCC of 0.69 and Cohen's Kappa of 0.63, outperforming LoRAS (MCC 0.54, Kappa 0.33) and ProWSyn (MCC 0.72, Kappa 0.71). Although AdaSyn-based models like LRSO consume more computational time of 560.22 seconds, the significant improvements in predictive performance justify the extra overhead, especially in critical applications such as ID in IoSTs.

AdaSyn works better than LoRAS and ProWSyn because it adaptively focuses on generating synthetic data in regions where learning is more difficult, such as near class boundaries or in sparse minority clusters. This enables

Table 12
Comparison of Different Data Balancing Techniques

Metric	LoRAS			ProwSyn		
	LRALE	LRALU	LRSO	LRALE	LRALU	LRSO
Accuracy	0.63	0.61	0.78	0.86	0.77	0.62
Precision	0.65	0.65	0.70	0.63	0.59	0.47
Recall	0.75	0.78	0.67	0.88	0.80	0.59
F1 Score	0.54	0.64	0.57	0.46	0.65	0.52
ROC-AUC	0.69	0.63	0.73	0.85	0.57	0.61
Log-Loss	0.70	0.73	0.63	0.78	0.76	0.45
MCC	0.78	0.63	0.54	0.38	0.50	0.72
Cohan Kappa	0.52	0.61	0.33	0.40	0.73	0.71
Execution Time(s)	45.21	9.78	490.61	12.22	2.78	380.07

the models to learn finer-grained decision boundaries and avoid overfitting on easily classified samples. Notably, in our experimental scenario, empirical results indicate that AdaSyn does not introduce overfitting or distort feature distributions. Classifier performance metrics including F1-score, MCC, and Cohen's Kappa remain consistently high across both minority and majority classes, demonstrating that the synthetic samples generated by AdaSyn are representative and beneficial. This confirms that, under our IoSTs ID setup, AdaSyn enhances minority class learning without compromising the integrity of the feature space or model generalization. In contrast, LoRAS and ProWSyn often fail to emphasize these challenging regions, resulting in weaker generalization. Therefore, AdaSyn not only improves class distribution but also enhances the model's discriminative power, making it the most effective balancing strategy among the three in this context.

4.7. Effect of Different Components of Proposed Models Performance

This section presents a comparative analysis of LR configured in multiple settings on the WSNs dataset, including models without preprocessing and balancing, with preprocessing alone, with preprocessing and ADASYN oversampling, and those incorporating AL strategies (LRALU, LRALE) and optimization (LRSO). The objective is to examine how each incremental component influences classification performance, minority-class sensitivity, and computational efficiency in order to identify the most effective configuration for the target IoST environment. The detailed results for all configurations across multiple evaluation metrics are summarized in Table 13.

The LR without preprocessing and without ADASYN configuration achieves high accuracy of 0.96 and ROC-AUC of 0.98; however, it records a lower recall value of 0.74 and an F1 score of 0.76, indicating a clear bias towards the majority class. The absence of preprocessing means that heterogeneous feature scales, irrelevant noise, and outliers remain untreated, leading to unstable weight estimation and suboptimal decision boundary formation in the multidimensional feature space. Furthermore, without any balancing mechanism, the model is inclined to favour the majority class, which inflates accuracy but results in an uneven trade-off between precision and recall, ultimately limiting the model's capability to generalize to minority-class instances. From a technical standpoint, LR is very sensitive to differences in feature scales. If features are not scaled, some can dominate the learning process and push the model coefficients in the wrong direction. This makes the optimization less reliable and the decision boundary unstable. Outliers and noise also disturb the weight estimation, making the model less accurate. Without balancing the dataset, the model still prefers the majority class, which causes the probability threshold to be biased and reduces the ability to correctly identify minority cases.

When preprocessing is introduced while still excluding ADASYN, as in the LR with preprocessing and without ADASYN variant, normalization and feature scaling improve the numerical stability of the optimization process towards a stable solution, producing slightly more uniform performance across metrics. However, the recall remains fixed at 0.74 because the underlying class distribution is still imbalanced, and the model continues to underrepresent minority-class predictions, demonstrating that preprocessing alone is insufficient to address skewed data distributions. Technically, scaling and normalization make the optimization smoother and more stable, since all features contribute more equally. However, the imbalance problem is still there, so the model continues to miss minority samples. Even though the model becomes computationally more stable, it is still biased because of the skewed data distribution.

In contrast, LR with preprocessing and with ADASYN demonstrates an improved recall of 0.77 and an F1 score of 0.82. Despite this gain, the overall accuracy drops to 0.77 and ROC-AUC declines to 0.82 because oversampling increases data diversity but also injects borderline and potentially noisy examples, which can cause the classifier to produce less distinct separation between classes. The LRALU model, combining preprocessing, balancing, and uncertainty-based AL, produces highly balanced precision of 0.94 and recall of 0.91, leading to a strong F1-score of 0.92. The uncertainty-based query strategy ensures that the model focuses training efforts on ambiguous samples that lie close to the classification boundary, thereby enhancing decision-making in difficult regions of the feature space and improving robustness to unseen data. This comes with only a minor reduction in accuracy to 0.91 compared to the baseline but significantly strengthens minority-class recognition.

The LRALE variant employs entropy-based querying, which actively selects the most information-rich instances based on the distribution of class probabilities. This approach achieves an accuracy of 0.92 and the highest recall of 0.92 while also delivering the fastest execution time of 0.08 seconds. The efficiency gain is achieved by drastically reducing redundant retraining cycles, which is especially valuable in resource-constrained IoST environments. Nevertheless, its precision of 0.87 is slightly lower than that of LRALU because the focus on highly uncertain samples often prioritize borderline cases over clearly separable positive predictions. Finally, LRSO integrates preprocessing, class balancing,

Table 13
Performance comparison of LR with and without preprocessing, without and with ADASYN balancing, and AL strategies LRSO, LRALU, and LRALE across multiple evaluation metrics.

Metric	LR without Preprocessing and without ADASYN	LR with Preprocessing and without ADASYN	LR with Preprocessing and with ADASYN	LRSO	LRALU	LRALE
Accuracy	0.96	0.95	0.77	0.84	0.91	0.92
Precision	0.79	0.78	0.78	0.94	0.94	0.87
Recall	0.74	0.74	0.77	0.84	0.91	0.92
F1 Score	0.76	0.76	0.82	0.88	0.92	0.89
ROC-AUC	0.98	0.98	0.82	0.90	0.94	0.92
Log-Loss	0.11	0.11	0.55	0.20	0.29	0.29
MCC	0.76	0.76	0.46	0.69	0.59	0.53
Cohen Kappa	0.76	0.75	0.36	0.63	0.57	0.53
Execution Time(s)	27.59	31.43	60.41	560.22	11.29	0.08

and a swarm optimization-based hyperparameter tuning process. The inclusion of a global search mechanism fine-tunes model parameters for better generalization, resulting in a precision of 0.94, a MCC of 0.69, and a Cohen's Kappa of 0.63, which represent stronger agreement and balanced predictive power. However, the model records a recall of 0.84, which is lower than that of LRALU and LRALE, suggesting a more conservative classification boundary. Moreover, its execution time of 560.22 seconds highlights the computational cost associated with meta-heuristic optimization. While such parameter tuning can yield an optimally configured model, the excessive time requirements limit its applicability in scenarios that demand real-time or near real-time predictions.

4.8. 10-Fold Cross Validation

10-FCV is a widely used resampling technique in ML and statistical modeling for assessing model generalizability and performance [47]. The dataset is randomly partitioned into ten equally sized folds, as shown in Figure 9. The model is trained and validated 10 times, each time using nine folds for training and the remaining fold for validation [48]. This rotation ensures each sample is used once for validation and nine times for training, reducing overfitting and yielding a more reliable performance estimate than a simple train-test split [49]. After all iterations, metrics (e.g., accuracy, precision, recall, F1-score, MSE) are averaged to produce a final score, reflecting expected performance on unseen data. 10-FCV is extensively applied in model selection, hyperparameter tuning, and benchmarking for both classification and regression tasks [50].

According to Table 14, the 10-FCV results provide a comprehensive and detailed evaluation of the proposed models LRSO, LRALU, and LRALE regarding their generalization, classification robustness, and operational efficiency in IoSTs ID. The average accuracy scores across ten folds on the IoSTs dataset are 0.91, 0.92, and 0.84 for LRSO, LRALU, and LRALE, respectively, indicating that LRALE achieves the highest consistency in correct classifications. In terms of precision, which quantifies the correctness among predicted positive samples, both LRALU and LRSO

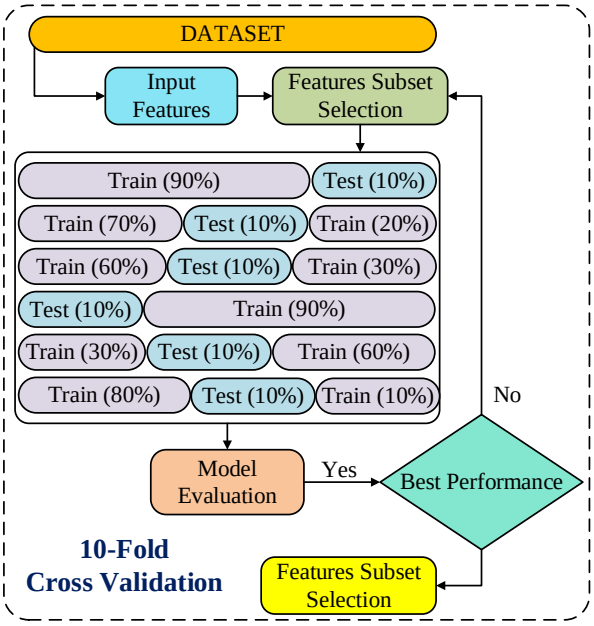


Figure 9: Work flow diagram of 10-Fold Cross Validation for Intrusion Detection in Internet of Sensor Things

demonstrate equally strong performances with an average of 0.94, while LRALE maintains a moderately lower yet acceptable precision of 0.87.

For recall, which measures the model's sensitivity to actual positive instances, LRALU and LRALE again deliver commendable results with averages of 0.92 and 0.91, respectively, surpassing LRSO's recall of 0.84. The F1-score, representing the harmonic mean of precision and recall, further supports these observations, with LRALU achieving the highest average F1-score of 0.92, followed by LRALE at 0.89 and LRSO at 0.88, demonstrating LRALU's superior balance between precision and recall. Moreover, the ROC-AUC values of 0.94 of LRALU, 0.92 of LRALE, and 0.90 of LRSO highlight the models' capabilities in distinguishing between attack and normal instances, reinforcing their classification power in highly imbalanced settings.

In terms of probabilistic confidence, the log-loss metric

Table 14
10-Fold Cross-Validation Results of the Proposed Models with IoSTs and CIC-IDS-DS Datasets.

Models	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Log-Loss	MCC	Cohen's Kappa	ET (s)
IoSTs Dataset									
LRSO	0.84, 0.85, 0.83, 0.84, 0.85, 0.83, 0.84, 0.84, 0.83, 0.84, **0.84**	0.94, 0.93, 0.94, 0.94, 0.95, 0.94, 0.94, 0.94, 0.94, 0.93, **0.94**	0.84, 0.85, 0.84, 0.83, 0.85, 0.84, 0.84, 0.83, 0.85, 0.84, **0.84**	0.87, 0.88, 0.89, 0.88, 0.87, 0.88, 0.88, 0.89, 0.88, 0.88, **0.88**	0.89, 0.90, 0.90, 0.91, 0.90, 0.90, 0.89, 0.90, 0.89, 0.90, **0.90**	0.19, 0.20, 0.21, 0.19, 0.20, 0.21, 0.20, 0.19, 0.21, 0.20, **0.20**	0.68, 0.68, 0.70, 0.69, 0.68, 0.70, 0.69, 0.68, 0.70, 0.69, **0.69**	0.63, 0.62, 0.64, 0.63, 0.62, 0.64, 0.63, 0.62, 0.64, 0.63, **0.63**	560.01, 561.22, 562.24, 562.19, 559.17, 560.21, 561.50, 561.52, 561.57, 560.09, **560.22**
LRALU	0.91, 0.90, 0.91, 0.92, 0.90, 0.91, 0.90, 0.92, 0.91, 0.91, **0.91**	0.94, 0.93, 0.94, 0.95, 0.94, 0.94, 0.93, 0.95, 0.94, 0.94, **0.94**	0.92, 0.91, 0.92, 0.93, 0.91, 0.92, 0.92, 0.93, 0.92, 0.92, **0.92**	0.92, 0.91, 0.92, 0.93, 0.92, 0.91, 0.92, 0.93, 0.92, 0.92, **0.92**	0.94, 0.93, 0.94, 0.95, 0.94, 0.94, 0.93, 0.95, 0.94, 0.94, **0.94**	0.29, 0.30, 0.28, 0.27, 0.29, 0.28, 0.30, 0.27, 0.28, 0.29, **0.29**	0.59, 0.60, 0.58, 0.59, 0.60, 0.58, 0.59, 0.60, 0.58, 0.59, **0.69**	0.57, 0.58, 0.56, 0.57, 0.58, 0.56, 0.57, 0.58, 0.56, 0.57, **0.53**	11.52, 11.08, 11.27, 11.43, 11.37, 11.11, 11.30, 11.20, 11.48, 11.57, **11.29**
LRALE	0.92, 0.91, 0.93, 0.92, 0.91, 0.93, 0.92, 0.93, 0.92, 0.92, **0.92**	0.86, 0.87, 0.88, 0.87, 0.86, 0.88, 0.87, 0.86, 0.88, 0.86, **0.87**	0.91, 0.90, 0.91, 0.92, 0.90, 0.91, 0.91, 0.92, 0.91, 0.91, **0.91**	0.88, 0.89, 0.90, 0.89, 0.88, 0.90, 0.89, 0.88, 0.90, 0.89, **0.89**	0.91, 0.92, 0.93, 0.92, 0.91, 0.93, 0.92, 0.91, 0.93, 0.92, **0.92**	0.28, 0.29, 0.30, 0.29, 0.28, 0.30, 0.29, 0.28, 0.30, 0.29, **0.29**	0.53, 0.54, 0.52, 0.53, 0.54, 0.52, 0.53, 0.54, 0.52, 0.53, **0.53**	0.53, 0.54, 0.52, 0.53, 0.54, 0.52, 0.53, 0.54, 0.52, 0.53, **0.53**	0.09, 0.08, 0.08, 0.08, 0.09, 0.07, 0.08, 0.08, 0.09, 0.08, **0.08**
CIC-IDS-DS Dataset									
LRSO	0.97, 0.98, 0.97, 0.98, 0.98, 0.97, 0.97, 0.98, 0.97, 0.98, **0.97**	0.97, 0.98, 0.97, 0.98, 0.98, 0.97, 0.97, 0.98, 0.97, 0.98, **0.97**	0.97, 0.98, 0.97, 0.98, 0.98, 0.97, 0.97, 0.98, 0.97, 0.98, **0.97**	0.97, 0.98, 0.97, 0.98, 0.98, 0.97, 0.97, 0.98, 0.97, 0.98, **0.97**	0.97, 0.98, 0.97, 0.98, 0.98, 0.97, 0.97, 0.98, 0.97, 0.98, **0.97**	0.16, 0.17, 0.18, 0.10, 0.16, 0.16, 0.18, 0.16, 0.17, 0.17, **0.16**	0.95, 0.96, 0.94, 0.97, 0.96, 0.95, 0.95, 0.96, 0.95, 0.96, **0.95**	0.95, 0.96, 0.94, 0.97, 0.96, 0.95, 0.95, 0.96, 0.95, 0.96, **0.95**	560.01, 561.22, 562.24, 562.19, 559.17, 560.21, 561.50, 561.52, 561.57, 560.09, **560.22**
LRALU	0.96, 0.97, 0.97, 0.96, 0.97, 0.97, 0.96, 0.97, 0.97, 0.97, **0.97**	0.97, 0.97, 0.98, 0.97, 0.97, 0.98, 0.97, 0.97, 0.98, 0.98, **0.97**	0.97, 0.96, 0.97, 0.96, 0.96, 0.97, 0.96, 0.96, 0.97, 0.97, **0.96**	0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, **0.97**	0.97, 0.98, 0.98, 0.97, 0.97, 0.98, 0.97, 0.97, 0.98, 0.98, **0.97**	0.21, 0.20, 0.18, 0.22, 0.19, 0.20, 0.21, 0.19, 0.20, 0.19, **0.20**	0.73, 0.74, 0.75, 0.73, 0.74, 0.74, 0.73, 0.74, 0.74, 0.75, **0.74**	0.71, 0.72, 0.73, 0.71, 0.72, 0.72, 0.71, 0.72, 0.72, 0.73, **0.72**	0.73, 0.71, 0.74, 0.72, 0.73, 0.72, 0.73, 0.73, 0.72, 0.74, **0.73**
LRALE	0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, **0.97**	0.98, 0.98, 0.98, 0.98, 0.98, 0.98, 0.98, 0.98, 0.98, 0.98, **0.98**	0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, **0.97**	0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, 0.97, **0.97**	0.98, 0.98, 0.98, 0.97, 0.98, 0.98, 0.98, 0.98, 0.98, 0.98, **0.98**	0.27, 0.26, 0.25, 0.27, 0.26, 0.25, 0.26, 0.27, 0.26, 0.25, **0.26**	0.74, 0.73, 0.74, 0.73, 0.74, 0.74, 0.73, 0.74, 0.73, 0.74, **0.74**	0.72, 0.72, 0.73, 0.72, 0.72, 0.72, 0.72, 0.73, 0.72, 0.73, **0.72**	0.04, 0.03, 0.03, 0.04, 0.03, 0.03, 0.03, 0.04, 0.03, 0.04, **0.03**

indicates that LRSO with 0.20 produces better-calibrated probability outputs compared to LRALU and LRALE both at 0.29, which is critically beneficial for risk-sensitive decision-making processes. Furthermore, the evaluation of MCC shows that LRSO achieves the highest average of 0.69, suggesting stronger performance in dealing with class imbalance, whereas LRALU and LRALE yield MCC values of 0.69 and 0.53 respectively. Similarly, Cohen's Kappa, which measures the agreement between predicted and actual classifications beyond chance, mirrors the MCC trends, with LRSO achieving 0.63, while LRALU and LRALE obtain 0.53, respectively.

Execution time analysis reveals a trade-off between classification performance and computational efficiency. LRALE exhibits extremely fast execution with an average time of merely 0.08 seconds per fold, highlighting its suitability for real-time or resource-constrained applications. LRALU maintains a reasonable execution time of 11.29 seconds, balancing between accuracy and computational cost. In contrast, LRSO, while delivering the best MCC, Cohen's Kappa, and log-loss, incurs a substantially higher computational cost, averaging 560.22 seconds per fold, primarily due to its complex optimization and AL operations. Additionally, the CIC-IDS-DS dataset as shown in Table 14, the LRSO model demonstrates stable predictive capability with an accuracy, precision, recall, and F1-score of 0.97. Despite these consistent values, the model exhibits a limitation in terms of ROC-AUC which drops to 0.16, highlighting a weaker ability to discriminate between attack and benign classes. Furthermore, the execution time of 560.22 seconds indicates a heavy computational burden, making LRSO less efficient even though its classification metrics remain stable. Cohen's Kappa and MCC values remain at 0.95, reflecting high agreement and correlation but at the expense of efficiency.

The LRALU model shows a slight reduction in classification accuracy with an average of 0.97. Precision, recall, and F1-score values remain strong at 0.97, confirming balanced detection capability across folds. The ROC-AUC increases to 0.20, which represents an improvement compared to LRSO, yet it remains limited in distinguishing classes at a higher granularity. Cohen's Kappa and MCC values are recorded as 0.72 and 0.74 respectively, indicating moderate agreement and predictive correlation. A major advantage of LRALU is its computational efficiency, as the execution time is reduced to 0.73 seconds, considerably lower than LRSO while still maintaining competitive performance.

The LRALE model achieves the most effective performance across the three approaches. Accuracy, F1-score and recall consistently reaches 0.97 and precision improves to 0.98. The ROC-AUC achieves a higher value of 0.26, surpassing both LRSO and LRALU, thereby enhancing class separability. The Cohen's Kappa and MCC values reach 0.72 and 0.74, respectively, confirming stronger inter-class agreement and correlation. The most remarkable improvement is observed in execution time, which is reduced to 0.03 seconds, the lowest among all three models. This demonstrates that LRALE combines high predictive accuracy with superior discriminative ability and extremely efficient computational performance.

4.9. Paired *t*-test Statistical Test

The paired t-test is a statistical technique used to compares the means of two related samples to determine if their difference is statistically significant [51]. In ML, it is commonly applied when evaluating two models on identical datasets or folds (e.g., K-FCV) to assess whether performance differences are due to chance or reflect a true disparity [52]. Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ denote performance metrics of two models across n paired observations, with $D_i = x_i - y_i$ [53]. The test statistic is:

$$t = \frac{\bar{D}}{s_D/\sqrt{n}} \tag{26}$$

where the mean difference is:

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \tag{27}$$

and the standard deviation of differences is:

$$s_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2} \tag{28}$$

where, n is the number of paired samples, D_i the i-th difference, and $\bar{D}$ their mean. The term $(D_i - \bar{D})^2$ measures squared deviation from the mean; summing over n and dividing by $(n-1)$ gives the sample variance, whose square root yields s_D. This test assumes that the differences D_i are approximately normally distributed. A low p-value (typically < 0.05) indicates that the null hypothesis (which states that the models perform equally) can be rejected, implying that there is a statistically significant difference between the two models. According to Table 15, the paired t-test results on IoSTS dataset comparing the baseline LR model with the improved models LRSO, LRALU, and LRALE across several performance metrics are reported. For all metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, the p-values are 0.0000, which are below the significance level of 0.05. The corresponding t-values are negative and large in magnitude (e.g., $t = -31.2064$ for accuracy in LR vs LRALU), indicating that the proposed models consistently outperform the baseline LR model. This consistent statistical significance across all comparisons and metrics demonstrates the robustness of the improvements. The proposed models, LRSO, LRALU, and LRALE, incorporate techniques such as uncertainty-based AL, entropy-based AL, and SO, which contribute to their enhanced generalization and classification performance. The paired t-test thus confirms that the improvements in predictive performance are not merely due to random fluctuations but are statistically significant, justifying their adoption over the traditional LR model in ID scenarios for IoSTs.

For the CIC-IDS-DS dataset, the statistical results presented

Table 15
Statistical Comparison of LR with LRSO, LRALU, and LRALE Models on IoSTs and CIC-IDS-DS Datasets

Dataset	Metric	LR vs LRALU (t, p)	LR vs LRALE (t, p)	LR vs LRSO (t, p)
IoSTs	Accuracy	$t = -31.2064$, $p = 0.0000$	$t = -33.5410$, $p = 0.0000$	$t = -34.0000$, $p = 0.0000$
	Precision	$t = -40.3190$, $p = 0.0000$	$t = -24.3312$, $p = 0.0000$	$t = -44.6389$, $p = 0.0000$
	Recall	$t = -32.1958$, $p = 0.0000$	$t = -34.6535$, $p = 0.0000$	$t = -24.8553$, $p = 0.0000$
	F1-Score	$t = -26.1494$, $p = 0.0000$	$t = -15.9231$, $p = 0.0000$	$t = -15.5840$, $p = 0.0000$
	ROC-AUC	$t = -31.9604$, $p = 0.0000$	$t = -26.6777$, $p = 0.0000$	$t = -28.5314$, $p = 0.0000$
CIC-IDS-DS	Accuracy	$t = -29.8876$, $p = 0.0000$	$t = -35.2214$, $p = 0.0000$	$t = -27.1145$, $p = 0.0000$
	Precision	$t = -33.4521$, $p = 0.0000$	$t = -30.7766$, $p = 0.0000$	$t = -28.9632$, $p = 0.0000$
	Recall	$t = -27.1189$, $p = 0.0000$	$t = -32.8450$, $p = 0.0000$	$t = -25.4720$, $p = 0.0000$
	F1-Score	$t = -24.5568$, $p = 0.0000$	$t = -28.1107$, $p = 0.0000$	$t = -22.9033$, $p = 0.0000$
	ROC-AUC	$t = -30.9821$, $p = 0.0000$	$t = -29.4459$, $p = 0.0000$	$t = -26.7715$, $p = 0.0000$

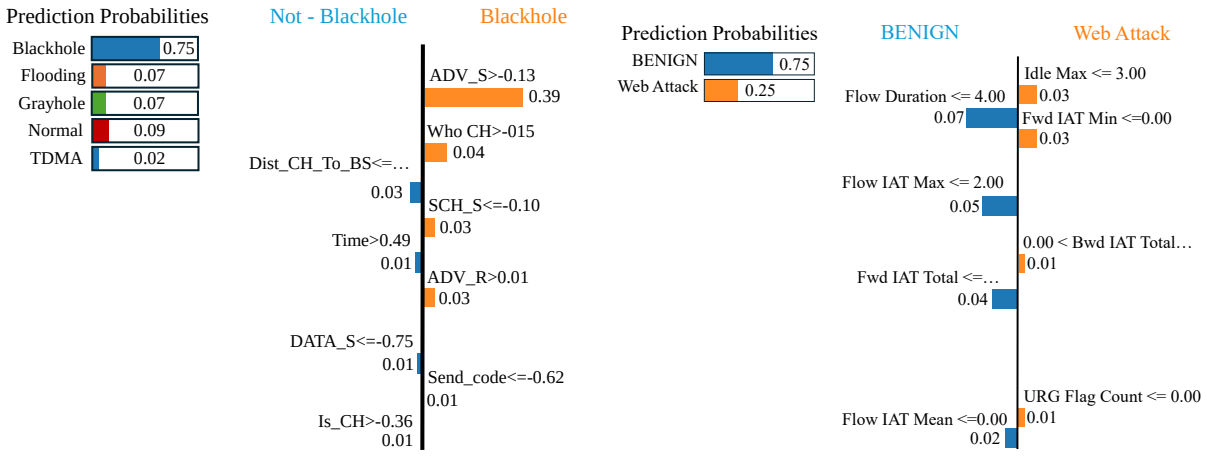
in Table 15 show a clear and consistent improvement of the proposed LRALU, LRALE, and LRSO models over the baseline LR model across all evaluation metrics. The accuracy comparison indicates highly significant differences, where LR vs LRALU produced $t = -29.8876$ with $p = 0.0000$, LR vs LRALE resulted in $t = -35.2214$ with $p = 0.0000$, and LR vs LRSO gave $t = -27.1145$ with $p = 0.0000$. These negative t-values reflect that the proposed models achieved significantly higher accuracy than the baseline LR model, which is crucial in intrusion detection to ensure correct classification of both normal and malicious traffic. Precision results also demonstrate strong statistical significance with $t = -33.4521$, $p = 0.0000$ for LR vs LRALU, $t = -30.7766$, $p = 0.0000$ for LR vs LRALE, and $t = -28.9632$, $p = 0.0000$ for LR vs LRSO. This indicates that the proposed models reduce false positives, an essential aspect in intrusion detection systems where minimizing false alarms is critical. Recall comparisons further support the superiority of the proposed models, with LR vs LRALU showing $t = -27.1189$, $p = 0.0000$, LR vs LRALE with $t = -32.8450$, $p = 0.0000$, and LR vs LRSO with $t = -25.4720$, $p = 0.0000$, meaning that these models achieve higher detection rates for malicious activities. The F1-score, which balances precision and recall, also reflects significant improvements, as shown by $t = -24.5568$, $p = 0.0000$ for LR vs LRALU, $t = -28.1107$, $p = 0.0000$ for LR vs LRALE, and $t = -22.9033$, $p = 0.0000$ for LR vs LRSO, confirming that the proposed approaches offer a more robust balance between correctly identifying attacks and avoiding false alarms. Finally, the ROC-AUC values confirm the overall discrimination ability of the models, with LR vs LRALU at $t = -30.9821$, $p = 0.0000$, LR vs LRALE at $t = -29.4459$, $p = 0.0000$, and LR vs LRSO at $t = -26.7715$, $p = 0.0000$, demonstrating that the proposed models have a much stronger ability to differentiate between normal and malicious traffic. Collectively, these results validate that LRALU, LRALE, and LRSO significantly outperform the baseline LR model on the CIC-IDS-DS dataset, thereby enhancing the reliability and effectiveness of intrusion detection systems.

4.10. Local Interpretable Model-agnostic Explanations

LIME is an effective XAI methodology aimed at improving the interpretability of intricate ML models. While conventional models such as deep neural networks, ensemble approaches, and support vector machines operate as opaque entities, LIME provides localized, understandable explanations for specific predictions without requiring an understanding of the model's fundamental structure [54] [55]. It achieves this by producing perturbed samples surrounding the input instance and acquiring predictions for these samples from the black-box model. Subsequently, LIME constructs an interpretable surrogate model, often a sparse linear model, based on these samples, weighted by their proximity to the original input. This surrogate model estimates the local decision boundary of the black-box model, enabling users to determine which features generally affected the particular prediction. LIME maintains model-agnosticism and emphasizes local fidelity, therefore fostering transparent, accountable, and trustworthy ML systems across various domains [56].

LIME prediction visualization and explanation of proposed models in Figures 10, 11, 12, 13, 14, and 15. Figure 10a illustrates a LIME prediction visualization for the LRALE model on the IoSTs dataset, developed for ID in IoSTs. This figure clarifies the process by which the model generates a certain forecast for a single test instance, categorizing it as a *Blackhole* attacks. It offers an analysis of the most significant features that influence the model's decision-making process. The predicted likelihood for the *Blackhole* class is 0.75, significantly exceeding the probability for other classes: *Flooding* is 0.07, *Grayhole* is 0.07, *Normal* is 0.09, and *TDMA* is 0.02. This result signifies that the LRALE model exhibits substantial confidence in classifying this sample as a *blackhole* attack. LIME assists in determining which input features corroborate or refute this prediction.

In Figure 10b, the LIME visualization highlights the feature-level contributions of the LRALE model for a specific prediction on the CIC-IDS-DS dataset. The model assigns a probability of *0.75* to the *Benign* class and *0.25*



(a) LIME Prediction for LRALE on WSN-DS

(b) LIME Prediction for LRALE on CIC-IDS-DS

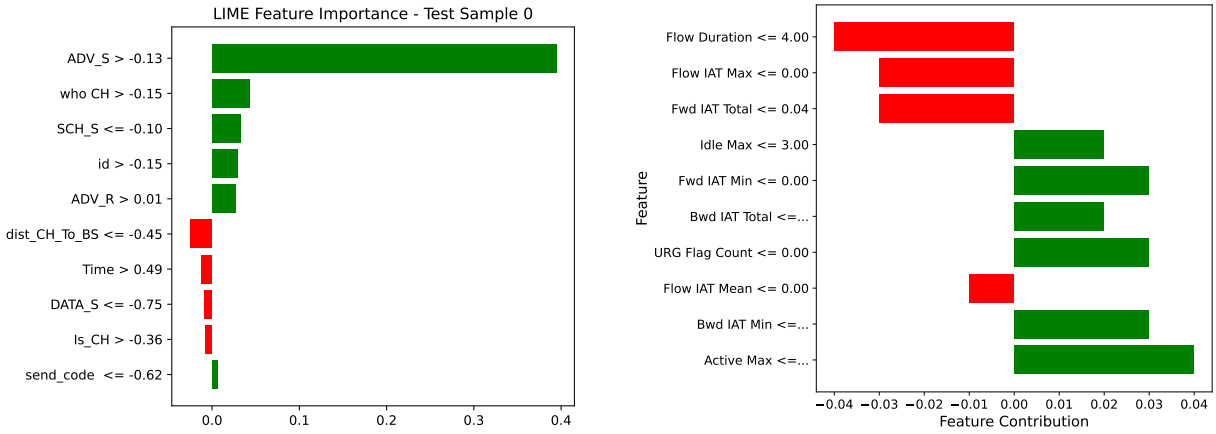
Figure 10: LIME Explanations for Intrusion Detection on WSN-DS and CIC-IDS-DS Datasets

to the *Web Attack* class. The LIME plot emphasizes the most influential features driving this prediction. For the *Benign* outcome, strong negative contributions arise from *Flow Duration* $\leq$ *4.00* and *Flow IAT Max* $\leq$ *2.00*, with the latter contributing a weight of 0.05. On the other hand, for the *Web Attack* outcome, the most relevant supporting features are *Idle Max* $\leq$ *3.00* with a contribution of 0.03, and *Fwd IAT Min* $\leq$ *0.00* with a contribution of 0.03. These results demonstrate that features related to flow duration and inter-arrival times have the most significant influence on the prediction, where shorter flow and lower IAT values tend to align with benign traffic, while idle and forward IAT indicators provide evidence towards potential web attack classification. This interpretability analysis validates that the LRALE model not only achieves high predictive accuracy but also offers transparent insights into the underlying decision process.

Figure 11a presents the LIME summary plot for the LRALE model on the IoSTs dataset, specifically within the context of ID in IoSTs. The summary plot offers a detailed interpretability analysis by quantifying the contribution of each feature towards the model's prediction of intrusion or benign behavior. Among the most influential features identified are *ADV_S*, *who*, *cluster head (CH)*, *SCH_S*, *id*, and *ADV_R*. While LIME is employed, their interpretations are extended beyond descriptive visualization to analytical insights. Specifically, *ADV_S* plays a critical role in distinguishing malicious behavior. In DoS attacks, adversaries flood the network with excessive advertisement messages, leading to abnormal spikes in *ADV_S* values. Similarly, in spoofing scenarios, attackers manipulate the *ADV_S* field to impersonate legitimate nodes, thereby misleading cluster heads during route formation. Hence, the prominence of *ADV_S* in feature importance directly reflects its strong correlation with these attack behaviors, confirming that the explainability results are not only statistical but also consistent with real-world intrusion patterns. The feature *ADV_S* appears to have the most substantial impact, suggesting that adversarial signal strength or source is a critical indicator of abnormal activity. Similarly, the *who* feature, likely representing the identity of the actor or device, strongly contributes to the model's capability to differentiate between normal and intrusive patterns. The features *CH* and *SCH_S*, possibly denoting communication channel characteristics and subchannel statistics respectively, further refine the model's detection sensitivity by capturing nuanced variations in IoST communication behaviors. The *id* feature reinforces device-specific profiling, aiding in differentiating legitimate from malicious entities. Finally, *ADV_R*, potentially indicative of the adversary's response or behavior, also plays a decisive role. Collectively, these features enhance the interpretability and reliability of the LRALE model by elucidating how local predictions are influenced, thus supporting transparency and trust in the detection system.

In IoSTs, *ADV_S*, *who*, *CH*, *SCH_S*, *id*, and *ADV_R* are critical, as they capture adversarial signal strength, device identity, communication channel structure, subchannel statistics, device profiling, and adversary behavior, respectively; these features not only enable accurate detection of anomalous activity but also provide actionable insights for security analysts, allowing prioritization of high-risk devices, identification of suspicious communication patterns, and targeted mitigation of potential attacks.

The LIME summary plot shown in Figure 11b further consolidates the feature contributions across multiple local explanations for the CIC-IDS-DS dataset using the LRALE model. It provides an aggregated view of the most influential features impacting classification decisions. The results indicate that *Flow Duration* $\leq$ *4.00* and *Flow IAT Max* $\leq$ *2.00 (0.05)* consistently push predictions towards the *Benign* class, supporting the higher probability of 0.75 for benign traffic. Conversely, features such as *Idle Max* $\leq$ *3.00 (0.03)* and *Fwd IAT Min* $\leq$ *0.00 (0.03)* contribute positively towards the *Web Attack* class, aligning with the



(a) LIME Summary Plot for LRALE on WSN-DS

(b) LIME Summary Plot for LRALE on CIC-IDS-DS

Figure 11: LIME Summary Plot for Intrusion Detection on WSN-DS and CIC-IDS-DS Datasets

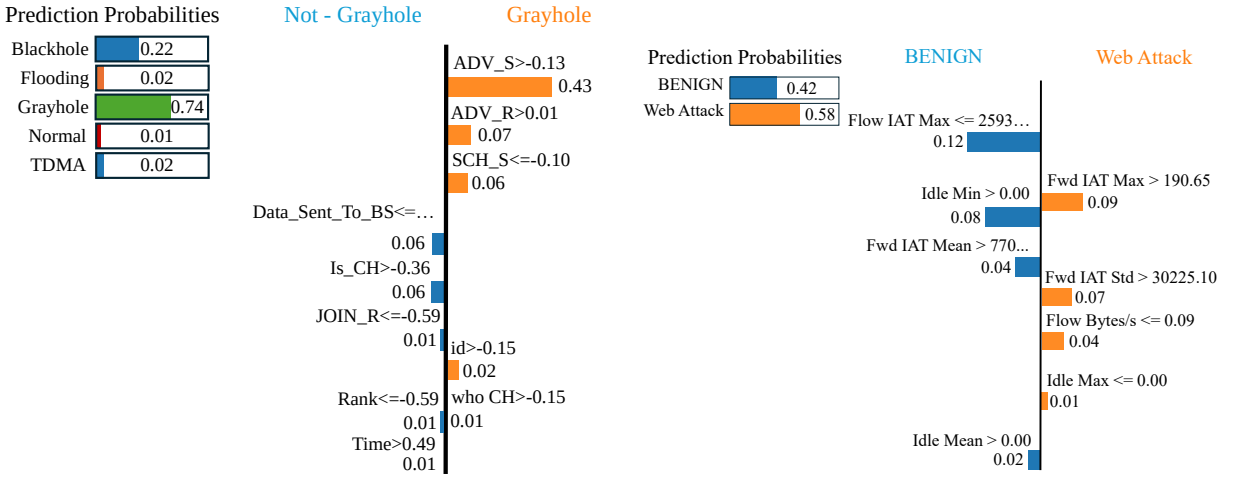
0.25 probability assigned to malicious traffic. This visualization confirms that shorter flow durations and reduced inter-arrival times act as strong benign indicators, while idle and forward IAT measures provide cues for potential web attack detection. By summarizing feature influences, the LIME summary plot enhances interpretability and ensures a transparent understanding of the decision boundaries learned by the LRALE model.

Figure 12a illustrates a LIME prediction explanation for the LRALU model on the IoSTs dataset, which categorizes a particular test occurrence as a *Grayhole* attack in IoSTs. This instance-level interpretability visualization demonstrates that the model, influenced by AL uncertainty-based, assigns a high probability of 0.74 to the *Grayhole* class. Conversely, the anticipated probabilities for the remaining classes are minimal: *Blackhole* at 0.02, *Flooding* at 0.02, *Normal* at 0.01, and *TDMA* at 0.02. This notable disparity indicates the model's effective reliability in categorizing the incident as a *Grayhole* intrusion, defined by selective packet dropping and misleading conduct.

The primary factor driving this prediction is $who\_CH > -0.15$, signifying an unusual pattern in the node's interaction with or identification of its CH. In IoSTs, a reliable connection with an authorized CH is crucial for secure routing. An anomalous value in this feature frequently indicates a hacked or misleading node, consistent with Grayhole attack patterns. A crucial attribute is $ADV\_R > 0.01$, which quantifies the frequency of advertisement responses obtained. An unexpectedly elevated value can indicate fraudulent transmission by the malicious node to erroneously cultivate confidence inside the network.

Indicators such as $is\_CH > -0.36$ and $SCH\_S \leq -0.10$ further substantiate the prediction. The first denotes whether the node is either a CH or not (an aberrant value often mean suspicious activity or role-shifts), whereas the latter will show us whether there was abnormal scheduling traffic (lack of coordination, miscommunication caused by Grayhole activity). Also, $JOIN\_R \leq -0.59$ and $Data\_Sent\_To\_BS \leq threshold$ affect the categorization with moderately strength which signifies compromised or weak data retransmission to the base station which is also the severed feature of Grayhole attacks. When a device or a node is able to connect in the network but does not present the packets properly; it is often an indication of a malicious online action. Moreover, the contextual easily controls the situation with such contextual features as $Time > 0.49$, $ADV\_S > -0.13$, and $Rank \leq -0.59$ indicating that the anomalous temporal behavior, high frequency of the advertisement dispatch, and status of the network being at the hierarchy are all factors that affect the final decision of the model. The identified features provide effective advice in IoST security surveillance: deviations in who_CH and is_CH would show compromised nodes, unusual patterns of ADV_R and ADV_S would reflect deceptive advertisement behaviour, and aberrations in SCH_S, JOIN_R, Data_Sent_To_BS, Time, Rank, which represent infra-structure characteristics, would demonstrate disrupted routing, scheduling, and network hierarchy and together supports analysts to be able to efficiently prioritize and respond to malicious activity.

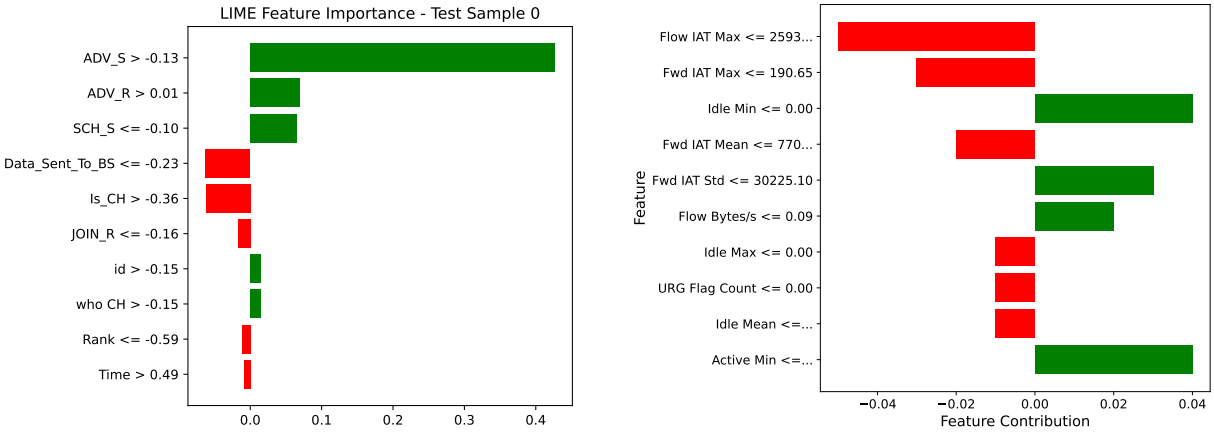
In Figure 12b, the LIME prediction plot illustrates the contribution of key features towards the classification made by the LRALU model on the CIC-IDS-DS dataset. For this specific instance, the model assigns a probability of *0.42* to the *Benign* class and *0.58* to the *Web Attack* class, indicating a stronger inclination towards malicious traffic. The dominant features supporting the *Benign* prediction include $Flow\ IAT\ Max \leq 2593$ with a weight of *0.12* and $Idle\ Min > 0.00$ with a weight of 0.08. In contrast, the most influential features contributing to the *Web Attack* classification are $Fwd\ IAT\ Std \leq 30225.10$ with a contribution of *0.07* and $Fwd\ IAT\ Max \leq 190.65$ with a contribution of *0.09*. These results highlight that while certain inter-arrival time and idle duration features favor benign traffic, variability and minimum values in forward inter-arrival times significantly push the decision towards identifying web attacks. This balance of feature contributions demonstrates how the



(a) LIME Prediction for LRALU on WSN-DS

(b) LIME Prediction for LRALU on CIC-IDS-DS

Figure 12: LIME Prediction for Intrusion Detection on WSN-DS and CIC-IDS-CS Datasets



(a) LIME Summary Plot for LRALU on WSN-DS

(b) LIME Summary Plot for LRALU on CIC-IDS-DS

Figure 13: LIME Summary Plot for Intrusion Detection on WSN-DS and CIC-IDS-DS Datasets

model discriminates between normal and attack traffic with transparent interpretability.

Figure 13a presents the LIME summary plot for IoSTs dataset for the LRALU model, highlighting the relative influence of key features such as *ADV_S*, *ADV_R*, and *SCH_S* within the IoSTs. This plot visualizes of each feature, contribution towards the model's prediction decisions by employing the LIME technique, which offers local fidelity around individual predictions. In the context of LRALU, a LR-based model tailored for user-level analytics in smart environments, *ADV_S* and *ADV_R* representing source and receiver advertisement preferences exert substantial influence, suggesting that personalized advertisement behavior strongly informs user-level learning outcomes. Similarly, *SCH_S*, denoting scheduling preferences or time allocations, plays a pivotal role, indicating its significance in modeling temporal aspects of user interaction. The color gradient in the plot reflects the feature values, while the x-axis denotes their impact on the model output. Features pushing the prediction toward a higher probability of the target class are positioned toward the right, whereas those with negative contributions lie to the left. The interpretability afforded by this visualization is crucial for understanding model behavior in complex, dynamic IoSTs environments, aiding in both debugging and transparency.

The features *ADV_S*, *ADV_R*, and *SCH_S* are domain-critical in IoSTs, as they directly capture the dynamics of user-level interactions and temporal coordination. Specifically, *ADV_S* and *ADV_R* reflect the directionality and frequency of advertisement exchanges, providing insight into abnormal or manipulated communication patterns between devices, while *SCH_S* encodes scheduling behavior, essential for detecting irregular time allocations or coordination anomalies. Collectively, these features enable the model to discern subtle deviations in personalized device behavior, enhancing both predictive accuracy and interpretability in complex, time-sensitive IoST environments.

The LIME summary plot in Figure 13b provides an aggregated view of the most influential features shaping the classification outcomes of the LRALU model on the CIC-IDS-DS dataset. The model outputs a probability of *0.42* for the *Benign* class and *0.58* for the *Web Attack* class, reflecting a stronger likelihood of malicious activity. Among the benign-supporting features, *Flow IAT Max* $\leq$ *2593 (0.12)* and *Idle Min* $>$ *0.00 (0.08)* emerge as the most significant contributors. Conversely, the web attack prediction is predominantly influenced by *Fwd IAT Std* $\leq$ *30225.10 (0.07)* and *Fwd IAT Max* $\leq$ *190.65 (0.09)*. This summary visualization reinforces the insight that while maximum inter-arrival time and idle duration are indicative of benign flows, variations and minima in forward inter-arrival times are strongly associated with attack behavior. By consolidating feature contributions across multiple explanations, the summary plot enhances interpretability and provides a transparent understanding of the model's decision-making process.

Figure 14a presents a comprehensive interpretative visualization for IoSTs dataset employing LIME for the LRSO model, which combines LR with SO for ID in IoSTs. This particular LIME explanation emphasizes the classification of a test event as a Grayhole attack, with the expected probability for the Grayhole class at 0.51, surpassing all other classes: Blackhole is 0.01, Flooding is 0.01, Normal is 0.02, and TDMA is 0.02. The heightened confidence in the Grayhole forecast is influenced by particular feature values that the model identifies as aberrant and suggestive of attack conduct. A significant contributing factor is *Data_Sent_To_BS* $\leq$ *threshold*, which robustly reinforces the categorization. This diminished number signifies an anomalous decrease in the data transmitted by the node to the base station, a trend characteristic of Grayhole attacks that selectively discard or interfere with packet delivery.

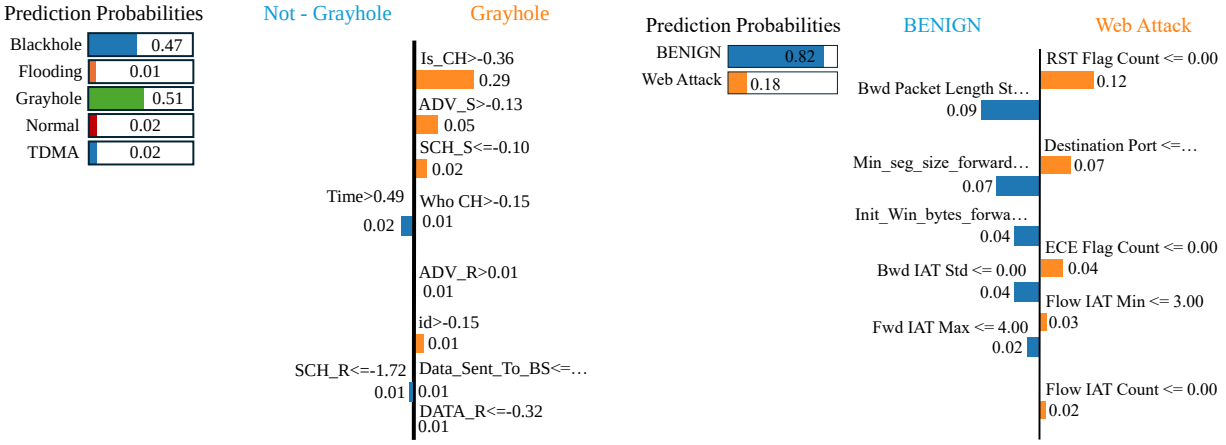
Additional features, like *Is_CH* $>$ *-0.36* and *ADV_S* $>$ *-0.13*, also play a substantial role in the model's decision-making process. The attribute *Is_CH* indicates if the node functions as a CH. Differences in this role, such as unauthorized nodes taking leadership within the cluster, are frequently linked to routing attacks. Likewise, *ADV_S*, which monitors advertisement broadcast activity, when exhibiting abnormal highs or lows, often indicate the fraudulent actions of nodes attempting to mimic real entities or obstruct authentic interactions. Further supporting features comprise *SCH_R* $\leq$ *-1.72*, *DATA_R* $\leq$ *-0.32*, and *CH* $>$ *-0.15*, all of which indicate anomalies in network coordination and communication. These characteristics collectively indicate that the node is engaging in anomalous network participation, either by receiving a reduced number of scheduled messages or by exhibiting dubious behavior within routing hierarchies.

The features *Data_Sent_To_BS*, *Is_CH*, *ADV_S*, *SCH_R*, *DATA_R*, and *CH* are domain-critical in IoSTs, as they capture anomalous packet delivery, cluster leadership deviations, fraudulent advertisement activity, scheduling and data reception irregularities, and suspicious channel interactions, collectively enabling accurate detection and actionable identification of malicious nodes within the network.

In Figure 14b, the LIME prediction plot illustrates the decision explanation of the LRSO model on the CIC-IDS-DS dataset. For this particular instance, the model predicts a probability of *0.82* for the *Benign* class and *0.18* for the *Web Attack* class, demonstrating a strong inclination towards normal traffic. The top contributing features supporting the benign outcome include *Bwd Packet Length Std* with a weight of *0.09* and *Min_seg_size_forward* with a weight of *0.07*. On the other hand, features influencing the web attack classification are *RST Flag Count* $\leq$ *0.00* with a contribution of *0.12* and *Destination Port* $\leq$ *...* with a contribution of *0.07*. These results highlight that backward packet length variability and forward segment size strongly favor benign traffic, whereas reset flag occurrences and destination port characteristics serve as indicators for attack behavior. The explanation confirms that the LRSO model not only achieves accurate predictions but also provides transparent interpretability by revealing the role of critical traffic-related features in shaping its decision.

Figure 15a illustrates the LIME summary plot for IoSTs dataset for the LRSO model, elucidating the influence of specific features on the model's decision-making process in the context of ID within the IoSTs framework. The plot visualizes how variations in individual features contribute to the classifier's prediction confidence, with both the magnitude and direction of the contributions clearly highlighted. Notably, the feature *Is_CH* exhibits a significant impact, often positively influencing predictions towards the detection of intrusion, suggesting that nodes assigned special network roles are critical indicators. Similarly, *ADV_S* and *ADV_R*, presumably referring to advertisement-sending and advertisement-receiving behaviors respectively, play pivotal roles highlighting that anomalous communication patterns during cluster formation stages can be indicative of intrusions. The *SCH_S* also contributes substantively, reflecting that temporal behaviors and transmission scheduling carry diagnostic value for the classifier. The feature *id*, while included, demonstrates comparatively less influence, implying that node identifiers alone are insufficient for robust detection but often provide auxiliary context. Collectively, this plot affirms that the LRSO model captures nuanced behavioral characteristics across the network, leveraging them effectively to discriminate between normal and malicious activities in IoSTs environments. The features *Is_CH*, *ADV_S*, *ADV_R*, *SCH_S*, and *id* are critical in IoSTs, as they capture deviations in cluster leadership roles, advertisement sending and receiving behaviors, scheduling and temporal transmission patterns, and node identity context. These characteristics enable the model to detect subtle anomalies in network coordination and communication, providing both accurate intrusion classification and actionable signals for monitoring potentially compromised nodes.

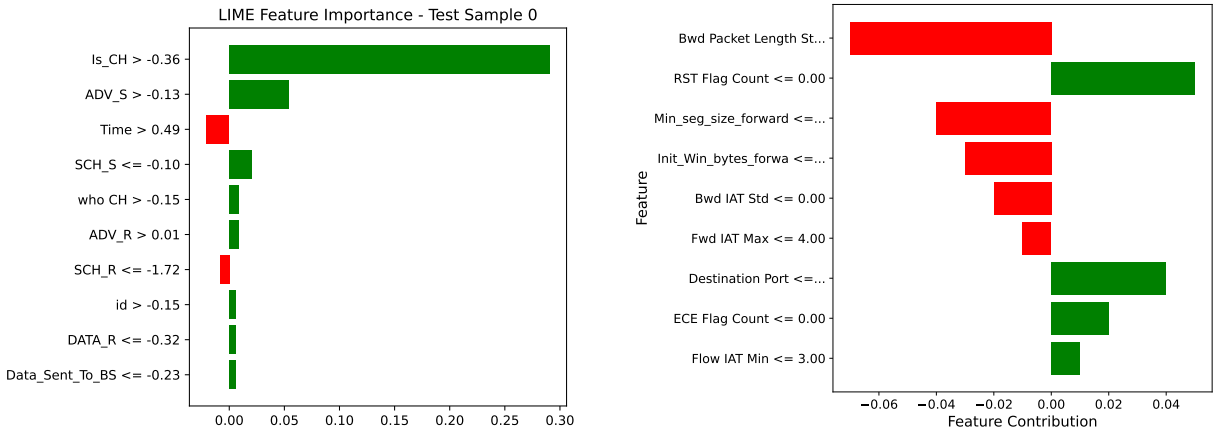
The LIME summary plot in Figure 15b presents an aggregated interpretation of the most influential features affecting the predictions of the LRSO model on the CIC-IDS-DS dataset. Among the top contributors favoring the



(a) LIME Prediction for LRSO on WSN-DS

(b) LIME Prediction for LRSO on CIC-IDS-DS

Figure 14: LIME Prediction for Intrusion Detection on WSN-DS and CIC-IDS-DS Datasets



(a) LIME Summary Plot for LRSO on WSN-DS

(b) LIME Summary Plot for LRSO on CIC-IDS-DS

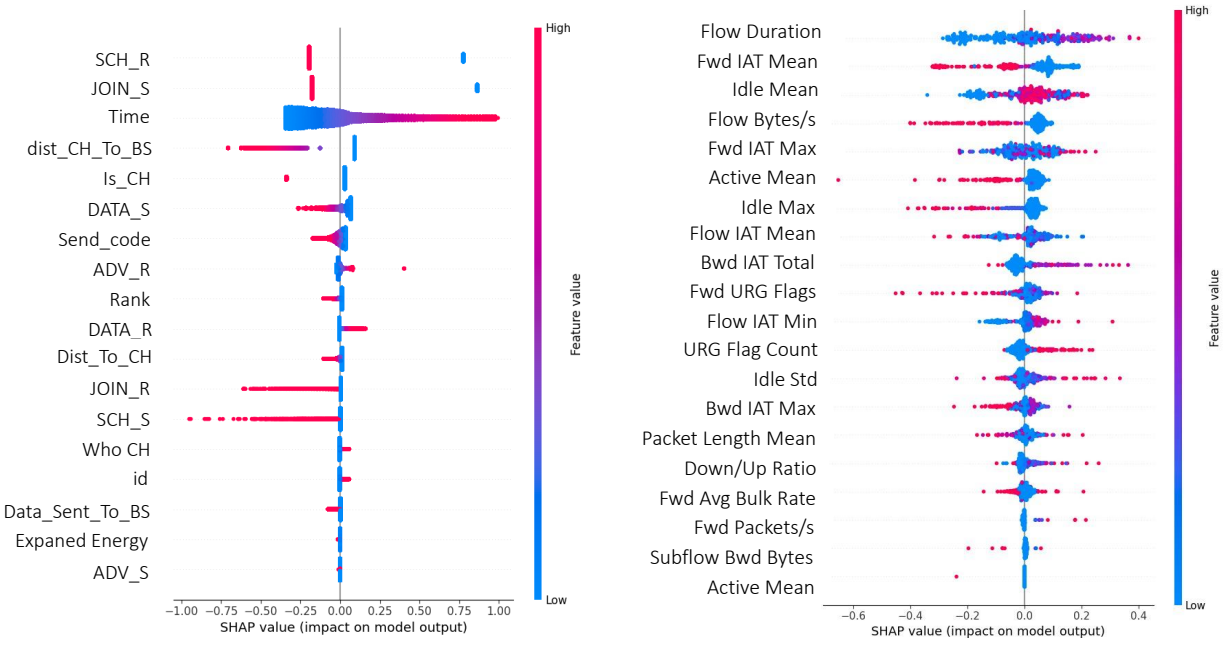
Figure 15: LIME Summary for Intrusion Detection on WSN-DS and CIC-IDS-DS Datasets

benign prediction are *Bwd Packet Length Std (0.09)* and *Min_seg_size_forward (0.07)*. Conversely, the most influential attack-oriented features include *RST Flag Count* $\leq$ *0.00 (0.12)* and *Destination Port* $\leq$ *... (0.07)*. By consolidating feature contributions across multiple instances, the summary plot confirms that variability in backward packet lengths and forward segment size predominantly support benign classification, while reset flag behavior and destination port patterns strongly align with attack traffic. This aggregated explanation enhances transparency by demonstrating the consistency of feature importance in guiding the LRSO model's decision-making process.

4.11. SHapley Additive exPlanations

SHAP is an explainable XAI method that helps interpret the output of ML models by attributing the prediction to individual features. It is based on the concept of Shapley values from cooperative game theory, where the goal is to fairly distribute the payout (i.e., the model's prediction) among all features (considered as players in a game) according to their contributions [57]. In this case, a feature's SHAP value indicates its contribution to the discrepancy between the actual prediction and the baseline, and average prediction [58]. The working of SHAP begins by calculating a baseline prediction, which is typically the expected value of the model output if no features are known. Then, for each feature, SHAP measures the change in the prediction when that feature is added to all possible combinations of the other features in Figure 16, 17, and 18. This helps compute the average marginal contribution of that feature [59]. The result is a set of SHAP values that are additive and consistent, meaning the sum of all SHAP values for a particular instance equals the difference between the instance's prediction and the baseline prediction. Furthermore, the direction of each SHAP value (positive or negative) indicates whether the feature increased or decreased the prediction [60].

While computing exact SHAP values are computationally expensive due to the need to consider all possible feature combinations, the SHAP framework includes efficient



(a) SHAP Violin Plot of LRALE on WSN-DS

(b) SHAP Violin Plot of LRALE on CIC-IDS-DS

Figure 16: SHAP Violin Plot of LRALE on WSN-DS and CIC-IDS-DS Datasets

model-specific implementations. For example, TreeSHAP is a fast and exact algorithm for tree-based models like eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and RF. These optimizations allow SHAP to be practically applied to complex models and large datasets. SHAP values are widely appreciated for their interpretability and transparency. They are used to explain individual predictions (local interpretability) and to understand overall feature importance (global interpretability). This is especially useful in high-stakes fields like healthcare, finance, and legal decision-making, where understanding why a model made a particular decision is just as important as the decision itself.

Figure 16a illustrates the SHAP Voilin plot for IoSTs dataset for the top contributing features in the LRALE model, providing a holistic view of how individual features influence the model's decision-making process in ID for IoSTs. The SHAP values quantify the magnitude and direction of each feature's contribution to the prediction for each data instance. The feature *SCH_R*, denoting the number of schedule messages received by a node, demonstrates a strong positive impact on the prediction of attacks. Higher values of *SCH_R* are often associated with elevated SHAP values, indicating that excessive or abnormal scheduling communication patterns are indicative of potentially malicious behavior. This is expected in IoSTs where tightly regulated communication schedules can be disrupted by attackers seeking to overload or confuse network operations. *JOIN_S*, representing the number of join requests sent by a node, also contributes significantly to the model's predictions. Instances with higher values of *JOIN_S* are frequently linked to positive SHAP values, suggesting that abnormal join activity such as excessive attempts to connect to the network can be symptomatic of compromised nodes or spoofing attacks.

The *Time* feature exhibits a more nuanced contribution. Its SHAP values are distributed across both positive and negative ends, reflecting the temporal complexity of intrusion events. Certain time intervals appear more susceptible to attacks, possibly due to scheduled system operations or routine vulnerabilities that attackers exploit, whereas other periods reflect normal behavior. This temporal sensitivity highlights the model's ability to capture time-dependent attack signatures. *dist_CH_To_BS*, which captures the distance from a CH to the Base Station, impacts the model based on its spatial context. Higher distances are generally associated with positive SHAP values, implying that nodes located farther from the central base station often be more vulnerable due to weaker monitoring or delayed communication, making them potential targets for intrusion.

Lastly, the binary feature *Is_CH*, indicating whether a node is serving as a CH, also has notable influence. The SHAP distribution shows that being a CH is a significant predictor in some cases, with both positive and negative contributions. As CHs manage critical routing and coordination tasks, their compromise can severely impact network integrity, making their operational status a valuable input for ID. Overall, Figure 16a confirms that the LRALE model effectively prioritizes temporally, spatially, and behaviorally relevant features in detecting threats, thereby offering both high predictive accuracy and interpretability suitable for real-world IoSTs security applications. The SHAP selected features in the LRALE model, namely *SCH_R*, *JOIN_S*, *Time*, *dist_CH_To_BS*, and *Is_CH*, are chosen based on

their significant contributions to ID. *SCH_R* captures abnormal scheduling patterns, *JOIN_S* reflects excessive join activity, *Time* encodes temporal variations, *dist_CH_To_BS* measures node proximity to the base station, and *Is_CH* indicates critical cluster roles. SHAP values demonstrate that these features substantially influence model predictions, with positive values increasing attack likelihood and negative values indicating normal behavior, justifying their selection for accurate and interpretable ID.

The SHAP violin plot of the LRALE model on the CIC-IDS-DS dataset as shown in Figure 16b provides a detailed understanding of how individual features influence the classification between benign and web attack traffic. The top five most contributing features are identified as *Flow Duration*, *Fwd IAT Mean*, *Idle Mean*, *Flow Bytes/s*, and *Fwd IAT Max*. The feature *Flow Duration* plays a critical role in ID, as shorter flow durations often correspond to abnormal and abrupt connections, which are commonly observed in malicious traffic. Similarly, *Fwd IAT Mean* captures the average inter-arrival time of forward packets, where unusually small values typically signal automated or scripted attack patterns rather than legitimate user activity. The feature *Idle Mean* is also significant, since lower idle times suggest continuous packet transmissions characteristic of attack flows, whereas higher idle periods are more aligned with normal user behavior.

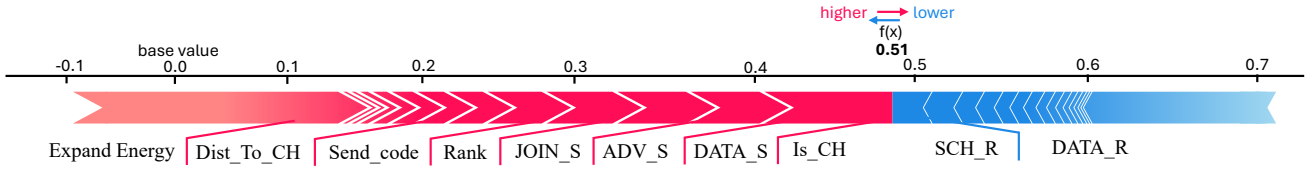
In addition, *Flow Bytes/s* provides strong discriminatory power, as attacks frequently generate atypical byte-per-second patterns, either abnormally high in cases such as brute-force or very low in stealthier attacks, thus serving as a reliable indicator of traffic anomalies. Finally, *Fwd IAT Max*, representing the maximum forward inter-arrival time, helps in distinguishing benign flows from attacks, where extreme values often imply irregular communication bursts associated with malicious activities. Together, these five features dominate the decision-making process of the LRALE model, highlighting that variations in flow duration, packet inter-arrival times, idle behavior, and byte transmission rates are pivotal in accurately detecting web-based attacks. The violin plot not only ranks the importance of these features but also reveals their distributional impact across benign and attack samples, thereby enhancing transparency and interpretability in the model's predictions.

The SHAP force plot for IoSTs dataset shown in Figure 17a demonstrates how various features influence the prediction made by the ML model, resulting in a final output of *0.51*. The base value is shifted according to the contribution of each feature. Features shown in red such as *Expaned Energy*, *Dist_To_CH*, *send_code*, *Rank*, *JOIN_S*, *ADV_S*, *DATA_S*, and *Is_CH* contribute positively, thereby increasing the model's prediction. Among these, *Expaned Energy* and *Dist_To_CH* have the most significant impact. On the other hand, features shown in such as *SCH_R* and *DATA_R* contribute negatively and pull the prediction downward. The net effect of all features yields a final predicted value of *0.51*. It visualizes how individual features collectively drive the model's prediction to *0.51*. Each feature exerts an additive effect on the base value, with *Expaned Energy* and *Dist_To_CH* pushing the prediction upward most strongly, while *SCH_R* and *DATA_R* pull it downward. The plot highlights both the direction and relative magnitude of each contribution, providing an interpretable decomposition of the final output and illustrating which features dominate the decision-making process in this instance.

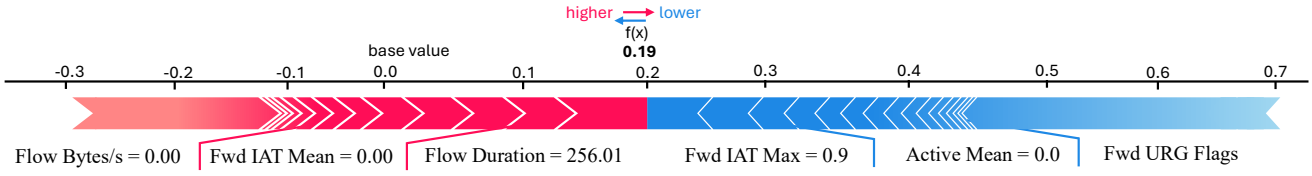
The SHAP force plots for the LRALU on the CIC-IDS-DS dataset as show in Figure 17b explain how individual feature values in a given flow push the model output $f(x)$ away from the dataset-wide base value (expected output) toward either the intrusion class (higher $f(x)$) or the benign class (lower $f(x)$). In these plots, red segments increase $f(x)$ while blue segments decrease it, so the final position of $f(x)$ relative to the base value visualizes the net evidence from all contributing features. For the illustrated flow instance, the temporal and rate-based statistics dominate the explanation. A relatively large value of *Flow Duration* $= 256.01$ contributes positively, indicating a longer-lived connection that is often consistent with attack sessions and therefore pushes $f(x)$ upward. A high *Fwd IAT Max* $= 0.9$ further increases the score by reflecting bursty or irregular packet timing in the forward direction, another hallmark of anomalous behavior that supports the intrusion hypothesis. The presence (non-zero state) of *Fwd URG Flags* also adds positive evidence, as unusual use of urgent flags can accompany crafted or reconnaissance traffic, again nudging $f(x)$ higher. In contrast, very low throughput and inactivity indicators, specifically *Flow Bytes/s* $= 0.00$ and *Active Mean* $= 0.0$, pull the prediction back toward benign by signaling little effective data movement or active periods in the flow. Finally, *Fwd IAT Mean* $= 0.00$ provides a mild stabilizing influence; in combination with the positive contribution from *Fwd IAT Max*, it suggests timing irregularities are driven by peaks rather than the average, leaving the net timing signal still skewed toward intrusion. Collectively, these contributions move $f(x)$ above the base value, yielding an intrusion-leaning decision for this particular flow.

Overall, the explanation is consistent with network ID intuition: longer flow lifetimes (*Flow Duration*) and spikes in packet inter-arrival timing (*Fwd IAT Max*), together with atypical control-flag usage (*Fwd URG Flags*), provide affirmative evidence for malicious activity, while low data rate and inactive periods (*Flow Bytes/s*, *Active Mean*) counterbalance by indicating benign-like behavior. The SHAP force plots thus decompose the LRALU decision into interpretable, flow-level signals directly tied to observable traffic characteristics in CIC-IDS-DS.

Figure 18a presents the SHAP summary plot for IoSTs dataset for the LRSO model, which integrates LR with the SO algorithm for enhanced ID in IoSTs. This plot provides a comprehensive overview of how the model attributes importance to different features across multiple predictions. Among the critical features influencing the model's decisions is *dist_CH_To_BS*, which measures the distance from a node's CH to the base station. High values in this feature can lead to reduced communication reliability and

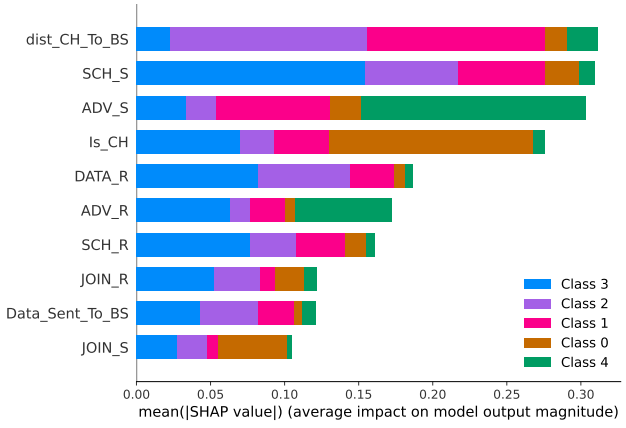


(a) SHAP Force Plot of LRALU on Wireless Sensor Networks Dataset

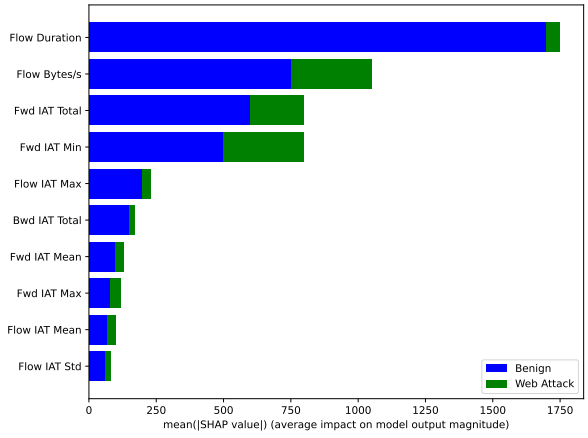


(b) SHAP Force Plot of LRALU on Network Intrusion (CIC-IDS-DS) Dataset

Figure 17: SHAP Force of LRALU for WSN-DS and CIC-IDS-DS Datasets



(a) SHAP Summary Plot of LRSO on WSN-DS



(b) SHAP Summary Plot of LRSO on CIC-IDS-DS

Figure 18: SHAP Summary of LRSO for WSN-DS and CIC-IDS-DS Datasets

increased vulnerability, particularly in distant nodes with weaker monitoring, thereby contributing positively to the prediction of intrusions.

The feature *SCH_S*, representing the number of scheduling messages sent by a node, also plays a significant role. Lower values of *SCH_S* are typically associated with suspicious behavior, such as interrupted or suppressed coordination signals, which are often exploited in network-layer attacks. This suppression often suggest misbehavior or malfunctioning due to malicious intent. Similarly, *ADV_S*, which denotes the frequency of advertisement messages sent by a node, shows clear patterns of influence. Elevated advertisement broadcasts often be indicative of a node attempting to assert false control or mislead neighbors common in deceptive routing attacks like Grayhole or Blackhole. The SHAP values for this feature illustrate both positive and negative effects depending on the intensity and context of the advertisement activity. *Is_CH*, a binary indicator denoting whether a node is currently acting as a CH, is another highly impactful attribute.

CHs are central to communication aggregation and routing, and any unauthorized or abnormal role assignment significantly increases the suspicion of compromise. Irregularities in this feature thus often push the model's prediction towards the intrusion class. Lastly, *DATA_R*, which quantifies the number of data packets received, provides insights into potential disruptions in data flow. Lower values in this feature often suggest the presence of selective forwarding or jamming behavior, where a node deliberately avoids receiving or processing expected messages. Such behavior aligns with the characteristics of several data-plane attacks in IoSTs. In the LRSO model, SHAP analysis highlights features that critically modulate prediction outcomes. *dist_CH_To_BS* identifies nodes with weakened connectivity prone to attacks, *SCH_S* signals irregularities in scheduling messages, *ADV_S* reflects abnormal advertisement behavior potentially linked to routing deception, *Is_CH* denotes central nodes whose compromise can affect network control, and *DATA_R* captures anomalies in data reception indicative of selective forwarding or jamming. The SHAP values quantify each feature's contribution, distinguishing which factors most strongly drive the model

toward classifying an instance as an intrusion versus normal activity, thereby supporting interpretable and targeted ID. Collectively, these features contribute to a transparent and reliable decision-making process in the LRSO model. The SHAP summary plot validates the model's sensitivity to temporal, spatial, and protocol-level behaviors critical for detecting anomalies in dynamic and resource-constrained IoSTs environments.

The SHAP summary plot as shown in Figure 18b highlights the six most influential features that drive the predictions of the LRSO model on the CIC-IDS-DS dataset. These features consistently exhibit the highest average impact magnitudes on the model output, reflecting their strong discriminative role in distinguishing between benign and attack traffic. The feature *Flow Duration* emerges as the most important factor. A higher standard deviation of inter-arrival times within a flow signals irregular or bursty traffic behavior, often associated with attack sessions. Consequently, this feature strongly increases the likelihood of predicting an intrusion when its values are large. Closely following in importance is *Flow IAT Mean*, which measures the average inter-arrival time of packets in a flow. Elevated mean values indicate slower, less uniform traffic patterns that deviate from typical benign flows and thus shift the prediction toward malicious activity.

The third-ranked feature, *Fwd IAT Max*, captures the maximum inter-arrival time between forward packets. Large spikes in this feature point to abnormal packet timing irregularities that align with the traffic characteristics of many web-based and flooding attacks. Similarly, *Fwd IAT Mean*, ranked fourth, reflects the overall average packet timing in the forward direction; unusually high values in this feature contribute to intrusion likelihood, whereas consistently low values tend to align with benign flows. The fifth most important feature is *Bwd IAT Total*, representing the cumulative inter-arrival time for backward traffic. Larger totals indicate extended delays or abnormal response patterns, which can be symptomatic of compromised or targeted hosts under attack. Finally, *Flow IAT Max* is identified as the sixth most influential feature. High maximum inter-arrival times for the overall flow reflect abrupt inactivity periods within communication, often associated with reconnaissance or slow-rate attacks. In summary, the dominance of inter-arrival time based features underscores that timing irregularities across both forward and backward directions serve as primary signals for ID. The LRSO model leverages these temporal dynamics to effectively differentiate normal user activity from malicious traffic in the CIC-IDS-DS dataset.

4.12. Comparative Analysis of LIME and SHAP

LIME and SHAP are two prominent methods used to interpret ML models, particularly in the domain of XAI. While both aim to provide local interpretability for individual predictions, their underlying methodologies and theoretical foundations differ significantly, impacting their behavior, fidelity, and computational efficiency. LIME approximates the local decision boundary of a complex *black-box* model by fitting an interpretable surrogate model, such as a linear regression, around the neighborhood of the instance being explained. It perturbs the input data and uses the corresponding predictions to train the surrogate model. The weights assigned to the perturbed samples are based on their proximity to the instance of interest, typically measured via a kernel function. LIME is model-agnostic and flexible in the choice of surrogate models and distance metrics, making it widely applicable. However, LIME suffers from instability due to its reliance on sampling and its sensitivity to the choice of neighborhood kernel and parameters. Moreover, it does not guarantee consistency or additivity, and its explanations often vary significantly for small perturbations of the input.

In contrast, SHAP operates as a global interpretability method grounded in cooperative game theory, attributing feature contributions based on the concept of Shapley values [61]. It provides a theoretically sound framework that ensures properties such as local accuracy (additivity), consistency, and missingness. SHAP computes the contribution of each feature by considering all possible subsets of features and averaging their marginal contributions. While this leads to a strong theoretical foundation, it is computationally expensive, especially for models with a large number of features. To address this, SHAP includes approximations such as KernelSHAP (a model-agnostic variant using weighted linear regression). Overall, SHAP offers more robust and theoretically consistent explanations than LIME, albeit at a higher computational cost. LIME, on the other hand, provides a quicker and more flexible solution suitable for exploratory analysis, but its lack of guarantees often limit its reliability in high-stakes applications. The choice between the two often depends on the trade-off between interpretability accuracy and computational feasibility, as well as the specific requirements of the application domain.

While SHAP and LIME provide valuable insights into model interpretability, their application in IoST environments presents certain limitations. LIME operates on a single instance at a time, making it relatively less computationally expensive, whereas SHAP is significantly more computationally intensive due to its requirement for exhaustive feature contribution calculations. Moreover, both SHAP and LIME typically require repeated perturbations of the input data, potentially leading to latency issues in real-time decision-making scenarios. The explanations generated are also sensitive to the quality and distribution of the input data; in highly heterogeneous IoST networks, this often results in inconsistent or misleading interpretations. Additionally, the local explanation focus of these methods often overlooks complex global patterns, which can be critical for system-wide IoST security and reliability. Recognizing these constraints is essential for contextualizing their effectiveness within IoST applications.

Table 16
Comparison of Optimized and Ablation Hyperparameters for Proposed Models

Hyperparameter	LR		LRALU		LRALE	
	Optimized	**Ablation**	**Optimized**	**Ablation**	**Optimized**	**Ablation**
Solver	liblinear	saga	liblinear	saga	liblinear	saga
Multi-class Strategy	OvR	OvR	OvR	OvR	OvR using OneVsRest-Classifier	OvR
Penalty	L2	None	L2	None	L2	None
Regularization Strength (C)	1.0	1000.0	1.0	1000.0	1.0	5000.0
Maximum Iterations (max_iter)	100	50	100	30	100	20
Tolerance (tol)	0.0001	0.01	0.0001	0.0001	0.0001	0.01
Initial Labeled Samples	–	–	100 samples randomly selected	50	–	30
Query Strategy	–	–	Uncertainty Sampling	Uncertainty-based	Entropy Sampling	Entropy-based
Samples Queried per Iteration (Top-N)	–	–	50	10	–	5
Active Learning Iterations	–	–	5	5	–	3

4.13. Ablation Study: Evaluating the Impact of Hyperparameter Variations on Model Performance

This ablation study provides a rigorous and comprehensive analysis of how different hyperparameter configurations impact model behavior and performance in the proposed ID framework. Table 16 reports the empirical outcomes resulting from the alternative hyperparameter settings. The models trained with these settings suffer from significantly deteriorated performance across all key evaluation metrics.

The accuracy, precision, recall, and F1-scores exhibit substantial drops compared to Table 8 results, highlighting the models' failure to effectively distinguish between normal and malicious activities. ROC-AUC scores decrease sharply, indicating weakened discriminative ability, and log-loss values increase, confirming poor probability calibration and heightened prediction uncertainty. The MCC and Cohen's Kappa values also decline, reflecting reduced robustness and greater misclassification rates, particularly in minority classes. Furthermore, the inefficacy of AL loops under these settings becomes apparent: due to miscalibrated uncertainty scores, the models query redundant or low-information samples, leading to slower and less effective model updates. These findings validate that careful hyperparameter optimization is not merely beneficial but fundamentally necessary for achieving stable, generalizable, and efficient ID models in IoSTs.

Table 16 presents the initial hyperparameter settings used for training the models, where key selections include the *liblinear* solver, a regularization parameter $C = 0.1$, the *ovr* strategy for multi-class handling, and a maximum iteration count of 1000. These hyperparameters are meticulously tuned to stabilize the optimization process, balance model complexity, and ensure computational efficiency.

Table 17
Performance comparison of Logistic Regression variants

Metric	LR	LRSO	LRALE	LRALU
Accuracy	0.93	0.90	0.79	0.74
Precision	0.94	0.92	0.67	0.88
Recall	0.93	0.90	0.79	0.74
F1 Score	0.93	0.91	0.72	0.77
ROC-AUC	0.99	0.99	0.96	0.92
PR-AUC	0.98	0.97	0.97	0.92
Cohen's Kappa	0.89	0.85	0.63	0.83
MCC	0.89	0.86	0.66	0.83
Log Loss	0.17	0.21	0.45	0.47
Execution Time (s)	51.30	35.59	0.20	11.49

The *liblinear* solver is particularly well-suited for small to medium-sized datasets with sparse features, as it guarantees convergence in convex optimization landscapes typical of LR. The $C = 0.1$ value provides moderate regularization, preventing overfitting while allowing sufficient model flexibility to capture subtle patterns in ID scenarios. Employing the *ovr* strategy simplifies multi-class classification into manageable binary classification tasks, improving sensitivity towards minority classes. Setting *max_iter* to 1000 ensures full convergence without premature stopping, particularly important when handling complex decision boundaries in high-dimensional feature spaces.

Table 16 presents a set of alternative hyperparameters employed to study the sensitivity of model performance to sub-optimal configurations and simulation results are shown in Table 17. In these settings, different solvers such as *saga* or *newton-cg* are utilized, regularization parameters are improperly adjusted, and alternative multi-class strategies like *multinomial* are experimented with. These choices

negatively impact the optimization dynamics. Specifically, solvers like *saga* introduce instability when applied to relatively small, sparse datasets, as they are originally designed for large-scale problems and non-convex settings. Inappropriate regularization either over-constrains the model (if C is too small), leading to underfitting, or makes the model too flexible (if C is too large), resulting in overfitting and poor generalization. Similarly, using *multinomial* instead of *ovr* distorts the model's ability to focus on per-class margins, weakening boundary separation, especially in imbalanced settings. These poor hyperparameter choices degrade both the calibration and generalization capabilities of the models. As a result, the AL process becomes inefficient: uncertainty or entropy estimations are no longer reliable, leading to the selection of redundant or uninformative samples during annotation.

5. Conclusion

This study proposes a robust and efficient ID framework tailored to the unique constraints and dynamic nature of IoSTs environments. To address class imbalance in the data, the framework incorporates the ADASYN technique, which generates class-sensitive synthetic instances to enhance minority class representation. Next, model efficiency is further improved through the integration of LR with two AL strategies: uncertainty-based LRALU and entropy-based LRALE. These strategies selectively enrich the training dataset with highly informative samples, thereby reducing manual annotation effort while maximizing learning effectiveness. Moreover, the LRSO model leverages SO to fine-tune LR hyperparameters, leading to improved accuracy and generalization capability. Extensive simulations demonstrate that the proposed models consistently outperform state-of-the-art approaches across multiple performance metrics on two benchmark datasets: WSN-DS and CIC-IDS-DS. On the IoSts dataset, the LRALE model achieved an accuracy of 0.87, recall and ROC-AUC of 0.92, while the LRALU model attained an accuracy and recall of 0.91, and ROC-AUC of 0.94. The LRSO model on the same dataset yielded an accuracy and recall of 0.84, and ROC-AUC of 0.90. On the CIC-IDS-DS dataset, the LRALE model reached an accuracy,recall, and ROC-AUC of 0.98, the LRALU model obtained an accuracy and recall of 0.97, and ROC-AUC of 0.98, while the LRSO model recorded an accuracy of 0.97, recall of 0.96, and ROC-AUC of 0.99. To ensure the generalizability of the proposed models, a 10-FCV technique was employed, validating performance consistency across diverse data splits. Furthermore, proposed models transparency is enhanced through the application of LIME and SHAP, which offer comprehensive insights into feature importance and decision application. The effectiveness of the proposed framework is statistically substantiated using paired t-tests, confirming its superiority over baseline methods with high confidence. These results underscore the potential of integrating synthetic oversampling, AL, and meta-heuristic optimization in developing intelligent and resource-efficient IDS, making the framework well-suited for real-world IoSTs applications.

5.1. Limitations

Although the proposed ID framework demonstrates promising results, certain limitations need to be addressed in future research. A primary concern is the challenge of handling highly imbalanced datasets, where traditional balancing methods often not fully mitigate the bias towards majority classes. Another limitation is the computational overhead introduced by the feature-rich dataset and the LRSO optimization, which increases execution time and often hinder scalability in real-time IoST environments. Moreover, the current framework has not been comprehensively tested for adaptability across diverse datasets, limiting its generalizability in heterogeneous IoT scenarios. In addition, resource utilization metrics such as CPU usage, memory consumption, and energy efficiency have not been explicitly evaluated, which restricts the assessment of its suitability for deployment in resource-constrained sensor nodes. Lastly, lightweight deployment aspects such as model compression, quantization, and pruning have not yet been explored, which often pose challenges for integration into tiny IoT devices without sacrificing detection accuracy.

5.2. Future Work

In future work, we plan to explore various advanced techniques to further enhance the performance of the proposed ID framework. One key direction is the application of alternative data balancing methods to address class imbalance more effectively in highly skewed datasets. Additionally, we aim to apply dimensionality reduction techniques to remove redundant features and reduce computational complexity, thereby improving model efficiency. Another important direction is the integration of transfer learning to evaluate the generalizability of the proposed models across different ID datasets [62]. This will help assess the adaptability of the framework to unseen data and real-world scenarios. Furthermore, to address the scalability concerns associated with the high execution time of LRSO, we plan to investigate optimization strategies such as parallelizing the search process, incorporating early stopping criteria based on convergence thresholds, and performing dimensionality reduction prior to optimization. In addition, future work will include a comprehensive evaluation of resource utilization metrics, including CPU usage, memory consumption, and energy efficiency, alongside execution time. These strategies and evaluations are expected to reduce latency and computational overhead while maintaining accuracy, thereby improving the suitability of LRSO for real-time and resource-constrained IoST applications. We also plan to perform a comprehensive complexity analysis, including precise measurements of execution time, memory requirements, and model storage size. Furthermore, we will explore lightweight model compression, quantization, and pruning techniques to enable seamless deployment on tiny sensor nodes without compromising detection accuracy.

Funding Statement

This study was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2503).

Acknowledgement

The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for funding this work through (grant number IMSIU-DDRSP2503).

References

[1] Khan, M.A. and Javaid, N., 2022. Computationally efficient topology optimization of scale-free IoT networks. Computer Communications, 185, pp.1-12.

[2] Alshehri, H.S. and Bajaber, F., 2024. A Cluster-Based Data Aggregation in IoT Sensor Networks Using the Firefly Optimization Algorithm. Journal of Computer Networks and Communications, 2024(1), p.8349653.

[3] Yu, S., Wang, X., Shen, Y., Wu, G., Yu, S. and Shen, S., 2024. Novel intrusion detection strategies with optimal hyper parameters for industrial internet of things based on stochastic games and double deep Q-networks. IEEE Internet of Things Journal, 11(17), pp.29132-29145.

[4] Javaid, N., Sher, A., Nasir, H. and Guizani, N., 2018. Intelligence in IoT-based 5G networks: Opportunities and challenges. IEEE Communications Magazine, 56(10), pp.94-100.

[5] Shafin, S.S., Karmakar, G., Mareels, I., Balasubramanian, V. and Kolluri, R.R., 2024. Sensor self-declaration of numeric data reliability in internet of things. IEEE Transactions on Reliability, 74(2), pp.2751-2765.

[6] Singh, A., Amutha, J., Nagar, J. and Sharma, S., 2023. A deep learning approach to predict the number of k-barriers for intrusion detection over a circular region using wireless sensor networks. Expert Systems with Applications, 211, p.118588.

[7] Shen, S., Cai, C., Shen, Y., Wu, X., Ke, W. and Yu, S., 2025. Joint Mean-Field Game and Multiagent Asynchronous Advantage Actor-Critic for Edge Intelligence-Based IoT Malware Propagation Defense. IEEE Transactions on Dependable and Secure Computing.

[8] Hasnain, M., Javaid, N., Khan, F.A., Nasser, N. and Imran, M., 2025. An AI-driven strategy for threat detection in Internet of Sensors Things using machine learning, active learning, and optimization techniques, accepted in IEEE Global Communications Conference (GLOBECOM) 2025.

[9] Manivannan, D., 2024. Recent endeavors in machine learning-powered intrusion detection systems for the internet of things. Journal of Network and Computer Applications, p.103925.

[10] Liu, J., Gao, Y. and Hu, F., 2021. A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM. Computers and Security, 106, p.102289.

[11] Ding, H., Chen, L., Dong, L., Fu, Z. and Cui, X., 2022. Imbalanced data classification: A KNN and generative adversarial networks-based hybrid approach for intrusion detection. Future Generation Computer Systems, 131, pp.240-254.

[12] Ahmed, U., Lin, J.C.W. and Srivastava, G., 2022. A resource allocation deep active learning based on load balancer for network intrusion detection in SDN sensors. Computer Communications, 184, pp.56-63.

[13] Vaarandi, R. and Guerra-Manzanares, A., 2024. Network IDS alert classification with active learning techniques. Journal of Information Security and Applications, 81, p.103687.

[14] Alkhonaini, M.A., Alohali, M.A., Aljebreen, M., Eltahir, M.M., Alanazi, M.H., Yafoz, A., Alsini, R. and Khadidos, A.O., 2025. Sandpiper optimization with hybrid deep learning model for blockchain-assisted intrusion detection in iot environment. Alexandria Engineering Journal, 112, pp.49-62.

[15] Aljehane, N.O., Mengash, H.A., Hassine, S.B., Alotaibi, F.A., Salama, A.S. and Abdelbagi, S., 2024. Optimizing intrusion detection using intelligent feature selection with machine learning model. Alexandria Engineering Journal, 91, pp.39-49.

[16] Almomani, I., Al-Kasasbeh, B. and Al-Akhras, M., 2016. IoSTs-DS: a dataset for intrusion detection systems in wireless sensor networks. Journal of Sensors, 2016(1), p.4731953.

[17] https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset/code

[18] Almomani, I., Al-Kasasbeh, B. and Al-Akhras, M., 2016. IoSTs-DS: a dataset for intrusion detection systems in wireless sensor networks. Journal of Sensors, 2016(1), p.4731953.

[19] Khan, Z.A., Javaid, N., Saeed, A., Ahmed, I. and Khan, F.A., 2025. Towards IoT device privacy & data integrity through decentralized storage with blockchain and predicting malicious entities by stacked machine learning. Internet of Things, p.101642.

[20] Alferaidi, A., Yadav, K., Alharbi, Y., Alreshidi, E.J., Alreshidi, A., Aboshosha, B.W., Sharma, R., Alkhayyat, A. and Aray, D.G., 2024. A novel hybrid, BERT and deep learning model network intrusion detection system for healthcare electronics. IEEE Transactions on Consumer Electronics.

[21] Hossain, M.A. and Islam, M.S., 2023. Ensuring network security with a robust intrusion detection system using ensemble-based machine learning. Array, 19, p.100306.

[22] Khan, M.A., Iqbal, N., Jamil, H. and Kim, D.H., 2023. An optimized ensemble prediction model using AutoML based on soft voting classifier for network intrusion detection. Journal of Network and Computer Applications, 212, p.103560.

[23] Samantaray, M., Barik, R.C. and Biswal, A.K., 2024. A comparative assessment of machine learning algorithms in the IoT-based network intrusion detection systems. Decision Analytics Journal, 11, p.100478.

[24] Abdul Salam, M., Fouad, K.M., Elbably, D.L. and Elsayed, S.M., 2024. Federated learning model for credit card fraud detection with data balancing techniques. Neural Computing and Applications, 36(11), pp.6231-6256.

[25] Abdulganiyu, O.H., Tchakoucht, T.A., Saheed, Y.K. and Ahmed, H.A., 2025. XIDINTFL-VAE: XGBoost-based intrusion detection of imbalance network traffic via class-wise focal loss variational autoencoder. The Journal of Supercomputing, 81(1), pp.1-38.

[26] Zhu, W., Lévy-Leduc, C. and Ternès, N., 2025. Variable selection in high-dimensional logistic regression models using a whitening approach. IEEE Transactions on Computational Biology and Bioinformatics.

[27] Raghavendra, T., Anand, M., Selvi, M., Thangaramya, K., Kumar, S.S. and Kannan, A., 2022. An intelligent RPL attack detection using machine learning-based intrusion detection system for Internet of Things. Procedia Computer Science, 215, pp.61-70.

[28] Gaur, P., Chowdhury, A., McCreadie, K., Pachori, R.B. and Wang, H., 2021. Logistic regression with tangent space-based cross-subject learning for enhancing motor imagery classification. IEEE Transactions on Cognitive and Developmental Systems, 14(3), pp.1188-1197.

[29] Widodo, A.O., Setiawan, B. and Indraswari, R., 2024. Machine learning-based intrusion detection on multi-class imbalanced dataset using SMOTE. Procedia Computer Science, 234, pp.578-583.

[30] Zhu, J., Chen, X., Hu, Q., Xiao, Y., Wang, B., Sheng, B. and Chen, C.P., 2024. Clustering environment aware learning for active domain adaptation. IEEE Transactions on Systems, Man, and Cybernetics: Systems.

[31] Zhang, H., Ye, J., Huang, W., Liu, X. and Gu, J., 2024. Survey of federated learning in intrusion detection. Journal of Parallel and Distributed Computing, p.104976.

[32] Yu, S., Zhai, R., Shen, Y., Wu, G., Zhang, H., Yu, S. and Shen, S., 2023. Deep Q-network-based open-set intrusion detection solution for industrial Internet of Things. IEEE Internet of Things Journal, 11(7), pp.12536-12550.

[33] Lin, J., Liang, Z., Deng, S., Cai, L., Jiang, T., Li, T., Jia, K. and Xu, X., 2024. Exploring diversity-based active learning for 3d object detection in autonomous driving. IEEE Transactions on Intelligent Transportation Systems.

[34] Kumar, P. and Gupta, A., 2020. Active learning query strategies for classification, regression, and clustering: A survey. Journal of Computer Science and Technology, 35, pp.913-945.

[35] Huang, J., Farpour, N., Yang, B.J., Mupparapu, M., Lure, F., Li, J., Yan, H. and Setzer, F.C., 2024. Uncertainty-based active learning by bayesian U-Net for multi-label cone-beam CT segmentation. Journal of endodontics, 50(2), pp.220-228.

[36] Hu, R., 2011. Active learning for text classification.

[37] Nirmala, B.V. and Selvaraj, K., 2024. Malicious node detection in wireless sensor network using modified sandpiper optimization algorithm. Wireless Networks, pp.1-22.

[38] Rajan, R., Raju, P.V.R., SVN, S.K., Selvi, M. and Shanmugapriya, R., 2025. RL-BOT—Reinforcement learning based billfish optimization technique for secured data aggregation in internet of things. Peer-to-Peer Networking and Applications, 18(3), p.136.

[39] Khan, I.U., Javeid, N., Taylor, C.J., Gamage, K.A. and Ma, X., 2021. A stacked machine and deep learning-based approach for analysing electricity theft in smart grids. IEEE Transactions on Smart Grid, 13(2), pp.1633-1644.

[40] Dangore, M., Bhaturkar, D., Bhale, K.M., Jadhav, H.M., Borate, V.K. and Mali, Y.K., 2024, June. Applying random forest for IoT systems in industrial environments. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

[41] Hossain, M.A., Saif, S. and Islam, M.S., 2024, often. Interpretable Machine Learning for IoT Security: Feature Selection and Explainability in Botnet Intrusion Detection using Extra Trees Classifier. In 2024 1st International Conference on Innovative Engineering Sciences and Technological Research (ICIESTR) (pp. 1-6). IEEE.

[42] Bharot, N. and Breslin, J.G., 2025, June. A Transfer Learning-Enhanced TabNet-Based Intrusion Detection System for IoMT Security. In Intelligent Computing-Proceedings of the Computing Conference (pp. 520-539). Cham: Springer Nature Switzerland.

[43] Wang, M., Yang, N. and Weng, N., 2023. Securing a smart home with a transformer-based IoT intrusion detection system. Electronics, 12(9), p.2100.

[44] Gopi, A.P., Gowthami, M., Srujana, T., Gnana Padmini, S. and Durga Malleswari, M., 2022. Classification of denial-of-service attacks in IoT networks using AlexNet. In Human-Centric Smart Computing: Proceedings of ICHCSC 2022 (pp. 349-357). Singapore: Springer Nature Singapore.

[45] Shen, S., Cai, C., Li, Z., Shen, Y., Wu, G. and Yu, S., 2024. Deep Q-network-based heuristic intrusion detection against edge-based SIoT zero-day attacks. Applied Soft Computing, 150, p.111080.

[46] Tahsien, S.M., Karimipour, H. and Spachos, P., 2020. Machine learning based solutions for security of Internet of Things (IoT): A survey. Journal of Network and Computer Applications, 161, p.102630.

[47] Fares, H., Vibhute, A.D., Mouniane, Y. and Bouijij, H., 2025. Intrusion Detection in Wireless Sensor Networks using Machine Learning. Procedia Computer Science, 252, pp.912-921.

[48] Hassini, K., Khalis, S., Habibi, O., Chemmakha, M. and Lazaar, M., 2024. An end-to-end learning approach for enhancing intrusion detection in Industrial-Internet of Things. Knowledge-Based Systems, 294, p.111785.

[49] Malik, J., Akhunzada, A., Al-Shaoftenleh, A.S., Zeadally, S. and Almogren, A., 2025. Hybrid deep learning based threat intelligence framework for Industrial IoT systems. Journal of Industrial Information Integration, p.100846.

[50] Gupta, K., Sharma, D.K., Gupta, K.D. and Kumar, A., 2022. A tree classifier based network intrusion detection model for Internet of Medical Things. Computers and Electrical Engineering, 102, p.108158.

[51] Cohen, L., Avrahami-Bakish, G., Last, M., Kandel, A. and Kipersztok, O., 2008. Real-time data mining of non-stationary data streams from sensor networks. Information Fusion, 9(3), pp.344-353.

[52] Jafarian, T., Ghaffari, A., Seyfollahi, A. and Arasteh, B., 2025. Detecting and mitigating security anomalies in software-defined networking (SDN) using gradient-boosted trees and floodlight controller characteristics. Computer Standards and Interfaces, 91, p.103871.

[53] Karthika, R.A. and Maheswari, M., 2022. Detection analysis of malicious cyber attacks using machine learning algorithms. Materials Today: Proceedings, 68, pp.26-34.

[54] Prasad, P., Tahir, M. and Isoaho, J., 2025. Enhancing Explainability of Artificial Intelligence for Threat Detection in SDN-based Multicast Systems. Procedia Computer Science, 257, pp.569-574.

[55] Nkoro, E.C., Nwakanma, C.I., Lee, J.M. and Kim, D.S., 2024. Detecting cyberthreats in metaverse learning platforms using an explainable DNN. Internet of Things, 25, p.101046.

[56] Alotaibi, S.R., Alkahtani, H.K., Aljebreen, M., Alshuhail, A., Saeed, M.K., Ebad, S.A., Almukadi, W.S. and Alotaibi, M., 2025. Explainable artificial intelligence in web phishing classification on secure IoT with cloud-based cyber-physical systems. Alexandria Engineering Journal, 110, pp.490-505.

[57] Wan, X., Xue, G., Zhong, Y. and Wang, Z., 2025. Separating Prediction and Explanation: An Approach Based on Explainable Artificial Intelligence for Analyzing Network Intrusion. Journal of Network and Systems Management, 33(1), p.16.

[58] Hooshmand, M.K., Huchaiah, M.D., Alzighaibi, A.R., Hashim, H., Atlam, E.S. and Gad, I., 2024. Robust network anomaly detection using ensemble learning approach and explainable artificial intelligence (XAI). Alexandria Engineering Journal, 94, pp.120-130.

[59] Shahzadi, N., Javaid, N., Akbar, M., Aldegheishem, A., Alrajeh, N. and Bouk, S.H., 2024. A novel data driven approach for combating energy theft in urbanized smart grids using artificial intelligence. Expert Systems with Applications, 253, p.124182.

[60] Aljabri, W., Hamid, M.A. and Mosli, R., 2025. Enhancing real-time intrusion detection system for in-vehicle networks by employing novel feature engineering techniques and lightweight modeling. Ad Hoc Networks, 169, p.103737.

[61] Calzarossa, M.C., Giudici, P. and Zieni, R., 2025. An assessment framework for explainable AI with applications to cybersecurity. Artificial Intelligence Review, 58(5), p.150.

[62] Shen, Yizhou, Carlton Shepherd, Chuadhry Mujeeb Ahmed, Shigen Shen, and Shui Yu. "SGD3QN: Joint stochastic games and dueling double deep Q-networks for defending malware propagation in edge intelligence-enabled Internet of Things." IEEE Transactions on Information Forensics and Security 19 (2024): 6978-6990.