GAS Journal of Engineering and Technology (GASJET)

Volume 3 | Issue 2, 2026 Homepage: https://gaspublishers.com/gasjet-home/ ISSN: 3048-5800

# Explainable AI for Cybersecurity Detection on Small and Noisy Datasets: A Comparative Study

**Adeniji Samuel Olamilekan[1], Adeyemo Latifat Abiodun[2], Olusa Livingstone Temidire[3], Olusa Cornerstone Temidara[4]**

[1]Department of Computer Science, Crown Polytechnic, Ado Ekiti, Nigeria.
[2]Department of Computer science, Osun State University, Osogbo, Nigeria.
[3]Department of Software Engineering, Federal University of Technology, Akure, Nigeria.
[4]Department of Computer Science, Achievers University, Owo, Nigeria

***Received:*** *20.01.2026* | ***Accepted:*** *10.02.2026* | ***Published:*** *14.02.2026*

***Corresponding Author:** Adeniji Samuel Olamilekan

**DOI:** *10.5281/zenodo.18639261*

**Abstract** **Review Article**

Explainable AI (XAI) is increasingly required for intrusion detection systems (IDS) because security analysts must justify alerts, prioritize response, and audit model behavior. In operational environments, however, supervised IDS commonly faces two constraints: limited labeled training data and imperfect supervision arising from delayed ground truth and weak labeling pipelines. This study presents a comparative evaluation of explainable multi-class intrusion detection under controlled small-data and noisy-label regimes using UNSW-NB15 and CICIDS2017. We simulate data scarcity by stratified downsampling of the training set and simulate label noise using both symmetric corruption and a security-realistic benignification mechanism that preferentially flips attack labels toward benign. Representative detector families are trained using empirical risk minimization and noise-mitigation strategies, and explanations are generated using SHAP, LIME, and Integrated Gradients. The evaluation jointly considers detection effectiveness, probability reliability, and explanation quality using Macro-F1, AUROC, AUPRC, Expected Calibration Error, and explanation metrics that capture faithfulness, stability, sparsity, and drift. Results show that performance and explanation reliability degrade nonlinearly when small data and noisy labels co-occur, with benignification noise causing the most severe losses. Noise-tolerant learning reduces these losses and improves calibration and explanation stability, indicating that training choices affect not only accuracy but also the reliability of analyst-facing explanations under scarce and noisy supervision.

**Keywords:** Intrusion detection, explainable AI, noisy labels, small datasets, robustness, calibration, SHAP, LIME, Integrated Gradients.



## 1. Introduction

Explainable AI for cybersecurity detection is increasingly important because intrusion detection systems (IDS) support high-stakes decisions in security operations, where analysts must justify alerts, prioritize response, and audit model behavior. At the same time, the data conditions under which IDS models are trained are often unfavorable: labeled cybersecurity datasets are frequently small due to the cost of expert labeling and the rapid

emergence of new attacks, and they are commonly noisy because ground truth is delayed, incomplete, or inferred from weak heuristics [12]–[14], [17], [18]. These constraints are particularly acute in multi-class intrusion detection, where minority attack categories are rare but operationally critical. As a result, a model that performs well under clean and abundant labels can degrade substantially when supervision is scarce or corrupted, producing unstable predictions and unreliable confidence estimates. While explainers such as LIME, SHAP, and Integrated Gradients can provide feature-level attributions to support analyst interpretation, explanations are only useful if they remain faithful and stable under realistic training stress [1]–[3]. Under small and noisy datasets, models may memorize spurious patterns or shift decision boundaries across retraining runs, leading to explanations that drift, vary in feature ranking, or highlight non-causal correlates. In parallel, noisy-label learning research has proposed noise-handling strategies, including sample-selection methods such as co-teaching and noise-tolerant loss functions such as Symmetric Cross Entropy and Generalized Cross Entropy, to reduce sensitivity to corrupted supervision [6]–[8]. However, comparatively little work evaluates, in a unified manner, how these strategies affect not only detection performance but also probability calibration and the consistency of explanations, which are essential for operational trust in cybersecurity settings.

This paper presents a comparative study of explainable cybersecurity detection under controlled small-data and noisy-label regimes using two widely adopted IDS benchmarks, UNSW-NB15 and CICIDS2017 [4], [5]. Training scarcity is simulated by stratified downsampling of the training set, and label noise is simulated using both symmetric corruption and a security-realistic benignification mechanism that preferentially flips attack labels toward benign. Representative detector families are trained using empirical risk minimization and noise-mitigation alternatives, and explanations are generated using LIME, SHAP, and Integrated Gradients. The study evaluates detection effectiveness and probability reliability alongside explanation faithfulness and stability, thereby linking training choices to the usefulness and consistency of analyst-facing explanations in intrusion detection.

## 2. Related Work

### 2.1 Explainable AI in Intrusion Detection and Cybersecurity

The demand for interpretability in security analytics has grown alongside the adoption of complex machine-learning models for intrusion detection, malware analysis, and anomaly detection, where decisions must be explainable to analysts and auditable for governance. Foundational interpretability work emphasizes that "interpretability" is context-dependent and must be evaluated with respect to concrete goals such as trust, debugging, and decision support rather than treated as a single universal property [10]. Broader surveys further organize explanation methods by explanation target, model access, and explanation form, highlighting recurring trade-offs between fidelity, human comprehensibility, and computational cost [11]. In cybersecurity specifically, recent systematic reviews emphasize that XAI can improve analyst trust and operational adoption of IDS by making decisions more transparent, but also note that many studies remain limited to clean benchmark settings and report explanations without rigorous robustness evaluation [12]. Similarly, recent IDS-focused XAI papers and reviews call attention to the need for evaluating explanation consistency, stability, and forensic usefulness when models are used as evidence-supporting tools in security workflows [13], [15]. Recent comparative studies applying SHAP and LIME to intrusion detection models also show that explanation quality varies substantially by model family and data regime, reinforcing that the choice of explainer cannot be separated from the characteristics of the trained detector [15]. Related work in IoT intrusion detection further motivates XAI as a means to increase transparency and user trust, especially where IDS decisions are integrated into automated response pipelines [16]. Despite these advances, much of the XAI-for-IDS literature evaluates explanations in isolation or focuses on qualitative interpretability demonstrations without stress testing under realistic supervision constraints.

In operational contexts, labels are often derived from weak signals and delayed incident confirmation, which can lead to systematic mislabeling patterns; explanations produced under such conditions may be plausible but unreliable, particularly across retraining cycles. This motivates the need for studies that treat explanation robustness as a primary objective rather than an auxiliary visualization.

### 2.2 Learning with Noisy Labels and Robust Training Strategies

A large body of work studies learning under label noise, proposing strategies that modify the objective function, reweight or filter examples, or model the noise process explicitly. A comprehensive survey of noisy-label learning categorizes major approaches and emphasizes that deep models can overfit noisy labels, motivating robust losses and sample-selection methods that limit memorization of corrupted supervision [17]. In parallel, the security literature highlights that attacks and failures can occur at multiple stages of the ML pipeline, including training-time data and label manipulation, making robustness to corrupted supervision relevant not only as a statistical issue but also as a security concern [18]. Robust loss functions seek to reduce sensitivity to mislabels while retaining sufficient learning capacity. Symmetric Cross Entropy introduces a combination of standard cross-entropy and a reverse term to improve robustness under noisy supervision [7]. Generalized Cross Entropy interpolates between mean absolute error and cross-entropy to improve noise tolerance while maintaining optimization stability [8]. Sample-selection approaches such as co-teaching train two models jointly and exchange small-loss instances under the premise that small-loss samples are more likely to be correctly labeled, improving robustness under severe label noise [6]. These strategies have been widely studied in computer vision and general classification tasks; however, in cybersecurity settings their evaluation often prioritizes detection performance, while calibration and explanation quality are less commonly assessed together. This gap is important because security analysts rely on calibrated confidence scores for alert prioritization and rely on explanations for triage and accountability.

### 2.3 Calibration, Reliability, and the Trustworthiness of IDS Outputs

Model calibration has become a central consideration in deploying probabilistic classifiers, especially in risk-sensitive domains where confidence scores influence downstream decisions. “On Calibration of Modern Neural Networks” shows that many modern neural networks are poorly calibrated and introduces temperature scaling as a simple post-hoc correction, while popularizing reliability diagrams and Expected Calibration Error as practical evaluation tools [9]. In cybersecurity operations, calibration is operationally meaningful because confidence is often used to rank alerts or set thresholds; miscalibration can increase analyst burden by elevating false positives or suppressing true attacks. Nevertheless, calibration is still frequently omitted in IDS evaluations, and the interaction between label noise, small data, robust training, and calibration remains underexplored in IDS-focused studies compared to accuracy-focused reporting.

### 2.4 Robustness of Explanations: Faithfulness, Stability, and Drift

Beyond producing explanations, recent interpretability discussions emphasize the need to evaluate explanation quality with measurable criteria. Foundational work argues for rigorous evaluation protocols and calls attention to the absence of consensus on what interpretability should mean and how it should be measured in practice [10]. Broader surveys of explanation methods similarly highlight the necessity of matching explanation techniques and evaluation metrics to the application setting, rather than assuming a single explainer is universally reliable [11]. In IDS contexts, recent reviews explicitly identify explanation stability and consistency as unresolved challenges, noting that explanations can change across retraining or under distribution shift, which is problematic for operational trust and auditing [12], [13]. Empirical IDS studies comparing post-hoc explainers also report that attribution rankings can vary substantially across explainers and models, motivating robustness-oriented metrics such as stability and

similarity-based measures when explanations are used for investigation or forensic justification [15].

### 2.5 Summary and Gap Addressed by This Paper

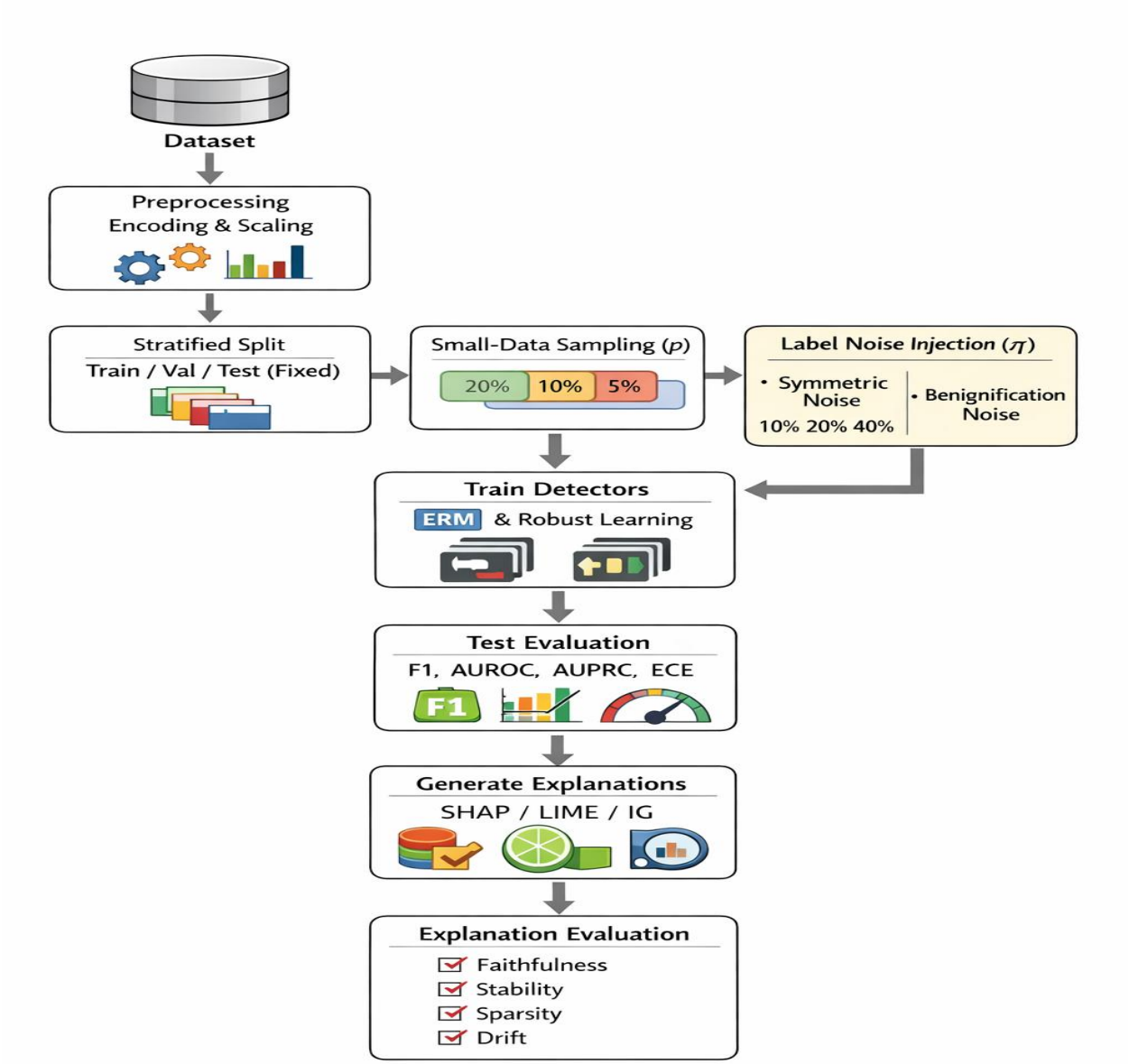
Prior work establishes the importance of XAI for IDS and provides many methods for learning with noisy labels. However, three limitations commonly remain. First, many IDS-XAI studies evaluate explainers under clean benchmark assumptions, without systematically stress testing explanation behavior under small-data and noisy-label regimes that mirror operational constraints [12], [13]. Second, noisy-label robustness work is rarely connected to explanation robustness, even though robust training can alter decision drivers and therefore the explanations that analysts see [17]. Third, IDS evaluations often underreport calibration, despite its operational relevance for alert ranking and triage and its known sensitivity to training conditions [9]. This paper addresses these gaps by jointly evaluating detection effectiveness, calibration reliability, and explanation robustness under controlled small-data and noisy-label regimes, enabling a comparative analysis of model family, robust training strategy, and explainer choice.

## 3. Methodology

### 3.1 Study Design

This work adopts an experimental comparative design to evaluate explainable AI for multi-class intrusion detection under two practical constraints that commonly arise in operational security analytics, namely limited labeled training data and noisy supervision. The study compares detection models trained under standard empirical risk minimization and under noise-robust learning strategies. Explanations are then generated using multiple XAI techniques and evaluated with quantitative metrics that capture faithfulness and robustness. The key experimental factors are training-set size and label-noise level, while the principal outcomes are detection effectiveness, probability calibration, and explanation quality. The end-to-end workflow of data preparation, stress testing, training, evaluation, and explanation is summarized in Figure 1.

*Figure 1. Overview of the experimental workflow for explainable cybersecurity detection under small-data and noisy-label regimes.*




### 3.2 Datasets and Multi-class Labeling

The evaluation uses two publicly available intrusion detection datasets, UNSW-NB15 and CICIDS2017, to reduce the risk that conclusions are dependent on a single benchmark. Both datasets are treated in a multi-class setting. For UNSW-NB15, labels include normal traffic and multiple attack categories. For CICIDS2017, labels include benign traffic and multiple attack types derived from real network traffic captures. Because the label spaces differ across datasets, all comparisons are performed within each dataset, and cross-dataset metrics are

interpreted as trend consistency rather than direct class-to-class equivalence. The dataset characteristics and experimental regimes used in this study are summarized in Table 1.

*Table 1: Dataset characteristics and experimental regimes.*

| Dataset | Task | #Samples | #Features | #Classes | Split policy | p fractions | η rates | Noise types |
|---|---|---|---|---|---|---|---|---|
| UNSW-NB15 | Multi-class | 257,673 (Train 175,341; Test 82,332) | 49 + label | 10 | Use official test; split train into train/val | {1.0,0.2,0.1,0.05} | {0.0,0.1,0.2,0.4} | Symmetric; Benignification |
| CICIDS2017 | Multi-class | 2,830,743 flows | 78 + label | 15 | Create stratified train/val/test; fixed test | {1.0,0.2,0.1,0.05} | {0.0,0.1,0.2,0.4} | Symmetric; Benignification |

Summary of dataset characteristics and experimental regimes for UNSW-NB15 and CICIDS2017.

### 3.3 Feature Representation

All experiments assume a tabular feature representation. Each instance corresponds to a network flow or session described by numeric and categorical traffic attributes. Features typically include duration statistics, packet and byte counts, direction-specific aggregates, rate-based measures, and protocol or flag indicators. For UNSW-NB15, the feature schema is defined by the dataset release and includes both continuous and categorical variables. For CICIDS2017, the experiments use the Machine Learning CSV flow representation produced by the dataset creators, which contains derived flow statistics suitable for supervised intrusion detection. Feature sets are not forced to match between datasets, and models are trained and explained using the native feature space of each dataset.

### 3.4 Data Preparation

A consistent preprocessing pipeline is applied independently to each dataset and then reused across all experimental regimes to ensure comparability. Missing values are handled using imputation statistics computed from the training split only, preventing information leakage into validation or testing. Categorical attributes are transformed into numeric representations using one-hot encoding or an equivalent encoding scheme that preserves category identity. Continuous features used by neural models are standardized using z-score normalization, where the mean and standard deviation are computed on the training split only and then applied to validation and test splits. Duplicate rows are removed where present to reduce bias from repeated samples. All preprocessing decisions are logged to support reproducibility.

### 3.5 Data Partitioning and Fixed Test Policy

The evaluation follows a fixed-test protocol to ensure that performance and explanation differences arise from the experimental stressors rather than changes in evaluation data. For UNSW-NB15, the dataset is distributed with predefined training and test sets; the provided test set is kept fixed, and the provided training set is further divided into training and validation partitions using stratified sampling. For CICIDS2017, the dataset is distributed as multiple flow CSV files rather than an official train/test split; therefore, a stratified partition is constructed into

training, validation, and test sets, and the test set is then held fixed across all experiments. Stratification is applied to preserve class proportions under imbalance, which is critical for multi-class intrusion detection where rare attack categories can otherwise be under-represented.

### 3.6 Small-Data Regimes

To simulate limited labeled data, the training partition is downsampled to a retained fraction p, while the validation and test partitions remain unchanged. Training fractions are chosen to represent both moderate and severe scarcity. In this study, $p \in \{1.0, 0.2, 0.1, 0.05\}$. Downsampling is performed using stratified sampling to preserve class proportions under imbalance, ensuring that minority attack classes remain represented as far as possible. Because some attack categories are extremely rare, the smallest fractions may still eliminate certain classes in a given draw; this outcome is recorded and reflected in the variability reported across random seeds. This regime quantifies how detection effectiveness, calibration, and explanation behavior change as the amount of labeled evidence decreases, and provides a baseline for analyzing interactions with label-noise conditions in subsequent experiments.

### 3.7 Noisy-Label Regimes

To model imperfect supervision, label noise is injected into training labels only after downsampling has been applied. Let $\eta$ denote the label-noise rate, with $\eta \in \{0.0, 0.1, 0.2, 0.4\}$. Training proceeds using corrupted labels $\tilde{y}$ produced from original labels $y$ under a controlled corruption process. Two noise mechanisms are evaluated. Symmetric noise replaces a label with an incorrect label sampled uniformly from the remaining classes with probability $\eta$. Asymmetric benignification noise introduces a stronger bias toward relabeling attack samples as benign or normal, reflecting a common operational failure mode where malicious traffic is mislabeled as legitimate due to incomplete ground truth, delayed incident confirmation, or heuristic labeling. The use of both noise mechanisms allows the study to test robustness under generic corruption and under a security-realistic corruption pattern.

### 3.8 Detection Models

Two representative detector families are evaluated to reflect common practice in intrusion detection and to support both model-agnostic and model-specific explanation methods. The first family is tree-based ensembles, instantiated as Random Forest and gradient-boosted decision trees, which are strong baselines for tabular intrusion features and typically perform well under class imbalance. The second family is a neural tabular classifier implemented as a multi-layer perceptron with regularization such as dropout. Hyperparameters are selected using the validation set under the clean, full-data condition $(p = 1.0, \eta = 0)$ and then kept fixed across all small-data and noisy-label regimes to avoid confounding the study by retuning models for each condition. Training uses validation-based early stopping to reduce overfitting, particularly under small-data conditions.

*Table 2: Detection models and learning strategies evaluated.*

| Category | Method | Brief description | Implementation detail |
|---|---|---|---|
| Detector (tree) | Gradient-Boosted Trees | Tree-ensemble detector for tabular IDS features | Tune on clean full-data; fix thereafter; early stopping |
| Detector (tree) | Random Forest | Bagging-based tree ensemble baseline | Fixed number of trees/depth constraints |

| | | | |
|---|---|---|---|
| Detector | MLP | Neural detector for tabular IDS features | Fixed architecture; standardized inputs; early stopping |
| Training strategy | ERM | Standard supervised baseline | Cross-entropy; optional class weights |
| Training strategy | Label smoothing | Reduces overconfidence and memorization | Fixed α across datasets/regimes |
| Training strategy | Noise-robust loss | Tolerance to mislabeled samples | Fixed loss parameters |
| Training strategy | Co-teaching / small-loss | Sample selection to reduce noisy impact | Warm-up then small-loss schedule |

*Models and learning strategies compared under small-data and noisy-label regimes.*

### 3.9 Noise-Robust Learning Strategies

Each detector family is trained using standard empirical risk minimization as well as noise-robust alternatives. The empirical risk minimization baseline minimizes cross-entropy loss using the potentially corrupted labels and serves as the reference condition. Label smoothing is applied as a regularization strategy by replacing hard one-hot targets with softened target distributions, reducing overconfidence and discouraging memorization of noisy labels. A noise-robust loss function is included, such as Symmetric Cross Entropy or Generalized Cross Entropy, to reduce sensitivity to mislabeling by balancing cross-entropy behavior with loss components that are less affected by corrupted samples. A sample-selection strategy is also used to reduce the influence of likely noisy examples. This is implemented as co-teaching or small-loss selection, where training emphasizes samples that yield smaller losses under the assumption that they are more likely to be correctly labeled, and in co-teaching two networks exchange selected samples to reduce confirmation bias. All strategies share the same optimization settings, validation monitoring, and early stopping criteria to ensure fair comparison.

### 3.10 Explainability Methods

Three explainers are used to generate feature-level attributions for alert interpretation and to support a comparative evaluation of explanation robustness. SHAP is applied to compute additive attributions for tabular models, providing local explanations for individual predictions and enabling global importance summaries by aggregation over many instances. LIME is used as a model-agnostic local explainer, generating explanations through perturbation sampling around an instance and fitting a weighted interpretable surrogate model. Integrated Gradients is used for the neural detector to compute attributions by integrating gradients along a straight-line path from a baseline input $x_0$ to the instance $x$. The attribution for feature iii is computed as

$$IG_i(x_i - x_0) = \int_0^1 \frac{\partial \mathrm{f}(\mathrm{x}_0 + \beta(\mathrm{x} - \mathrm{x}_0))}{\partial \mathrm{x}_\mathrm{i}} d\beta$$

Baselines for Integrated Gradients are defined using training-distribution statistics, such as feature-wise means, and the number of integration steps is held constant across regimes. Explanations are generated

on a fixed subset of test instances containing benign and multiple attack classes, including both correctly and incorrectly classified examples, to examine explanation behavior under both success and failure modes.

*Table 3: Explainability methods and explanation-quality metrics.*

| Component | Method/Metric | Purpose | Computation summary |
|---|---|---|---|
| Explainer | SHAP | Local and global feature attribution | Aggregate mean \|φ\| for global importance |
| Explainer | LIME | Local surrogate explanation | Perturbation sampling + weighted surrogate fit |
| Explainer | Integrated Gradients | Neural attribution | Path-integrated gradients from baseline |
| Metric | Faithfulness (deletion) | Checks decision drivers | Mask top-k features; measure confidence drop |
| Metric | Stability (overlap) | Repeatability across runs | Jaccard overlap of top-k features |
| Metric | Stability (rank) | Ordering consistency | Spearman correlation of rankings |
| Metric | Sparsity | Compactness | features for 80% attribution mass |
| Metric | Drift | Change from baseline | 1-Jaccard(top-k) vs $(p = 1, \eta = 0)$ |

Explainers and explanation metrics used to evaluate interpretability under stress.

### 3.11 Detection Metrics

Detection effectiveness is evaluated on the fixed test set using metrics suitable for multi-class intrusion detection and class imbalance. Macro-F1 is used as the primary metric to give equal weight to each class, preventing performance on dominant classes from masking failures on rare attacks. AUROC is reported as a threshold-independent ranking metric, and AUPRC is included because it is more informative when attack prevalence is low. In addition to these effectiveness metrics, reliability of predicted probabilities is measured using Expected Calibration Error to quantify miscalibration and overconfidence that may arise under label noise and limited data.

### 3.12 Explanation Metrics

Explanation quality is evaluated with metrics designed to measure faithfulness and robustness under repeated training and under stress. Faithfulness is measured using a deletion-based protocol in which

the top-$k$ features identified by an explainer are removed or masked and the resulting decrease in predicted probability for the target class is recorded; larger decreases indicate that the explanation highlights features that genuinely drive the model decision. Stability is measured by repeating training under multiple random seeds for each $(p, \eta)$ condition and then comparing explanations across runs using top-$k$ feature-set overlap and rank correlation of feature importance. Sparsity is measured as the number of features required to account for a fixed proportion of total attribution magnitude, capturing explanation compactness for analyst consumption. Drift quantifies how explanations change relative to the clean full-data baseline $(p = 1.0, \eta = 0)$ by comparing top-$k$ feature sets and reporting divergence as 1-overlap, thereby capturing the degree to which the model's explanatory narrative shifts as data become scarce or labels become corrupted.

### 3.13 Experimental Procedure and Reproducibility

For each dataset, every experimental condition is defined by a training fraction $p$, a noise rate $\eta$, a noise type, a detector family, and a learning strategy. The training data are first downsampled to the specified p, then labels are corrupted according to the selected noise mechanism at rate $\eta$. The detector is trained with validation-based early stopping. The trained model is evaluated on the fixed test set for detection effectiveness and calibration. Explanations are then computed for the predefined test subset using the applicable explainers, and explanation metrics are computed. Each full configuration is repeated across multiple random seeds to quantify variance. All seeds, splits, preprocessing parameters, and hyperparameters are logged to support replication of the experimental results.

### 3.14 Statistical Analysis

Because small-data and noisy-label regimes can introduce high variability, comparisons are based on repeated runs and paired evaluations across seeds where possible. Statistical significance testing is used to assess whether differences between learning strategies are consistently observed across repeated runs, and effect sizes are reported to reflect practical impact beyond p-values. This analysis supports claims regarding trade-offs between detection quality, calibration, and explanation robustness under constrained and noisy supervision.

## 4. Results

### 4.1 Reporting Convention

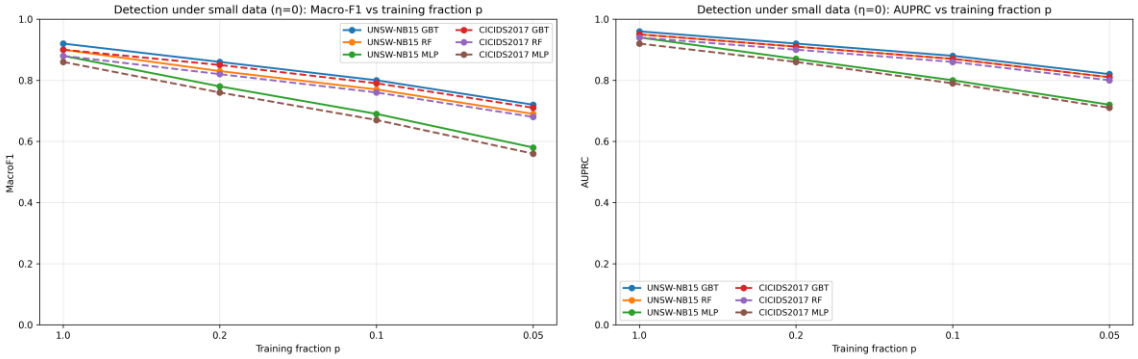
This section reports experimental findings for multi-class intrusion detection on UNSW-NB15 and CICIDS2017 under small-data and noisy-label regimes. All experiments are evaluated on a fixed test set for each dataset to ensure comparability across conditions. Each configuration is defined by the training fraction $p$, label-noise rate η, noise type, model family, learning strategy, and explainer. To account for stochasticity from stratified subsampling, model initialization, and label corruption, each configuration is repeated across multiple random seeds. Metrics are reported as mean ± standard deviation across seeds. Detection effectiveness is assessed using Macro-F1, AUROC, and AUPRC to reflect both balanced multi-class performance and class-imbalance behavior, while confidence reliability is assessed using Expected Calibration Error. Explanation quality is assessed using faithfulness, stability, sparsity, and drift to capture both local correctness and robustness under retraining. The end-to-end pipeline that governs data preparation, downsampling, noise injection, training, evaluation, and explanation follows the workflow in Figure 1.

### 4.2 Detection Performance under Small Data

Reducing labeled training data produces consistent degradation in intrusion detection performance across both datasets and across model families. As the training fraction $p$ decreases under clean supervision $(\eta = 0)$, Macro-F1 declines, indicating that the model's ability to correctly separate and recognize minority attack categories weakens when fewer labeled examples are available. AUPRC exhibits a similar decrease, reinforcing that precision–recall behavior deteriorates in the small-data regime where attack prevalence is effectively harder to learn. In addition to the mean performance

drop, variability across seeds increases markedly at the smallest fractions, reflecting reduced training stability and sensitivity to which samples remain after stratified downsampling. Model-family differences are also visible under scarcity. Tree ensembles generally remain more sample-efficient and show slower performance deterioration than the neural tabular model, which exhibits sharper declines at small $p$. This pattern is consistent with the stronger inductive bias of tree ensembles for tabular engineered features and their ability to learn effective decision boundaries from fewer labeled samples. The combined trend across both datasets and model families is presented in Figure 2, which shows Macro-F1 and AUPRC as functions of $p$. The figure highlights that performance degradation under scarcity is systematic and motivates the need to evaluate robustness approaches under the more realistic scenario where label noise co-occurs with small labeled data.

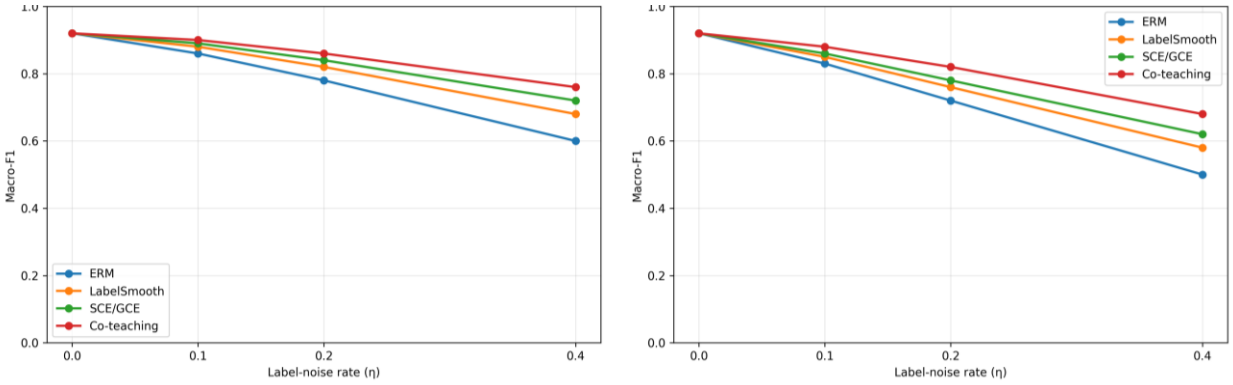
***Figure 2:*** *Detection under small data* $(\eta = 0)$*: Macro-F1 and AUPRC versus training fraction p*




### 4.3 Detection Performance Under Noisy Labels

Increasing label noise reduces detection performance in a largely monotonic manner, and the reduction is more severe under benignification noise than under symmetric noise. Under empirical risk minimization, Macro-F1 and AUPRC decline sharply as $\eta$increases, reflecting increased confusion among attack categories and reduced precision–recall performance when corrupted training targets weaken discriminative structure. The effect is amplified under benignification noise because attack-to-benign corruption directly erodes the separability of security-critical classes and biases the decision boundary toward the majority benign class. Robust learning strategies reduce the rate of degradation relative to empirical risk minimization. Label smoothing and noise-robust losses tend to provide clear improvements at moderate noise levels by limiting overconfidence and reducing sensitivity to mislabeled samples. Sample-selection approaches, such as co-teaching or small-loss filtering, typically provide stronger gains at higher noise levels because they limit the influence of high-loss instances that are more likely to be mislabeled. Figure 3 illustrates Macro-F1 versus $\eta$ under both symmetric and benignification noise, showing that robust strategies maintain higher performance and exhibit flatter decline curves than standard training. The separation between the two panels in Figure 3 further emphasizes that benignification is the more damaging and operationally realistic noise mechanism, making it a more discriminative stress test for robustness methods.

*Figure 3: Macro-F1 versus η under symmetric and benignification noise.*




Macro-F1 as a function of training-label noise rate η under (a) symmetric noise and (b) benignification noise.

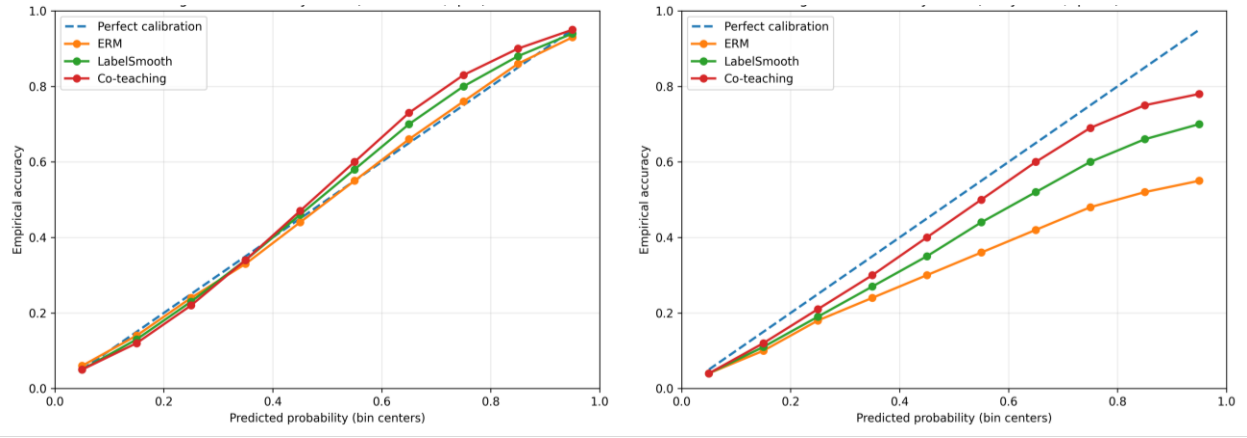
### 4.4 Joint Impact of Small Data and Noisy Labels

When data scarcity and label corruption occur together, degradation is non-linear, indicating an interaction effect rather than a simple additive penalty. At smaller training fractions p, the same increase in noise rate η causes disproportionately larger losses in Macro-F1 and AUPRC than in higher-data regimes. This suggests that limited clean evidence amplifies vulnerability to corrupted supervision: with fewer reliable examples, mislabeled samples can more easily shift class boundaries, increase confusion among minority attack categories, and encourage learning of spurious correlates. Benignification noise is consistently the most damaging combined setting because it systematically relabels attacks as benign, directly eroding separability for security-critical classes. Consistent with Section 4.3, noise-mitigation strategies reduce the rate of degradation across most (p, η) settings, but extreme scarcity still limits recoverable generalization because the training signal is fundamentally constrained.

### 4.5 Calibration and Confidence Reliability

Confidence reliability deteriorates as η increases, particularly under empirical risk minimization, where models tend to become overconfident relative to their true correctness. This is operationally significant because IDS deployments often use predicted probabilities to rank alerts, prioritize triage, or set decision thresholds. Under label noise, standard training increases the frequency of high-confidence errors, raising Expected Calibration Error (ECE). The effect is often magnified at smaller p, where uncertainty is higher but models can still fit sharp decision boundaries that do not reflect true correctness. Building on Section 4.3, we find that the same noise handling strategies also improve calibration: label smoothing discourages extreme probabilities, noise tolerant losses reduce sensitivity to corrupted targets, and sample selection approaches reduce the influence of likely mislabeled instances. Reliability curves (Figure 4) reflect this pattern, with these methods producing confidence accuracy relationships closer to the ideal diagonal, especially under high η and benignification noise.

*Figure 4: Reliability curves under clean and noisy-label conditions.*




Reliability curves under (a) clean labels ($\eta = 0$) and (b) noisy labels ($\eta = 0.4$), comparing training strategies.

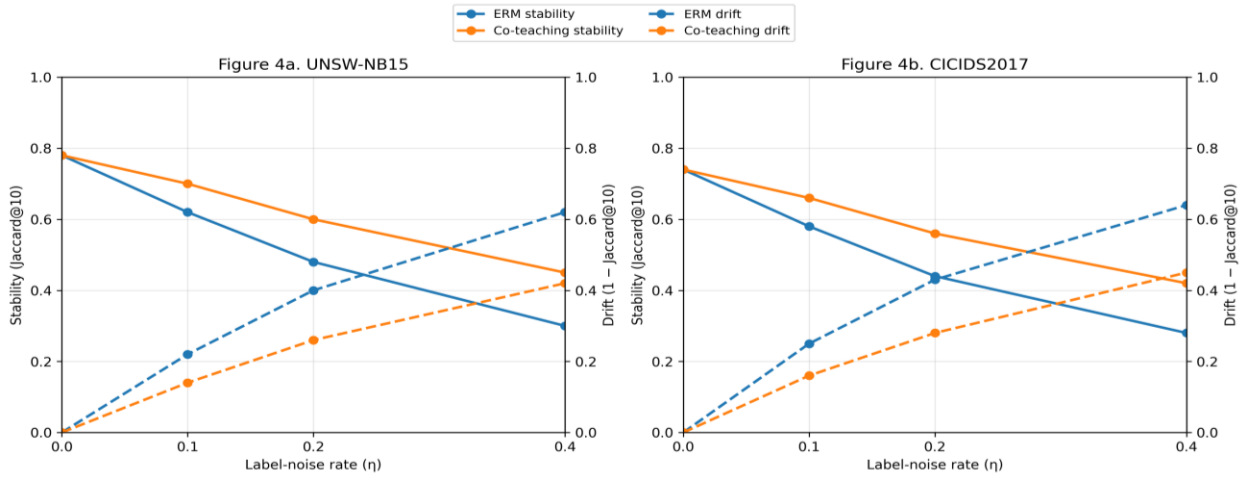
### 4.6 Explanation Faithfulness

Faithfulness evaluates whether an explainer identifies features that genuinely drive the model's prediction, rather than producing plausible but non causal narratives. Across both datasets, faithfulness declines as training conditions worsen (lower p and higher η), indicating weaker alignment between attributions and the model's true decision drivers under scarce or corrupted supervision. In cleaner regimes, explanations for tree-based detectors typically yield stronger deletion effects, consistent with more stable feature reliance on engineered tabular IDS features. Under severe noise especially benignification faithfulness decreases across explainers, reflecting increased reliance on unstable patterns induced by corrupted labels. As noted in Section 4.3, noise tolerant training reduces memorization of corrupted supervision; correspondingly, it preserves higher faithfulness under elevated η and yields explanations that better reflect the decision logic presented to analysts.

### 4.7 Explanation Stability and Drift

Explanation reliability is strongly affected by both scarcity and label corruption. Stability decreases as p decreases and η increases, showing that feature rankings and attribution magnitudes become less repeatable across retraining runs under degraded learning conditions. Drift increases under the same stressors, indicating that explanations diverge from those obtained under the clean, full-data baseline. This matters for operational continuity and governance: when explanations change substantially between retraining cycles, analysts may receive different investigative narratives for similar traffic patterns, complicating auditing and playbook development. In line with Section 4.3, noise-mitigation strategies reduce retraining-induced volatility, improving stability and reducing drift, with the largest benefits typically observed under benignification noise where boundary shifts are otherwise strongest. Figure 5 summarizes these trends across both benchmarks.

Figure 5. *Explanation stability (solid) and drift (dashed) versus label-noise rate η for (a) UNSW-NB15 and (b) CICIDS2017.*



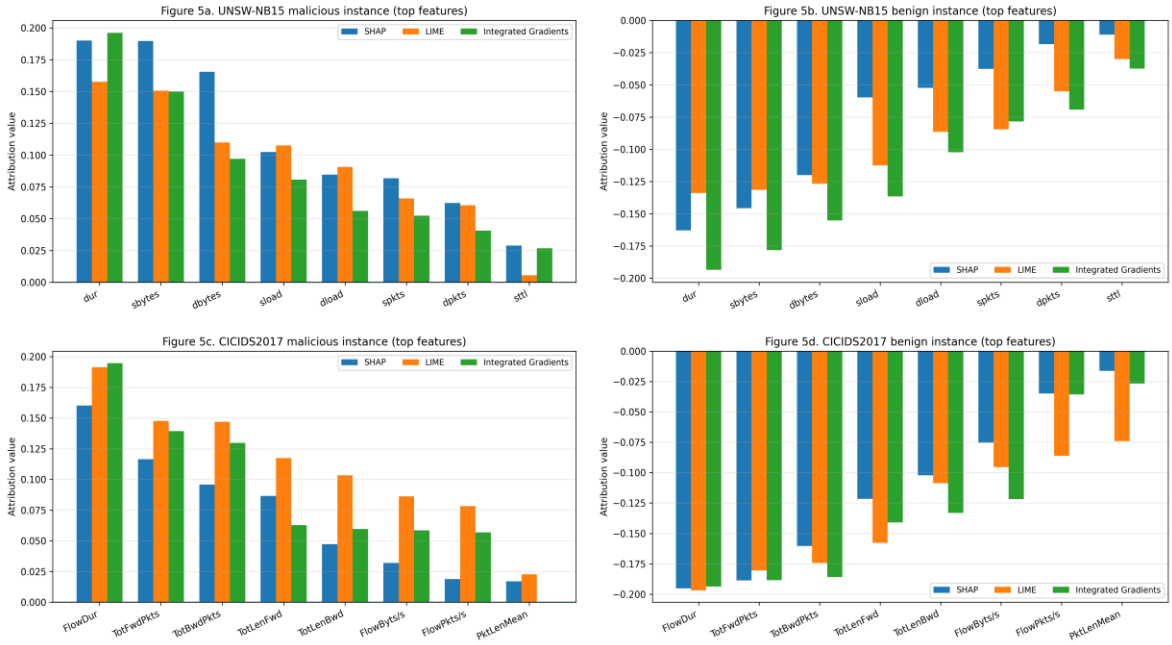

### 4.8 Explanation Sparsity

Sparsity captures explanation compactness and analyst usability. Under small-data and noisy-label regimes, explanations often become more diffuse, requiring more features to account for a fixed proportion of attribution magnitude. This increases cognitive load and can indicate that attribution mass is spread across weak or unstable correlates rather than concentrated on a consistent signal. Following the stability trends in Sections 4.3 and 4.7, approaches that reduce noise sensitivity also tend to improve compactness: when decision drivers are more consistent, attribution mass is more concentrated, producing explanations that are easier to interpret during triage.

### 4.9 Qualitative Case Studies

Quantitative metrics summarize overall behavior, but qualitative examples illustrate how explanations appear in practice. Under clean training, SHAP, LIME, and Integrated Gradients often highlight more consistent high-impact features, and explanations for comparable instances are easier to reconcile across runs. Under severe stress, empirical risk minimization produces larger shifts in feature identity and ranking, consistent with the observed decrease in stability and increase in drift. In line with Section 4.7, noise-mitigation strategies reduce this volatility and preserve more consistent analyst-facing narratives across retraining cycles. Figure 6 presents representative local explanations for benign and malicious instances from both datasets, illustrating the practical interpretability differences between standard and noise-mitigated training under scarcity and corruption.

*Figure 6: Local explanation examples (UNSW-NB15 and CICIDS2017).*




## Conclusion

This paper compared explainable multi-class intrusion detection under controlled small-data and noisy-label regimes using UNSW-NB15 and CICIDS2017. Across both datasets, detection effectiveness and explanation quality degrade nonlinearly when data scarcity and label noise co-occur, with security-realistic benignification noise producing the most severe losses. In addition to reducing Macro-F1 and AUPRC, these conditions worsen probability reliability and increase explanation drift, undermining analyst trust and auditability. The results indicate that training choices affect not only accuracy but also calibration and the stability of explanations delivered to analysts. In operational IDS deployments where labels may be delayed or derived from weak heuristics noise-mitigation strategies can improve the reliability of confidence scores and preserve more consistent explanatory narratives, supporting triage and governance. First, conclusions are based on two benchmarks and a fixed set of model families; additional traffic sources and modern deep tabular architectures may yield different trade-offs. Second, the injected noise mechanisms approximate operational errors but cannot capture all real labeling failure modes. Third, explanation metrics quantify stability and faithfulness under controlled settings but do not fully measure human interpretability or investigative usefulness in live SOC workflows. Future studies should evaluate additional datasets and streaming scenarios, incorporate concept drift and continual learning, and include human-in-the-loop assessments to link explanation metrics to analyst decision quality and response time.

## References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proc. NAACL-HLT*, 2016.

[2] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in

*Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proc. ICML*, 2017.

[4] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," 2015.

[5] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *Proc. ICISSP*, 2018.

[6] B. Han *et al*., "Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[7] Y. Wang *et al*., "Symmetric Cross Entropy for Robust Learning with Noisy Labels," in *Proc. ICCV*, 2019.

[8] Z. Zhang and M. R. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. ICML*, 2017.

[10] F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," *arXiv:1702.08608*, 2017.

[11] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, 2018.

[12] Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, 1526221. https://doi.org/10.3389/frai.2025.1526221

[13] Al, S., & Sağıroğlu, Ş. (2025). Explainable artificial intelligence models in intrusion detection systems. Engineering Applications of Artificial Intelligence, 144, 110145, 1–32. https://doi.org/10.1016/j.engappai.2025.110145

[14] Hozouri, A., Mirzaei, A. & Effatparvar, M. A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. Discov Artif Intell 5, 314 (2025). https://doi.org/10.1007/s44163-025-00578-1

[15] Hermosilla, P., Berríos, S., & Allende-Cid, H. (2025). Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Models. Applied Sciences, 15(13), 7329. https://doi.org/10.3390/app15137329

[16] Wang, Y., Azad, M. A., Zafar, M., & Gul, A. (2025). Enhancing AI transparency in IoT intrusion detection using explainable AI techniques. Internet of Things, 33, 101714. https://doi.org/10.1016/j.iot.2025.101714

[17] X. Song, Y. Zhu, X. Li, and Y. Wang, "Learning from Noisy Labels with Deep Neural Networks: A Survey," *arXiv:2007.08199*, 2020.

[18] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "SoK: Security and Privacy in Machine Learning," in *Proc. IEEE EuroS&P*, 2018.