IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 59, No. 6, June, 2026, pp. 1563-1571
Journal homepage: http://iieta.org/journals/jesa

An Explainable TabNet Framework for Cyberattack Detection in IIoT Networks

Amine Brikci-Nigassa[1*], Hamed Benahmed[1], Akkacha Bekaddour[2], Amira N. E. H. Ghezzar[1], Abderrahmane Brikci-Nigassa[1], Mohammed M'hamedi[2], Mohammed A. Abderrahim[3]

[1] LRIT Laboratory, CS Department, Faculty of Sciences, University of Abou Bekr Belkaïd, Tlemcen 13000, Algeria
[2] École Supérieure en Sciences Appliquées de Tlemcen, Tlemcen 13000, Algeria
[3] LTALA Laboratory, CS Department, Faculty of Sciences, University of Abou Bekr Belkaïd, Tlemcen 13000, Algeria

Corresponding Author Email: amine.brikcinigassa@univ-tlemcen.dz

https://doi.org/10.18280/jesa.590605

Received: 13 April 2026
Revised: 20 May 2026
Accepted: 11 June 2026
Available online: 30 June 2026

Keywords:
cyberattacks, DataSense, explainable AI, feature engineering, Intrusion Detection System, Industrial Internet of Things, SHapely Additive exPlanations, TabNet

ABSTRACT

The rapid expansion of the Industrial Internet of Things (IIoT) field has brought forward many changes and, alongside them, significant cybersecurity challenges, particularly regarding the detection and classification of attacks within industrial network traffic. The existing intrusion detection approaches struggle with class imbalance and lack interpretability, effectively limiting their deployment in safety-critical industrial environments where transparency is primordial. This paper presents a comprehensive attack detection and classification approach that addresses said challenges through explainable Machine Learning (ML) techniques and the Synthetic Minority Oversampling Technique (SMOTE). We evaluate our approach on the DataSense IIoT dataset. Our method outmatches state-of-the-art methods by providing interpretability, which makes it a transparent solution for strengthening the security posture of IIoT networks against emerging cyber threats.

1. INTRODUCTION

With the advent of the "Industry 4.0" paradigm came a revolution in the industrial sector; what used to be disconnected machines working on their own or in small clusters has now become a huge network of interconnected sensors and Internet of Things (IoT) devices. This change has brought forward what is now commonly referred to as Industrial Internet of Things (IIoT), which is, at its core, the integration of physical industry assets with intelligent electronics for the sake of centralized control, information collection, inter-communication, and data analysis, thus providing a better grasp on the machines' state and efficiency.

This growth, however, does not come without its own set of issues, as exposing Industrial Control Systems (ICS) to a networked environment poses a big cybersecurity risk as these systems are especially susceptible to common IoT vulnerabilities [1]. An example of such a case would be the CrashOverride-based "Industroyer" malware attacks on Ukraine's electricity grid in 2016 [2] perfectly showcasing how cyberthreats can compromise an ICS by exploiting communication protocols [3] resulting in widespread blackouts and affecting thousands of people. These incidents emphasise the urgent need for resilient cybersecurity measures in IIoT systems to ensure operational continuity and safeguard critical industrial infrastructure even under an attack scenario.

To secure an IIoT network, the use of an Intrusion Detection System (IDS) is crucial, especially given the inherent vulnerabilities of IoT devices with limited computing power and storage capacity, which complicates the use of standard encryption techniques. Traditional IDS implementations employ two primary methods: signature-based detection, which compares observed activities against a database of known threat patterns, offering a higher accuracy for recognized threats but proving ineffective against zero-day exploits, and anomaly-based detection, which establishes a baseline for normal network behaviour and flags deviation as potential threats but may generate excessive false positives. Modern approaches have evolved to leverage Machine Learning (ML) to classify network traffic into specific attack categories. Deep Learning (DL) techniques enable more sophisticated pattern recognition than rule-based classification while also providing interpretable results that can identify the specific type of attack rather than simply flagging it as such. This multi-class classification approach combines the specificity of signature-based detection with the adaptability of ML, and has proven highly effective in detecting and classifying complex network traffic patterns [4-6].

Despite their success, current approaches face significant limitations due to outdated datasets that do not adequately represent modern IIoT cyberattacks [7]. The lack of realistic, comprehensive benchmarks that capture both network-level traffic and sensor-level telemetry across diverse attack types has slowed the development of ML-based detection systems. Furthermore, classification methods used to identify specific attack types often struggle with class imbalance, where the minority classes are underrepresented, this results in a real problem that affects model accuracy and leads to biased performance, DL models especially have difficulties learning to identify them which results in poor performance in these

areas [8, 9]. Existing systems perform poorly on 50-class attack classification due to the imbalanced state of the dataset and the lack of representation of rare attacks. The inherent complexity of real-world data makes feature extraction challenging, which in turn can bias classification models and hinder effective deployment.

In the context of an IDS, misclassifying malicious traffic as benign presents a far greater risk than blocking legitimate traffic, as it could cause significant damage or complete system failures.

To address the class imbalance problem, data sampling techniques, such as oversampling and under-sampling, have been proposed to balance training data before training [10]. The Synthetic Minority Over-Sampling Technique (SMOTE) [11] has emerged as a rather effective solution, generating synthetic minority examples by interpolating between existing samples in the feature space. However, the effectiveness of resampling techniques varies considerably depending on the degree of imbalance and the diversity of attack types [12].

Additionally, the interpretability of IDS remains a critical concern in industrial environments where transparency is essential for regulatory compliance and operational trust. Recent work has explored eXplainable Artificial Intelligence (XAI) techniques such as SHapely Additive exPlanations (SHAP) [13] to provide feature-level interpretability in intrusion detection systems, enabling security analysts to understand which network characteristics drive the classification decisions [14-17].

In this study, we address these challenges by proposing a comprehensive attack detection and classification pipeline, which we evaluated on the recently introduced DataSense IIoT dataset [7]. This dataset represents a significant advancement over previous benchmarks such as Edge-IIoTset [18] and WUSTL-EHMS [19], providing realistic data spanning 50 distinct attack types across 7 major categories: reconnaissance, Denial-of-Service (DoS), Distributed DoS (DDoS), web exploitation, Man-in-the-Middle (MitM), brute force, and malware. Our approach integrates domain-informed feature engineering, per-class interpretable feature selection via SHAP and conservative SMOTE-based oversampling to mitigate the somewhat extreme class imbalance while still preserving the natural distribution for the evaluation.

We employ TabNet [20], an attention DL architecture explicitly designed for tabular data, which provides both high classification Accuracy and built-in interpretability through feature importance mechanisms. Unlike black-box neural networks, TabNet uses sequential attention to select relevant features at each decision step, offering transparency in feature utilisation. Our methodology demonstrates that combining interpretable feature selection, targeted resampling and an interpretable classifier can effectively address the challenges of explainability and class imbalance in multi-class IIoT intrusion detection.

The main contributions of this work are as follows:

1) We demonstrate that per-class SHAP-based feature selection preserves discriminative power for minority attack classes while avoiding the majority-class bias inherent in global feature importance methods.
2) We show that using conservative SMOTE on minority classes strikes an effective balance between improving Recall on rare attacks and avoiding over-reliance on synthetic data.
3) We employ TabNet's built-in feature importance mechanisms to provide model interpretability without requiring post-hoc explanation methods.

2. RELATED WORKS

ML and DL techniques for attack detection and classification in IoT and IIoT networks have evolved significantly in recent years. Existing approaches range from computationally efficient models optimised for deployment on resource-limited devices to complex hybrid architectures that combine high classification performance with interpretable decision-making mechanisms essential for industrial application.

Lightweight and resource-efficient intrusion detection has been heavily prioritised for IIoT deployments due to stringent hardware and energy constraints. Mendonca et al. [21] proposed a sparse evolutionary training (SET) based prediction model for a lightweight IDS, evaluated on two IoT datasets, namely, DS2OS and CICIDS2017; the SET-based model achieved an average Accuracy of 99.89% with an average testing time of 2.29ms. Similarly, Peng et al. [22] proposed a framework combining Random Forest (RF), SMOTE and Bayesian-Optimised eXtreme Gradient Boosting (XGBoost) for resource constrained environments, achieving 99.99% Accuracy on the CIC-IDS2017 dataset, demonstrating the effectiveness of tree-based models combined with resampling for deployable IIoT IDS. Beyond algorithmic optimisation, modern edge computing architectures have incorporated model compression techniques such as structured pruning, integer quantisation [23], and knowledge distillation [24] alongside edge-cloud orchestration and federated learning paradigms [25], to support efficient on-device deployment.

DL–based intrusion detection for industrial IoT has been addressed using recurrent and hybrid sequential models. Soliman et al. [26] proposed a DL–based approach using Singular Value Decomposition (SVD) for feature selection combined with Gated Recurrent Unit (GRU) networks on the ToN_IoT dataset, achieving 99.99% Accuracy for binary classification and 99.98% for multi-class classification; the same study reported strong multi-class performance for Bidirectional Long-Short Term Memory (Bi-LSTM) and GRU, confirming the suitability of recurrent architectures for IIoT traffic. Explainability and ensemble methods have been combined to improve both performance and transparency. Shtayat et al. [14] presented an explainable ensemble DL approach for intrusion detection in the Industrial IoT, combining an ensemble of Convolutional Neural Network (CNN) models with an Extreme Learning Machine (ELM) baseline and employing SHAP and Local Interpretable Model-agnostic Explanations (LIME) for interpretability; on the ToN_IoT dataset the ensemble achieved 99.69% Accuracy for binary classification and 99.63% for multi-class classification, while the ELM reached 98.77% (binary) and 88.23% (multi-class), highlighting the benefit of ensemble and explainable design in IIoT IDS.

Class imbalance and minority-class detection in IIoT have been tackled with multi-stage and multi-layer DL. Multi-stage deep learning (MSDL) approaches using Deep Neural Networks (DNNs) have been evaluated on X-IIoTID and WUSTL-IIoT datasets [27], achieving average F1-scores of 99.78% and 99.71%, respectively and specifically improving detection for minority attack types such as Command Injection and Backdoor compared to standard sampling, which aligns with the need for robust performance across rare attack classes

in industrial settings. CNN-based multiclass attack classification in IoT and IIoT has been explored using the DNN-EdgeIIoT (Edge-IIoTset) dataset [28]: a CNN for spatial feature extraction from network traffic achieved 100% Accuracy and F1-score for binary classification, 99.12% Accuracy and 0.946 F1-score for 6-class classification, and 95.5% Accuracy for 15-class classification, outperforming traditional ML classifiers. Transformer-based architectures have also been applied to multi-class intrusion detection in IoT. Tseng et al. [29] evaluated DNN, CNN, and Transformer models on the CIC-IoT-2023 dataset, reporting accuracies of 99.2%, 99.0%, and 98.8% for binary classification and approximately 93% for 12-class classification, showing that both convolutional and transformer-based models can handle complex multi-class IoT threat taxonomies. However, standard Transformers remain constrained by quadratic attention complexity and high memory demands, limiting their direct deployment for fine-grained multiclass traffic classification. Recent work has addressed these bottlenecks through efficiency-oriented variants such as hybrid CNN–Transformer pipelines [30] and edge-optimised inference runtimes [31]. While these adaptations improve scalability for sequential workloads, they typically require extensive architectural tuning and still depend on post-hoc methods for interpretability, leaving a practical gap for structured, tabular-form network telemetry where computational efficiency and transparent decision-making remain critical.

Hybrid DL models have been used to improve robustness and generalisation across datasets and attack types. A robust IDS using an improved hybrid CNN-MultiLayer Perceptron (CNN-MLP) model [32] was evaluated on IoT-23 and NF-BoT-IoT-v2, achieving 99.94% binary and 99.99% multi-class Accuracy on IoT-23 and 99.96% binary and 98.02% multi-class on NF-BoT-IoT-v2 (excluding normal class), demonstrating the advantage of hybrid architectures for binary and multi-class settings. Ghosh et al. [33] proposed TACNet (Temporal Attention Convolutional Network), which integrates multi-scale CNNs, Long-Short Term Memory (LSTM) networks, and temporal attention mechanisms for intrusion detection in IoT and IIoT networks; evaluated on CICIDS 2018, N-BaIoT, CIC IoT-DIAD 2024, TabularIoTAttack-2024, and DNN-EdgeIIoT, it achieved accuracies ranging from 98.56% to 99.98%, with 99.98% on CICIDS 2018 and 99.97% on CIC IoT-DIAD 2024. Such combinations of convolutional feature extraction and recurrent temporal modelling remain a common choice for IIoT IDS, with recent variants reporting high multiclass Accuracy on Edge-IIoTset and related benchmarks.

A comparison of the existing studies is presented in Table 1. While they demonstrate strong performance on binary or moderate multi-class intrusion detection tasks, relatively few address the challenge of highly imbalanced, fine-grained multi-class classification with dozens of distinct attack categories. Our work addresses this gap by evaluating on the DataSense dataset, which comprises 50 attack types across seven major categories. We combine per-class SHAP feature selection to preserve discriminative features for minority classes, conservative SMOTE to balance training data without excessive synthetic oversampling, and TabNet to achieve 82.93% Accuracy and 89.09% top-3 Accuracy while maintaining transparency in feature utilisation and decision-making.

Table 1. Summary of related works in Industrial Internet of Things (IIoT) intrusion detection

Ref.	Authors	Year	Datasets	Models	Classification	Accuracy
[21]	Mendonca et al.	2022	DS2OS CICIDS2017	SET	Multiclass	99.89%
[22]	Peng et al.	2025	CIC-IDS2017 CIC-DDoS2019	XGBoost-assisted Random Forest	Multiclass	99.99% 99.85%
[26]	Soliman et al.	2023	ToN_IoT	SVD-GRU	Multiclass	99.98%
[14]	Shtayat et al.	2023	ToN_IoT	Ensemble Model	Multiclass	99.63%
[27]	Popoola et al.	2025	X-IIoTID WUSTL-IIoT	DNN	Multiclass	99.78% 99.71%
[28]	Seyedkolaei et al.	2025	DNN-EdgeIIoT	CNN	Binary 6 class 15 class	100% 99.12% 95.50%
[29]	Tseng et al.	2024	CIC-IoT-2023	Transformer	Multiclass	99.40%
[32]	Kamal & Mashaly	2025	IoT-23 NF-BoT-IoT-v2	CNN-MLP	Multiclass	99.99% 98.02%
[33]	Ghosh et al.	2025	CICIDS 2018 DNN-EdgeIIoT	CNN-LSTM + Attention	Multiclass	99.98% 98.56%
[34]	Akhi et al.	2026	DataSense	TCN	Multiclass	65.80%

Note: SET = Sparse Evolutionary Training; XGBoost = eXtreme Gradient Boosting; SVD = Singular Value Decomposition; GRU = Gated Recurrent Unit; DNN = Deep Neural Network; CNN = Convolutional Neural Network; MLP = MultiLayer Perceptron; LSTM = Long-Short Term Memory; TCN = Temporal Convolutional Network.

3. METHODOLOGY

This section details the end-to-end pipeline developed for multi-class attack detection on IIoT network traffic. Our approach combines domain-informed feature engineering, class-aware feature selection, and a deep tabular learning architecture to address the challenges posed by high-dimensional data and severe class imbalance.

We first describe the DataSense IIoT dataset and our feature extraction strategy, then present the SHAP-based per-class feature selection method, data partitioning and preprocessing procedures, conservative resampling strategy, and conclude it with the TabNet classifier training, as shown in Figure 1.

3.1 Dataset

We evaluated our proposed pipeline on the DataSense IIoT dataset introduced by Firouzi et al. [7] at the Canadian Institute for Cybersecurity (CIC). This dataset is a realistic sensor and network-centric benchmark made specifically for security research on real time IIoT network attack analysis. The testbed spans five logical layers (IoT/IIoT, network infrastructure,

edge, cloud, and attacker) using 40 heterogeneous devices, including industrial sensors, IP cameras, embedded IoT devices, and network components [7]. The dataset regroups sensor logs and network traffic traces for 50 different attacks distributed across 7 high-level categories: reconnaissance, DoS, DDoS, web exploitation, MitM, brute force and malware. On top of that, this dataset also includes benign operational traffic generated under normal industrial operation to provide a balanced representation of a real-world industrial network under both attack and non-attack conditions. Compared to previous IoT/IIoT benchmarks such as Edge-IIoTset [18] or WUSTL-EHMS [19], DataSense has a broader attack coverage, integrates both sensor and network-level views, and has a more realistic device diversity

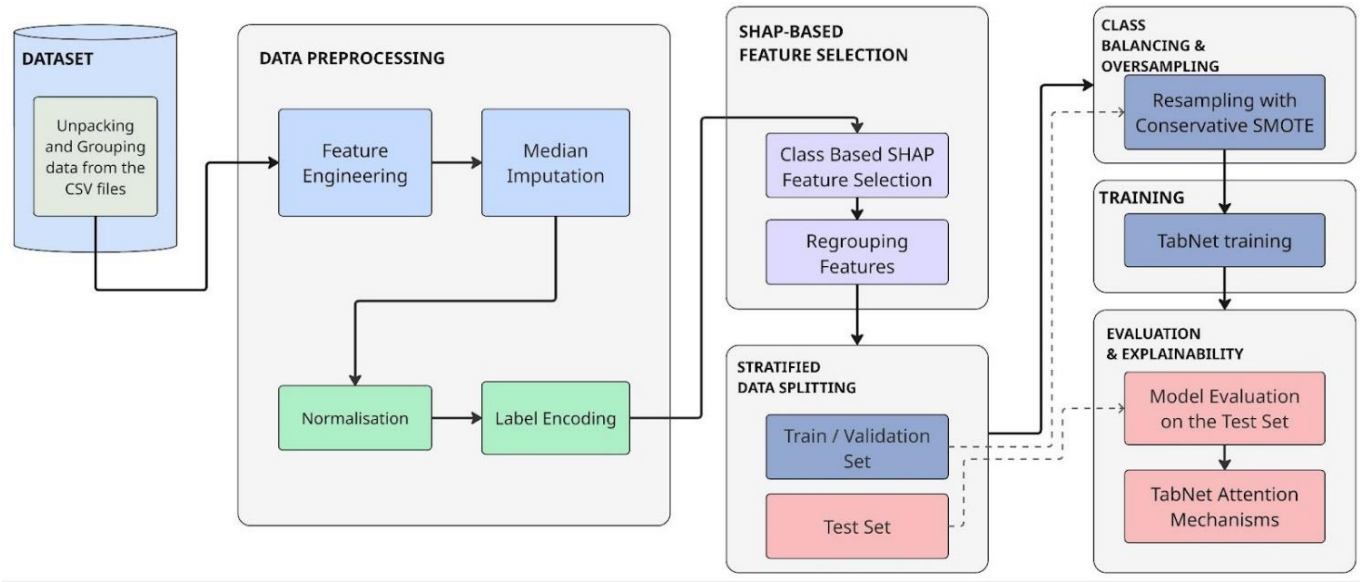

Figure 1. Proposed methodology pipeline
Note: CSV = Comma-Separated Values; SHAP = SHapeley Additive exPlanations; SMOTE = Synthetic Minority Oversampling Technique

3.2 Domain features and feature space

After doing our initial feature exclusion, we have 84 features, on top of which we extracted 19 hand-crafted domain-informed features (protocol diversity, flow asymmetry, flow ratio src/dst, etc.). These features capture protocol and port diversity (e.g., unique destination ports per flow, protocol counts) to distinguish multi-stage scans and lateral movement; flow asymmetry and directionality (e.g., packet and byte ratios src/dst, asymmetry scores) temporal burstiness and inter-arrival variability (e.g., burstiness indices, inter-arrival standard deviation), which help capture flooding and on–off attack patterns; and header and payload-level statistics such as min/mean/std of window size, Time To Live (TTL), IP length, and payload length. All ratio-based features were regularised by adding small constants to the denominator to avoid division by zero. After this augmentation, the full feature space is now made up of 103 dimensions in total. This addition was made with the aim of enriching the feature space before selectively filtering it down.

3.3 Feature selection

Operating directly in 103 dimensions can increase model complexity, training time and risk of overfitting, especially under strong class imbalance [7], so in order to compact down the feature subset we employed SHAP with an XGBoost surrogate model. The procedure we implemented is as follows: we pre-trained the surrogate model on the full 103 features for use as the SHAP surrogate model, then computed SHAP values using the TreeExplainer (for XGBoost) on a subset of the training set to keep the computation tractable. SHAP values were obtained per output class, yielding one importance array per class (each of shape samples × features). For each of the 50 classes, we took the mean absolute SHAP value per feature, ranked features by this importance, and kept only the top 20 features for that specific class. We then made the union of all the features that appeared in one or more classes' top-20, and ranked the features in this union by their average importance across classes' averages, resulting in 73 features that are listed alongside their average importance, of which we then selected the top-40 as shown in Figure 2.

3.4 Train/validation/test split

Because DataSense is provided as multiple CSV (Comma-Separated Values) files grouped by scenario, we first concatenated all attack and benign files, then formed a single labeled table. We then performed a stratified split to preserve class proportions and avoid discarding rare attacks: 80% of the data was allocated to a training set and 20% to the test set, and the 80% training set was further split into 70% train / 10% validation (overall 70/10/20 train/val/test), again using stratification.

3.5 Preprocessing

Preprocessing was fitted only on the training subset, and the fitted transformers were subsequently applied to validation and test sets to avoid information leakage: missing values in numerical features were imputed using median imputation, all continuous features were standardised via StandardScaler to stabilise gradient-based optimisation, and class labels were mapped to integer indices using a LabelEncoder fitted on the training labels.

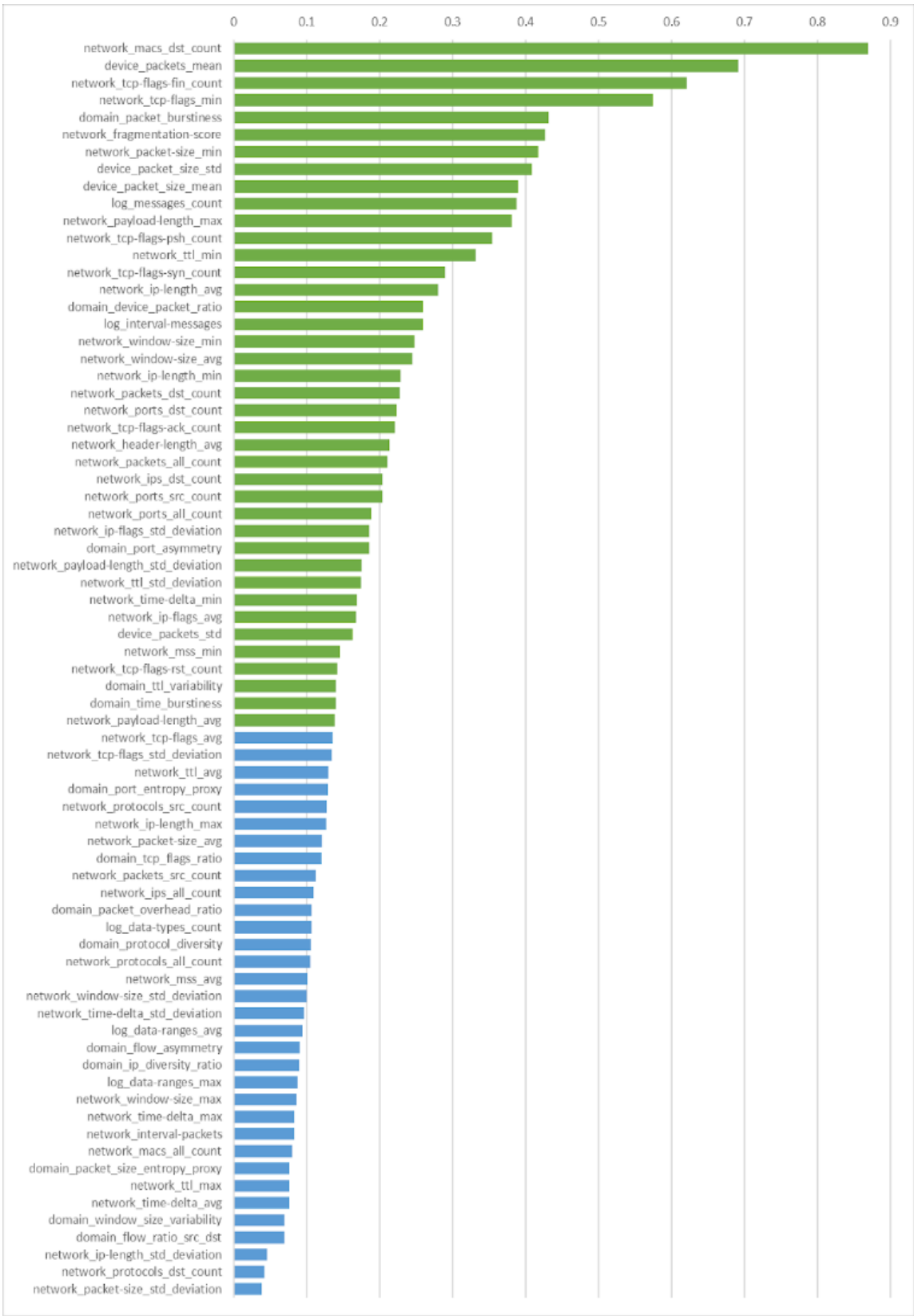

Figure 2. Global SHapley Additive exPlanations (SHAP) feature ranking after joining and aggregating the top-20 per class SHAP importance (in green: the top-40 features selected)

3.6 Resampling: Conservative Synthetic Minority Over-Sampling Technique

The DataSense dataset exhibits a class imbalance, with the largest class having orders of magnitude more samples than the smallest. Highly skewed label distributions are known to degrade classifier performance on minority classes and bias decision thresholds towards the majority [35]. To mitigate this, we applied SMOTE [11] to the training split only: classes with fewer than 500 instances were oversampled up to 500 samples, while validation and test sets were not resampled, preserving the original imbalanced distribution for evaluation. SMOTE generates artificial minority examples by interpolating between a point and its nearest neighbours in feature space,

which has been repeatedly shown to improve minority-class recall in IDS and fraud-detection tasks [11]. We chose a conservative target of 500 samples rather than full balancing to keep the total training set size manageable and avoid over-representing synthetic data.

3.7 Training

For our final classifier, we decided to use TabNet, which is a DL architecture explicitly designed for tabular data [20]. It uses a sequence of decision steps with sparse feature-selection masks from an attentive mechanism, allowing the model to focus on different feature subsets at different steps and to provide feature-level importance scores [20].

We then trained a TabNetClassifier using the Pytorch-TabNet implementation, initialised from the pretrained weights on the SMOTE-resampled, standardised training set with labels. The hyperparameters were empirically tuned on the validation set: decision and attention dimensions were set to n_d = 96 and n_a = 96, the number of decision steps to n_steps = 7, and sparsity and relaxation parameters to gamma = 1.5 and lambda_sparse = 1e-4. Optimisation was performed using Adam with a learning rate of 0.005 and a ReduceLROnPlateau scheduler (patience 10, factor 0.5, min_lr 1e-5). Batching used a mini-batch size of 256 and virtual batch size of 128 (ghost batch normalisation) [36], with a training budget of maximum 200 epochs and early stopping on validation Accuracy with patience 30. The loss function was cross-entropy for 50-class classification, and a random seed of 42 was used for reproducibility.

4. RESULTS AND DISCUSSION

This section presents the experimental results obtained from the evaluation of the proposed model, discusses its performance in comparison with some baseline models, and explains the importance scores of the features we selected.

We evaluated the model on the test set according to four metrics – Accuracy, Precision, Recall and F1-Score – defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\_score = \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

where,

TP: True Positives, *TN*: True Negatives, *FP:* False Positives, *FN*: False Negatives.

Using the 40-feature subset from SHAP-per-class selection and conservative SMOTE on the training set, we compared our model with a few relevant baseline models in Figures 3 and 4.

In these two figures, we can see that TabNet achieves 82.93% Accuracy and 81.18% F1-Score, and demonstrates consistent performance across both evaluation metrics and while the difference between KNN (K-Nearest Neighbours) and TabNet is relatively modest we decided to use TabNet because of its explainability through built-in feature importance mechanisms, as shown in Figure 5.

Figure 5 presents TabNet's sparse attention mechanisms through the aggregated feature importance score of all 40 of our selected features. The top-20 features account for 93.77% of the model's attention, with the most discriminative indicators being notably device packet statistics, network diversity metrics, and packet size characteristics.

To further contextualise our results, we compare the performance of our TabNet-based approach against other methods evaluated on the DataSense IIoT dataset. Table 2 presents a comparison with the baseline models reported by Firouzi et al. [7] in their original dataset introduction, as well as the Temporal Convolutional Network (TCN) approach proposed by Akhi et al. [34]. All methods were evaluated on the same 50-class classification task.

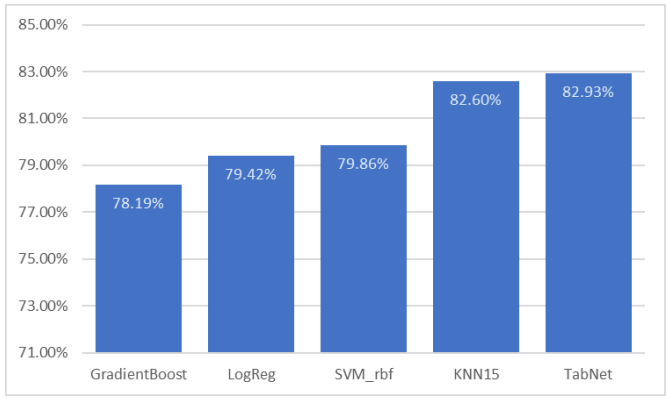

Figure 3. Accuracy of TabNet compared to other models
Note: LogReg = Logistic Regression; SVM_rbf = Support Vector Machine radial basis function; KNN = K-Nearest Neighbours

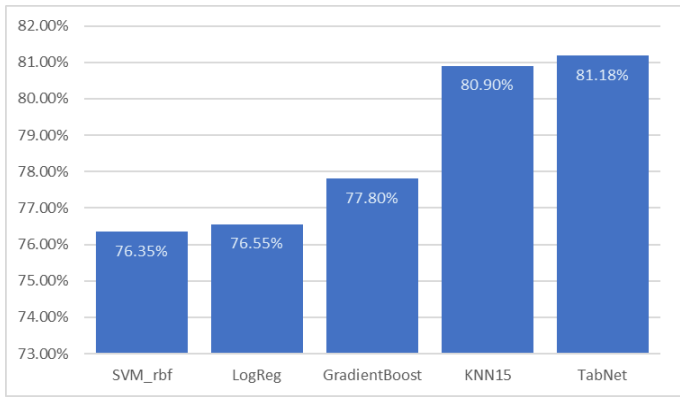

Figure 4. F1-score of TabNet compared to other models
Note: LogReg = Logistic Regression; SVM_rbf = Support Vector Machine radial basis function; KNN = K-Nearest Neighbours

Our TabNet-based approach achieves 82.93% Accuracy, exceeding the RF baseline at 82.69% and significantly outperforming the CNN baseline at 73.51%. This result is particularly noteworthy as it demonstrates that TabNet's attention-based architecture can compete with traditional ensemble methods on tabular data while providing interpretability advantages that RF lacks. In terms of Precision-Recall balance, our approach maintains 81.15% Precision and 82.93% Recall, closely aligning with the RF baseline and substantially outperforming CNN across all metrics. In an IIoT environment, high Recall is important

because missed attacks may lead to safety hazards or costly system failures; however, excessive false alarms can also disrupt operations by overwhelming security analysts, triggering unnecessary investigations, or interrupting legitimate industrial traffic.

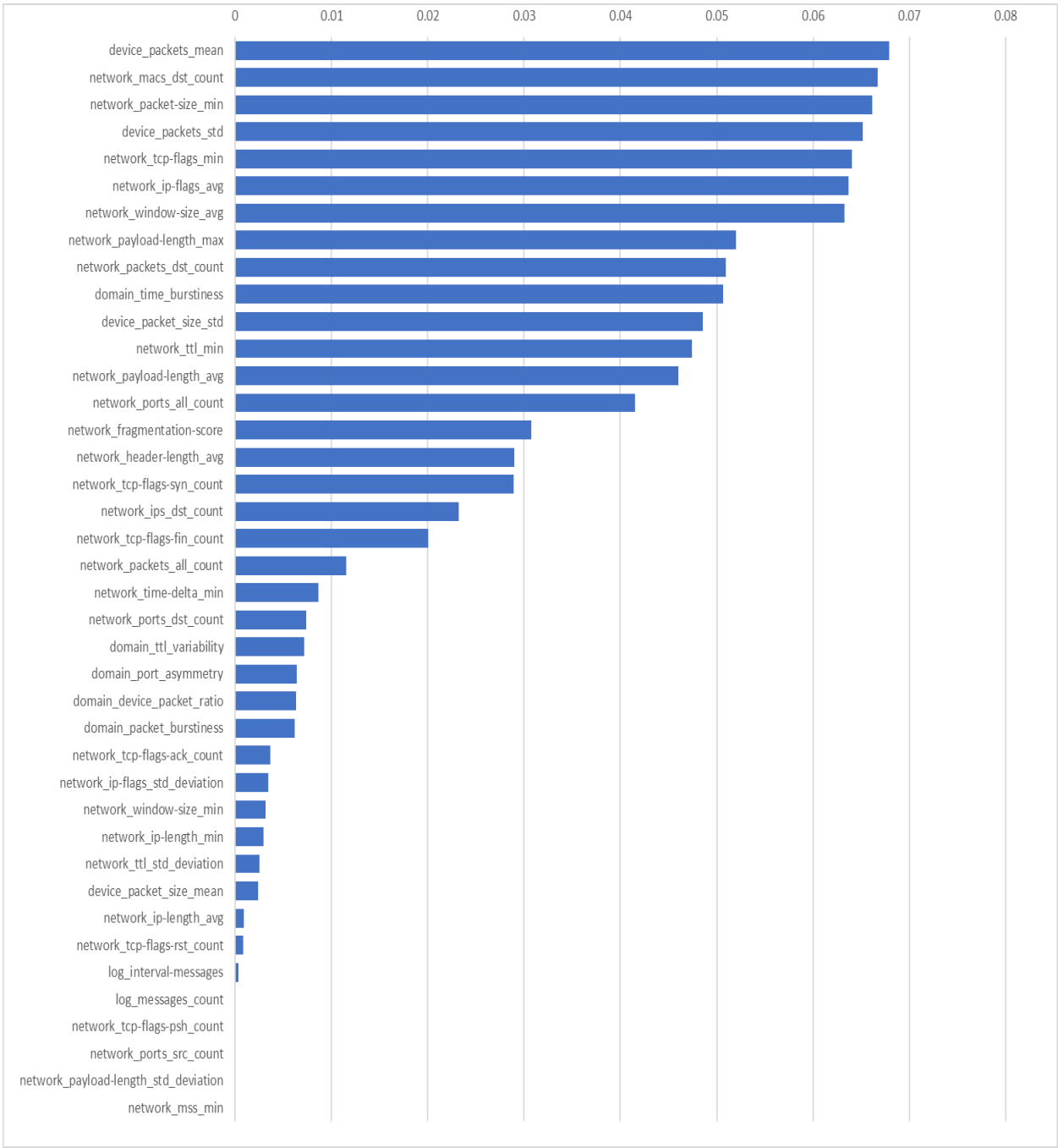

Figure 5. TabNet's aggregated feature importance ranking

Table 2. Comparative performance of different models on the DataSense IIoT dataset (on 50-class classification)

Author	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Firouzi et al. [7]	CNN	73.51	70.88	73.51	72.28
Firouzi et al. [7]	RF	82.69	81.83	82.69	82.21
Akhi et al. [34]	TCN	65.80	62.99	89.26	73.86
Our work	TabNet	82.93	81.15	82.93	81.18

Note: IIoT = Industrial Internet of Things; CNN = Convolutional Neural Network; RF = Random Forest; TCN = Temporal Convolutional Network

The TCN approach by Akhi et al. [34] presents an interesting contrast: while it achieves the highest Recall at 89.26%, this comes at the cost of significantly lower Precision (62.99%) and Accuracy (65.80%). This trade-off suggests the TCN model is overly aggressive in flagging traffic as malicious, resulting in numerous false positives that would be problematic in production IIoT environments where false alarms can disrupt legitimate industrial processes.

Our model of choice strikes a more balanced approach, maintaining competitive Accuracy and a healthy Precision-

Recall trade-off critical for real-world deployment. Additionally, we achieve 89.09% top-3 Accuracy, meaning that in nearly 9 out of 10 cases, the correct attack type appears among the model's top three predictions, which is a valuable characteristic for security operations where analysts investigate multiple potential classifications. Unlike RF and TCN, TabNet provides built-in interpretability through sequential attention mechanisms, allowing security analysts to understand which features drive classification decisions, a critical capability for safety-critical industrial environments.

5. CONCLUSION

This study presents a comprehensive attack detection and classification pipeline for IIoT networks, addressing the dual challenges of extreme class imbalance and model interpretability in fine-grained multi-class intrusion detection. Our approach integrates SHAP-based feature selection, reducing the augmented 103 domain-informed features to the 40 most relevant ones, with TabNet's sequential attention mechanisms and conservative SMOTE oversampling. It should be noted that SMOTE is applied exclusively during the training phase, where synthetic minority-class samples are generated to enhance model learning without introducing computational overhead during inference. Evaluated on the DataSense IIoT dataset, which encompasses 50 distinct attack types across seven major categories, the proposed pipeline achieves 82.93% overall Accuracy and 89.09% top-3 Accuracy. These results demonstrate that interpretable DL architectures can achieve competitive performance on complex multi-class intrusion detection tasks without sacrificing explainability for classification accuracy. From a deployment perspective, the primary computational burden stems from TabNet inference. Although inherently more resource-intensive than conventional tree-based models, the architecture remains viable for fog or edge-server environments rather than low-power sensor nodes. Consequently, deployment at an IIoT gateway level is recommended to balance computational demands with real-time operational constraints. Future work will focus on validating the pipeline in real-world industrial testbeds and conducting user studies with security analysts to evaluate operational utility. Furthermore, given the decentralized and sensitive nature of IIoT data, subsequent research should investigate edge-centric deployment strategies that prioritize data privacy, alongside model compression techniques such as pruning, quantization, and knowledge distillation to further optimize inference efficiency in resource-constrained settings.

REFERENCES

[1] Serror, M., Hack, S., Henze, M., Schuba, M., Wehrle, K. (2020). Challenges and opportunities in securing the industrial Internet of Things. IEEE Transactions on Industrial Informatics, 17(5): 2985-2996. https://doi.org/10.1109/TII.2020.3023507

[2] Cherepanov, A. (2017). Win32/Industroyer: A new threat for industrial control systems. White paper, ESET. https://web-assets.esetstatic.com/wls/2017/06/Win32_Industroyer.pdf.

[3] Sarjan, H., Ameli, A., Ghafouri, M. (2022). Cyber-security of industrial Internet of Things in electric power systems. IEEE Access, 10: 92390-92409. https://doi.org/10.1109/ACCESS.2022.3202914

[4] Ni, C., Li, S.C. (2024). Machine learning enabled industrial IoT security: Challenges, trends and solutions. Journal of Industrial Information Integration, 38: 100549. https://doi.org/10.1016/j.jii.2023.100549

[5] Chaabouni, N., Mosbah, M., Zemmari, A., Sauvignac, C., Faruki, P. (2019). Network intrusion detection for IoT security based on learning techniques. IEEE Communications Surveys & Tutorials, 21(3): 2671-2701. https://doi.org/10.1109/COMST.2019.2896380

[6] Popoola, S.I., Adebisi, B., Hammoudeh, M., Gui, G., Gacanin, H. (2020). Hybrid deep learning for botnet attack detection in the internet-of-things networks. IEEE Internet of Things Journal, 8(6): 4944-4956. https://doi.org/10.1109/JIOT.2020.3034156

[7] Firouzi, A., Dadkhah, S., Maret, S.A., Ghorbani, A.A. (2025). DataSense: A real-time sensor-based benchmark dataset for attack analysis in IIoT with multi-objective feature selection. Electronics, 14(20): 4095. https://doi.org/10.3390/electronics14204095

[8] Mirsadeghi, S.M.H., Bahsi, H., Vaarandi, R., Inoubli, W. (2023). Learning from few cyber-attacks: Addressing the class imbalance problem in machine learning-based intrusion detection in software-defined networking. IEEE Access, 11: 140428-140442. https://doi.org/10.1109/ACCESS.2023.3341755

[9] Johnson, J.M., Khoshgoftaar, T.M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1): 27. https://doi.org/10.1186/s40537-019-0192-5

[10] Abdelkhalek, A., Mashaly, M. (2023). Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning. The Journal of Supercomputing, 79(10): 10611-10644. https://doi.org/10.1007/s11227-023-05073-x

[11] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357. https://doi.org/10.1613/jair.953

[12] Eid, A.M., Soudan, B., Nassif, A.B., Injadat, M. (2024). Enhancing intrusion detection in IIoT: Optimized CNN model with multi-class SMOTE balancing. Neural Computing and Applications, 36(24): 14643-14659. https://doi.org/10.1007/s00521-024-09857-x

[13] Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[14] Shtayat, M.B.M., Hasan, M.K., Sulaiman, R., Islam, S., Khan, A.U.R. (2023). An explainable ensemble deep learning approach for intrusion detection in industrial Internet of Things. IEEE Access, 11: 115047-115061. https://doi.org/10.1109/ACCESS.2023.3323573

[15] Zhang, S., Liu, S. (2025). An interpretable IoT intrusion detection method based on a hybrid deep learning model. In Proceedings of the 4th International Conference on Computer, Artificial Intelligence and Control Engineering, pp. 204-212. https://doi.org/10.1145/3727648.3727684

[16] Ahakonye, L.A.C., Nwakanma, C.I., Lee, J.M., Kim, D.S. (2024). Machine learning explainability for intrusion detection in the industrial Internet of Things.

IEEE Internet of Things Magazine, 7(3): 68-74. https://doi.org/10.1109/IOTM.001.2300171

[17] Attique, D., Hao, W., Ping, W., Javeed, D., Kumar, P. (2024). Explainable and data-efficient deep learning for enhanced attack detection in IIoT ecosystem. IEEE Internet of Things Journal, 11(24): 38976-38986. https://doi.org/10.1109/JIOT.2024.3384374

[18] Ferrag, M.A., Friha, O., Hamouda, D., Maglaras, L., Janicke, H. (2022). Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. IEEE Access, 10: 40281-40306. https://doi.org/10.1109/ACCESS.2022.3165809

[19] Hady, A.A., Ghubaish, A., Salman, T., Unal, D., Jain, R. (2020). Intrusion detection system for healthcare systems using medical and network data: A comparison study. IEEE Access, 8: 106576-106584. https://doi.org/10.1109/ACCESS.2020.3000421

[20] Arik, S.Ö., Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8): 6679-6687. https://doi.org/10.1609/aaai.v35i8.16826

[21] Mendonca, R.V., Silva, J.C., Rosa, R.L., Saadi, M., Rodriguez, D.Z., Farouk, A. (2022). A lightweight intelligent intrusion detection system for industrial Internet of Things using deep learning algorithms. Expert Systems, 39(5): e12917. https://doi.org/10.1111/exsy.12917

[22] Peng, S., Han, Y., Li, R., Liu, L., Liu, J., Gu, Z. (2025). ROSE-BOX: A lightweight and efficient intrusion detection framework for resource-constrained IIoT environments. Applied Sciences, 15(12): 6448. https://doi.org/10.3390/app15126448

[23] Bella, K., Guezzaz, A., Benkirane, S., Azrour, M., Ennajar, S. (2024). Pruning and quantization of CNN-based intrusion detection systems for cyber-physical systems. In the International Conference on Artificial Intelligence and Smart Environment, pp. 348-354. https://doi.org/10.1007/978-3-031-88304-0_48

[24] Almabdy, S., Ullah, A. (2025). Optimising intrusion detection systems in cloud-edge continuum with knowledge distillation for privacy-preserving and efficient communication. In 2025 IEEE 9th International Conference on Fog and Edge Computing (ICFEC), Tromso, Norway, pp. 1-5. https://doi.org/10.1109/ICFEC65699.2025.00016

[25] Albogami, N.N. (2025). Intelligent deep federated learning model for enhancing security in Internet of Things enabled edge computing environment. Scientific Reports, 15(1): 4041. https://doi.org/10.1038/s41598-025-88163-5

[26] Soliman, S., Oudah, W., Aljuhani, A. (2023). Deep learning-based intrusion detection approach for securing industrial Internet of Things. Alexandria Engineering Journal, 81: 371-383. https://doi.org/10.1016/j.aej.2023.09.023

[27] Popoola, S.I., Tsado, Y., Ogunjinmi, A.A., Sanchez-Velazquez, E., Peng, Y., Rawat, D.B. (2025). Multi-stage deep learning for intrusion detection in industrial Internet of Things. IEEE Access, 13: 60532-60555. https://doi.org/10.1109/ACCESS.2025.3557959

[28] Seyedkolaei, A.A., Mahmoudi, F., García, J. (2025). A deep learning approach for multiclass attack classification in IoT and IIoT networks using convolutional neural networks. Future Internet, 17(6): 230. https://doi.org/10.3390/fi17060230

[29] Tseng, S.M., Wang, Y.Q., Wang, Y.C. (2024). Multi-class intrusion detection based on transformer for IoT networks using CIC-IoT-2023 dataset. Future Internet, 16(8): 284. https://doi.org/10.3390/fi16080284

[30] Vijayakumar, B., Yogini, S., Lucia, R.J., Neha, B. (2026). A survey on a hybrid CNN+ transformer with genetic algorithm for intrusion detection in wireless IoT devices. In International Conference on Artificial Intelligence and Secure Data Analytics (ICAISDA 2025), pp. 223-237. https://doi.org/10.2991/978-94-6239-616-6_19

[31] Umar, H.G.A., Yasmeen, I., Aoun, M., Mazhar, T., Khan, M.A., Jaghdam, I.H., Hamam, H. (2025). Energy-efficient deep learning-based intrusion detection system for edge computing: A novel DNN-KDQ model. Journal of Cloud Computing, 14(1): 32. https://doi.org/10.1186/s13677-025-00762-9

[32] Kamal, H., Mashaly, M. (2025). Robust intrusion detection system using an improved hybrid deep learning model for binary and multi-class classification in IoT networks. Technologies, 13(3): 102. https://doi.org/10.3390/technologies13030102

[33] Ghosh, K.P., Hasan, M., Robin, M.T.I., Hossain, M.A., Islam, M.S. (2025). A novel deep learning framework with temporal attention convolutional networks for intrusion detection in IoT and IIoT networks. Scientific Reports, 15(1): 44624. https://doi.org/10.1038/s41598-025-32697-1

[34] Akhi, M., Eising, C., Dhirani, L.L. (2026). Securing IoT using lightweight TCN for edge deployment on raspberry Pi 4. IEEE Open Journal of the Communications Society, 7: 442-460. https://doi.org/10.1109/OJCOMS.2025.3649498

[35] Lemaître, G., Nogueira, F., Aridas, C.K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17): 1-5.

[36] Hoffer, E., Hubara, I., Soudry, D. (2017). Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. arXiv preprint arXiv:1705.08741. https://doi.org/10.48550/arXiv.1705.08741