WOCI: Weighted Operational Cost Index A New Evaluation Metric for Intrusion Detection

Vagner E. Quincozes[1], Silvio E. Quincozes[2], Célio Albuquerque[1], Diego Passos[3], Daniel Mossé[4]

[1]Universidade Federal Fluminense - UFF
[2]PPGES, Universidade Federal do Pampa - UNIPAMPA
[3]Instituto Superior de Engenharia de Lisboa - ISEL
[4]Universidade de Pittsburgh - PITT

vequincozes@midiacom.uff.br, silvioquincozes@unipampa.edu.br,
celio@ic.uff.br, diego.passos@isel.pt, mosse@pitt.edu

Abstract. *This paper proposes the Weighted Operational Cost Index (WOCI), a post-hoc metric for evaluating binary detection configurations under asymmetric error costs. WOCI combines averaged counts of false positives and false negatives through a configurable cost ratio $R = C_{FN}/C_{FP}$, normalized by the number of positive instances. We instantiate WOCI in an Intrusion Detection System (IDS) case study using the ERENO and 5G-NIDD datasets with CatBoost and XGBoost, comparing configurations from Built-in importance, SHAP, and Mutual Information across multiple feature subset sizes. Results show that 13 of 20 configurations keep the same preferred strategy across $R \in [1, 100]$, while seven change at a crossover point R^*. The main transitions occur in 5G-NIDD with XGBoost near $R \approx 10$, a plausible enterprise cost scenario.*

1. Introduction

Intrusion Detection Systems (IDS) are a key component of enterprise cybersecurity, especially in environments with high-volume network traffic, heterogeneous devices, cloud services, and increasingly complex attack surfaces. Machine learning models are widely used to detect attacks and anomalies, but their evaluation is often centered on predictive metrics such as F1-Score, False Negative Rate (FNR), and False Positive Rate (FPR). Although these metrics are informative, they do not directly express the operational cost imposed on an organization when a model generates false alarms (i.e., false positives) or misses real attacks (i.e., false negatives). This limitation has motivated cost-sensitive approaches to IDS evaluation, which argue that detection models should be assessed in terms of statistical accuracy as well as operational, response, and damage-related costs [Cha et al. 2026, Gombar et al. 2026].

In intrusion detection, this limitation matters because False Positives (FP) and False Negatives (FN) have asymmetric consequences. A false positive may trigger unnecessary investigation, consume analyst time, and increase alert fatigue, whereas a false negative may allow an intrusion to remain undetected, causing downtime, incident response costs, data exposure, regulatory consequences, and reputational damage. In this setting, two configurations with similar predictive metrics may still produce different operational consequences if their FP/FN profiles differ. This is consistent with recent IDS literature showing that false positives directly affect Security Operations Center (SOC)

workflows by increasing analyst fatigue and reducing trust in IDS outputs, while false negatives may leave attacks undetected and increase incident response and recovery impacts [Skrodelis and Romanovs 2026, Hozouri et al. 2025, Nelson et al. 2025].

The economic relevance of this asymmetry is supported by external evidence. The IBM *Cost of a Data Breach Report 2025* estimates the global average cost of a data breach at USD 4.4 million, while the U.S. Bureau of Labor Statistics reports a 2024 median pay of USD 60.05 per hour for information security analysts [IBM 2025, U.S. Bureau of Labor Statistics 2025]. These values should not be interpreted as universal costs for a single false positive or false negative. They show why alert investigation and missed intrusions should not be treated as equivalent operational events.

Cost-sensitive IDS research has addressed this problem through model training, class-imbalance handling, threshold optimization, alert management, and resource-aware feature selection [Telikani and Gandomi 2021, Gupta et al. 2022, Cha et al. 2026, Chen et al. 2024]. These approaches are valuable, but they typically modify the learning process, the decision threshold, or the selection algorithm. A complementary gap remains: evaluating already trained binary detection configurations under different enterprise cost assumptions, without changing the model or the decision threshold.

To address this gap, this work proposes the *Weighted Operational Cost Index* (WOCI), a post-hoc cost-sensitive evaluation index for binary detection tasks. WOCI converts averaged counts of false positives and false negatives observed across experimental runs into a normalized operational cost measure. In this paper, WOCI is instantiated through an IDS case study using two datasets (ERENO [Quincozes et al. 2023], a smart-grid intrusion detection dataset, and 5G-NIDD [Siriwardhana et al. 2025], a 5G network intrusion detection dataset), two gradient-boosting classifiers (CatBoost and XGBoost), and three feature-ranking strategies (Built-in importance, SHAP, and Mutual Information) across multiple retained feature subset sizes. The implementation is publicly available for reproducibility and external cost-ratio analysis[1].

2. Related Work

Cost-sensitive evaluation has been widely studied as a response to the limitations of treating all classification errors as equivalent. Foundational work on cost-sensitive learning and cost curves showed that classifier comparison may change when false positives and false negatives have different consequences [Elkan 2001, Drummond and Holte 2006]. Cost curves provide a threshold-oriented view of classifier performance across the ROC space. WOCI follows the same general motivation of unequal error costs, but addresses a different evaluation setting: it operates on observed confusion-matrix outcomes at a fixed operating point, making it applicable to already evaluated configurations without probability calibration or threshold sweeping.

In intrusion detection, cost-sensitive reasoning has traditionally been motivated by the operational asymmetry between missed attacks and false alarms. Earlier IDS-oriented studies framed intrusion detection as a cost-sensitive decision problem, considering operational costs, damage costs, and response costs rather than relying only on predictive performance [Lee et al. 2002, Stolfo et al. 2000]. More recent work has incorporated cost sensitivity into IDS training, class-imbalance handling, and decision-threshold

[1] `https://github.com/vagnerereno/woci`

adjustment. For example, Telikani and Gandomi proposed cost-sensitive stacked autoencoders for intrusion detection in IoT environments, while CSE-IDS uses cost-sensitive deep learning and ensemble methods to address class imbalance in network-based intrusion detection [Telikani and Gandomi 2021, Gupta et al. 2022].

More recently, Cha proposed a cost-sensitive threshold optimization approach for NIDS with XGBoost, explicitly using asymmetric false-negative and false-positive costs to adjust decision thresholds [Cha et al. 2026]. These studies recognize that missed attacks and false alarms should not be treated as operationally equivalent. However, they mainly incorporate cost into the learning process or threshold selection, whereas WOCI evaluates already trained configurations.

A complementary issue arises when different IDS configurations produce similar predictive performance but different false-positive and false-negative profiles. Such configurations may result from alternative models, thresholds, preprocessing strategies, or feature subsets. In this paper, feature selection is used in this role: not as the proposed contribution, but as a controlled way to generate multiple IDS configurations whose operational cost can be compared post hoc. This is relevant because feature reduction can affect both efficiency and detection behavior, especially in network traffic datasets with redundant, noisy, or costly-to-process attributes.

Recent IDS studies have explored feature selection mainly to improve efficiency, reduce dimensionality, and support deployment in resource-constrained environments. Examples include ensemble-based feature selection for lightweight IDS in IoT devices and information-theoretic feature selection for network intrusion detection [Fatima et al. 2024, Westphal et al. 2025]. However, selected subsets are usually compared through conventional metrics and computational criteria, while the operational cost induced by their specific FP/FN profiles remains implicit.

The closest work to this study is RecoSelector, which proposes cost-sensitive feature selection for network intrusion detection in resource-constrained IoT environments [Chen et al. 2024]. RecoSelector is highly related because it combines NIDS, feature selection, and cost awareness. Its objective is to select feature subsets under resource and deployment constraints during the selection process. WOCI addresses a different stage of the pipeline: it does not propose a classifier, thresholding strategy, or feature-selection algorithm, but evaluates the FP/FN cost profile after training. In this sense, a cost-sensitive selector may generate candidate configurations, while WOCI can compare their post-hoc operational cost under different enterprise risk scenarios.

While prior studies have incorporated cost sensitivity into IDS training, threshold optimization, alert management, and feature-selection algorithms, there is still a need for simple post-hoc evaluation indices that translate observed FP/FN outcomes into an interpretable operational cost measure. WOCI addresses this gap by operationalizing it as a normalized post-hoc evaluation layer applicable to any binary detection configuration with observable FP/FN outcomes. In this study, feature-ranking methods are used to instantiate this layer over multiple IDS configurations generated from Built-in importance, SHAP, and Mutual Information, as summarized in Table 1.

Table 1. Positioning of WOCI in relation to cost-sensitive IDS evaluation and feature-selection studies.

Work / line	Main focus	Relation to WOCI	Main difference
[Elkan 2001, Drummond and Holte 2006]	Cost-sensitive learning and cost curves	Establish the basis for unequal error costs	General classifier evaluation; not IDS-oriented operational analysis
[Lee et al. 2002, Stolfo et al. 2000]	Cost-based IDS and intrusion/fraud detection	Frame detection as an operational cost problem	Broader cost modeling and response; not a normalized post-hoc index
[Telikani and Gandomi 2021, Gupta et al. 2022]	Cost-sensitive IDS learning	Incorporate cost into model training and imbalance handling	Cost is embedded in learning; WOCI evaluates trained configurations
[Chen et al. 2024]	Cost-sensitive feature selection for NIDS/IoT	Closest related line combining cost, NIDS, and feature selection	Selects features under resource constraints; WOCI evaluates observed FP/FN outcomes
[Fatima et al. 2024, Westphal et al. 2025]	Feature selection for efficient IDS	Address feature reduction in IDS/NIDS	Focus on efficiency and predictive performance, not FP/FN operational cost
[Cha et al. 2026]	Cost-sensitive threshold optimization	Uses asymmetric false-negative and false-positive costs	Optimizes thresholds; WOCI evaluates post-hoc operational cost
Proposed WOCI	Post-hoc operational evaluation	Converts observed FP/FN outcomes into normalized enterprise cost	Enables sensitivity analysis over $R = C_{FN}/C_{FP}$ across binary detection configurations

3. Proposed Methodology: Weighted Operational Cost Index

This section presents the proposed WOCI, a post-hoc cost-sensitive evaluation index for binary detection tasks. Although the index can be applied to any domain in which false positives and false negatives have asymmetric operational consequences, this study instantiates it in the context of intrusion detection. In this setting, a false positive corresponds to a false alarm that may consume analyst time and increase alert fatigue, whereas a false negative corresponds to a missed intrusion that may lead to incident response costs, downtime, data exposure, regulatory consequences, and reputational damage.

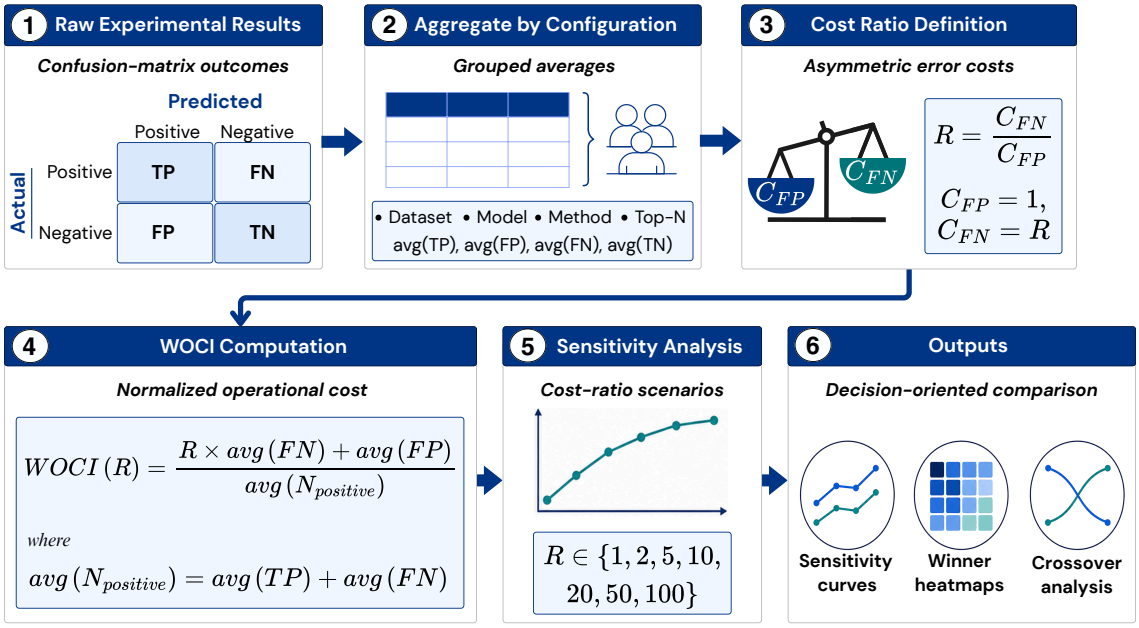

Figure 1. Workflow for computing WOCI, from confusion-matrix aggregation to cost-ratio sensitivity analysis and decision-oriented outputs.

Figure 1 summarizes the proposed workflow in six steps. In Step 1, the method

starts from raw experimental results containing confusion-matrix outcomes for each evaluated configuration. In Step 2, these outcomes are aggregated by configuration. In Step 3, the asymmetric cost relation between false negatives and false positives is defined through a relative cost ratio. In Step 4, the WOCI score is computed as a normalized operational cost. In Step 5, the computation is repeated under multiple cost-ratio scenarios. Finally, in Step 6, the results are organized into sensitivity curves, winner summaries, and crossover analyses to support decision-oriented comparison among evaluated configurations.

3.1. Aggregation of confusion-matrix outcomes

As shown in Step 1 of Figure 1, WOCI requires the confusion-matrix outcomes obtained for each experimental run or fold. These outcomes are associated with the corresponding dataset, classifier, evaluated configuration, and run or fold. This makes it possible to evaluate any configuration that produces binary classification outcomes, including different models, feature subsets, preprocessing pipelines, or decision thresholds.

In Step 2, the confusion-matrix components are averaged across runs or folds for each configuration:

$$\overline{TP},\ \overline{TN},\ \overline{FP},\ \overline{FN}. \tag{1}$$

Each component reflects the expected count per test fold. Averaging rather than summing is used so that WOCI remains comparable across experimental designs with different numbers of folds or repetitions, and across datasets of different sizes.

The average number of positive instances per fold is defined as:

$$\overline{N}_{positive} = \overline{TP} + \overline{FN}. \tag{2}$$

False negatives are instances that belong to the positive class (real attacks) but were incorrectly classified as negative. Therefore, $\overline{TP}$ (detected attacks) and $\overline{FN}$ (missed attacks) together account for all positive instances in the test fold. As shown in Step 4 of Figure 1, this quantity serves as the normalization factor in the WOCI formula, so that the score reflects the weighted operational cost per relevant event independently of dataset scale. Normalizing by $\overline{N}_{positive}$ rather than by the negative class size is intentional: the operational question is how much cost is incurred per real attack, not per benign instance.

3.2. Cost-Ratio Definition

In Step 3, the framework defines the relative cost between the two types of operational error. Let C_{FP} be the cost of a false positive and C_{FN} be the cost of a false negative. The relative severity of false negatives is represented by:

$$R = \frac{C_{FN}}{C_{FP}}. \tag{3}$$

Without loss of generality, the false-positive cost is set as the unit reference:

$$C_{FP} = 1. \tag{4}$$

Therefore:

$$C_{FN} = R. \tag{5}$$

With this formulation, the analysis depends on the relative severity of the two errors. The parameter R represents how many times a false negative is assumed to cost more than a false positive. For example, $R = 10$ means that a missed intrusion is considered ten times more costly than a false alarm.

3.3. WOCI Computation

In Step 4, WOCI is computed for each experimental configuration and each cost ratio R:

$$WOCI(R) = \frac{R \cdot \overline{FN} + \overline{FP}}{\overline{N}_{positive}}. \tag{6}$$

The numerator represents the weighted misclassification cost. False negatives are multiplied by R, while false positives are assigned unit cost. The denominator normalizes this weighted cost by the average number of positive instances, as defined in Equation 2. Thus, WOCI can be interpreted as the weighted operational cost per positive instance. Lower WOCI values indicate lower expected operational cost under the specified cost ratio.

The normalization by $\overline{N}_{positive}$ supports comparisons across datasets with different sizes and class distributions. However, in datasets with very few positive instances per fold, small variations in $\overline{FN}$ may be amplified, especially for high values of R. This limitation is addressed explicitly in Section 7.

3.4. Sensitivity Analysis over Cost Ratios

In Step 5, WOCI is evaluated over a range of cost ratios:

$$R \in \{1, 2, 5, 10, 20, 50, 100\}. \tag{7}$$

When $R = 1$, false positives and false negatives are treated as equally costly. Larger values represent environments where missed intrusions are increasingly more severe than false alarms. In principle, $0 < R < 1$ is also mathematically valid and would represent settings where false alarms are more costly than missed detections; however, in the IDS domain, $R \gg 1$ is the operational norm, since an undetected intrusion typically causes substantially more harm than a spurious alert. This sensitivity analysis examines whether the preferred configuration remains stable across cost assumptions or changes as false negatives become more expensive.

3.5. Decision-Oriented Outputs

Finally, Step 6 produces three complementary outputs. First, sensitivity curves show how WOCI changes as R increases. Second, winner summaries identify the configuration with the lowest WOCI for each evaluated cost scenario. Third, crossover analysis identifies cost-ratio thresholds at which the ranking between configurations changes.

For two configurations A and B, the crossover point is found by solving $WOCI_A(R) = WOCI_B(R)$:

$$R^* = \frac{\overline{FP}_A - \overline{FP}_B}{\overline{FN}_B - \overline{FN}_A}, \tag{8}$$

provided that $\overline{FN}_A \neq \overline{FN}_B$ and that R^* is positive and finite. Because $WOCI(R)$ is linear in R for any fixed configuration (it is a first-degree polynomial in R), the curves for any two configurations A and B are two straight lines. Two distinct straight lines intersect in at most one point, so there is at most one crossover point for any pair of configurations. When $\overline{FN}_A = \overline{FN}_B$, the two curves have the same slope (parallel lines) and the ranking is determined entirely by the FP counts, independently of R.

These outputs support a decision-oriented interpretation: instead of selecting a configuration by conventional metrics such as F1-Score alone, WOCI allows the analyst to identify the option that minimizes operational cost under the organization's risk assumptions, and to determine whether that choice is robust to uncertainty in those assumptions. Section 6 provides examples of both robust and cost-sensitive decisions.

4. Interpreting the Cost Ratio in Enterprise IDS

The cost ratio $R = C_{FN}/C_{FP}$ can be interpreted as an enterprise-specific calibration parameter rather than as a fixed universal constant. In WOCI, the false-positive cost is set as the unit reference, $C_{FP} = 1$, and the false-negative cost is expressed relatively as $C_{FN} = R$. Thus, R indicates how many times a missed intrusion is assumed to be more costly than a false alarm.

The cost of a false positive can be estimated from the operational effort required to investigate an incorrect alert. A simple approximation is:

$$C_{FP} = n \times h \times c_h, \tag{9}$$

where n is the number of professionals involved, h is the number of hours spent on triage and escalation, and c_h is the average hourly cost per professional. This cost may include analyst time, overtime, on-call activation, incident documentation, network or infrastructure support, and managerial coordination. Public labor data provide a reference[2] for the magnitude of this cost: the U.S. Bureau of Labor Statistics reports that information security analysts had a median annual wage of USD 124,910 in May 2024 [U.S. Bureau of Labor Statistics 2025].

The cost of a false negative is typically more difficult to estimate because a missed intrusion may have multiple layers of impact. These may include incident response, service downtime, data exposure, regulatory penalties, contractual consequences, reputational damage, and business interruption. Industry reports reinforce the financial relevance of this type of event. For instance, the IBM *Cost of a Data Breach Report 2025* estimates the global average cost of a data breach at USD 4.4 million [IBM 2025]. This value should not be interpreted as the cost of a single false negative in all contexts, but it illustrates that undetected or late-detected security incidents can produce consequences far beyond the cost of investigating a false alarm.

In regulated critical infrastructures, part of the financial impact of a missed detection may be estimated from sector-specific rules. In the Brazilian electric power transmission sector, ANEEL publishes the Annual Allowed Revenue (*Receita Anual Permitida*, RAP) of transmission companies, including official data for the 2025–2026 cycle [ANEEL 2025, ANEEL 2026]. The Brazilian National System Operator (ONS) defines the *Parcela Variável por Indisponibilidade* (PVI), a monthly value deducted from the RAP of a transmission utility due to outages in its facilities [Operador Nacional do Sistema Elétrico 2019]. These mechanisms do not define a universal cost for cyber incidents, but they provide a transparent way to illustrate how R can be calibrated when service unavailability has measurable financial consequences.

[2]Note that this estimate should be interpreted as a calibration example rather than a general rule, since the actual false-positive cost depends on each organization's staffing model, escalation procedures, tooling, and incident response workflow.

Let $PB = RAP/12$ be the monthly payment base of a transmission function. For a month with D days, the base hourly value is $PB/(D \times 24)$. If a non-programmed unavailability event is subject to a regulatory multiplier K defined by the applicable rule, the hourly financial impact can be approximated as:

$$C_{FN}^{hour} \approx \frac{PB}{D \times 24} \times K. \tag{10}$$

Given an estimated false-positive investigation cost C_{FP} and an unavailability duration t in hours, the corresponding cost ratio can be expressed as:

$$R = \frac{C_{FN}^{hour} \times t}{C_{FP}}. \tag{11}$$

Thus, values such as $R = 20$, $R = 50$, or $R = 100$ can be plausible when the cost of a missed intrusion includes service unavailability, incident response, or regulatory exposure. For example, with RAP of R$ 12 million, $K = 150$, $D = 30$, $t = 1$ hour, and $C_{FP} =$ R$ 10,000, Equations (10) and (11) yield $R \approx 20$, placing this scenario above the crossover thresholds observed in Section 6.

5. Experimental Design

Figure 2 summarizes the experimental set up used in this paper to generate WOCI inputs. The study focuses on two intrusion detection datasets, ERENO and 5G-NIDD, both formulated as binary classification tasks. The positive class represents malicious, anomalous, or attack-related traffic, while the negative class represents normal or benign traffic.

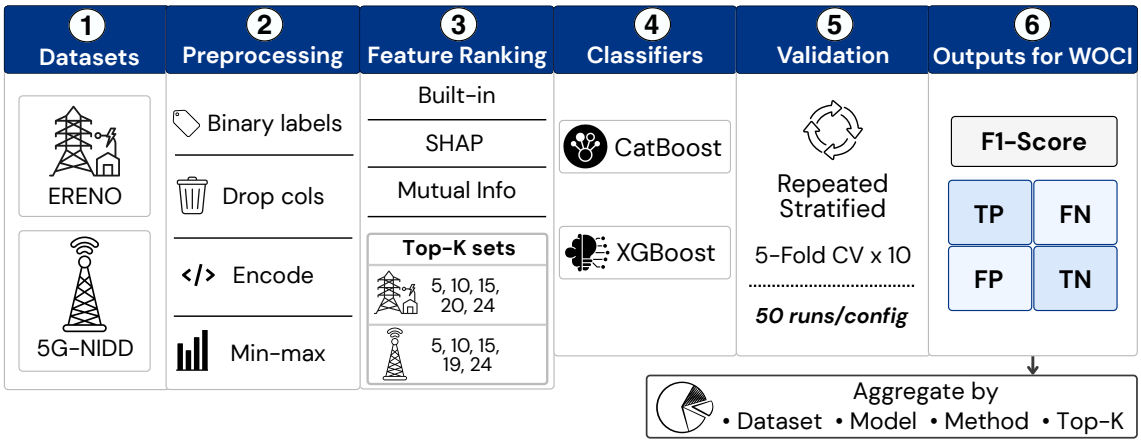

Figure 2. Experimental instantiation used to generate WOCI inputs: dataset preprocessing, feature-ranking configurations, validation protocol, and confusion-matrix aggregation for WOCI computation.

For ERENO [Quincozes et al. 2023], records labeled as *normal* were mapped to class 0, while all remaining attack classes were mapped to class 1. The columns *Time*, *GooseTimestamp*, and *t* were removed to prevent trivial memorization of temporal ordering, since these fields encode the sequence of events rather than properties of the traffic itself. For 5G-NIDD [Siriwardhana et al. 2025], records labeled as *Benign* were mapped to class 0, while all remaining records were mapped to class 1. The columns *Attack Type*,

Attack Tool, *Seq*, and *Offset* were removed to avoid using attack descriptors or sequence-related fields as predictive features, since they directly encode the class label.

After dataset-specific preprocessing, categorical features were encoded using label encoding, missing values were imputed with column means, and all features were scaled using Min-Max normalization. This resulted in a feature pool of 48 predictive features for ERENO and 47 predictive features for 5G-NIDD. The evaluated retained feature subsets correspond to approximately 10%, 20%, 30%, 40%, and 50% of the available feature pool, yielding Top-5, Top-10, Top-15, Top-20, and Top-24 for ERENO (48 features), and Top-5, Top-10, Top-15, Top-19, and Top-24 for 5G-NIDD (47 features), where Top-19 reflects the 40% threshold rounded to the nearest integer.

The WOCI analysis focuses on two gradient-boosting classifiers: CatBoost and XGBoost, both widely used in tabular intrusion detection and well-suited as tree-based baselines for evaluating feature-ranking configurations under asymmetric costs.

Three strategies generated the evaluated configurations. Built-in importance [Wang et al. 2024] is extracted from fitted tree-based models as the weighted reduction in node impurity. SHAP [Lundberg and Lee 2017] rankings use mean absolute Shapley values from TreeExplainer. Mutual Information [Papaioannou et al. 2025] is computed independently of the classifier as a statistical filter using a non-parametric estimator. For each strategy, features are ordered by the corresponding criterion, and each Top-N configuration uses the N highest-ranked features. Thus, Top-N refers to method-specific ranked subsets, not a shared feature set across strategies.

Each configuration was evaluated using repeated stratified 5-fold cross-validation with 10 repetitions, resulting in 50 test evaluations per configuration. For each fold, the pipeline recorded the F1-score and the four confusion-matrix components: true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). These outputs were aggregated by dataset, classifier, feature-ranking method, and retained subset size to compute the WOCI scores under different cost-ratio scenarios.

6. Results and Discussion

6.1. WOCI Sensitivity to Cost Ratio

Figure 3 shows the WOCI sensitivity curves for 5G-NIDD. Since WOCI grows linearly with R, the slope of each curve is directly associated with $\overline{FN}/\overline{N}_{positive}$. Therefore, curves that increase faster indicate configurations that become more costly when missed intrusions are penalized.

In 5G-NIDD, Mutual Information is heavily penalized in compact subsets, especially at Top-5 and Top-15. As R increases, its WOCI grows much faster than Built-in and SHAP, reflecting a larger average false-negative count per positive instance under feature-constrained settings. One possible explanation is that Mutual Information, which evaluates each feature independently of the classifier, identifies features that are marginally informative in isolation but fail to capture the interaction structure needed for effective intrusion detection with very few features; however, confirming this mechanism would require a separate feature-level analysis.

With Top-24 features, the gap among methods decreases for both CatBoost and XGBoost, and Mutual Information becomes competitive under higher cost ratios. At

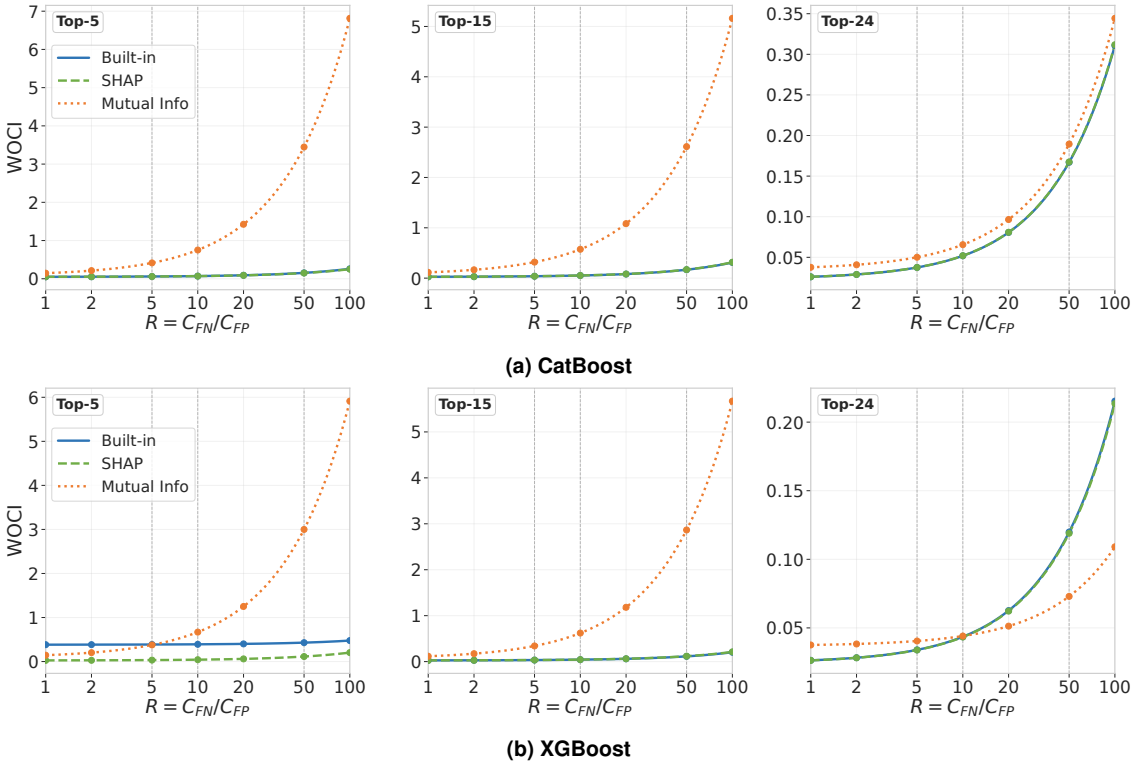

Figure 3. WOCI sensitivity curves for 5G-NIDD, comparing Built-in importance, SHAP, and Mutual Information across representative retained feature subsets. Note the different scale for Top-24.

Top-24, the absolute WOCI differences among methods are substantially smaller than in compact subsets, so the crossover analysis in Section 6.3 provides the most informative comparison.

Figure 4 shows a different behavior for ERENO. In this dataset, Mutual Information performs well with Top-5 features for both CatBoost and XGBoost, producing lower WOCI than the other methods. This suggests that, for ERENO, Mutual Information identifies a compact subset with a favorable FP/FN cost profile. At Top-5, the WOCI gap between Mutual Information and the other methods can exceed 4 units at high cost ratios, while at Top-15 and Top-24 the curves converge to differences below 1 unit, indicating that the advantage of MI diminishes substantially as the feature budget increases.

This advantage does not persist in larger subsets. With Top-15 and Top-24 features, Mutual Information becomes more costly as R increases, while Built-in importance and SHAP become more competitive. In several larger-subset settings, Built-in and SHAP show very similar curves, indicating comparable FP/FN cost profiles. For practitioners, these curves support a direct decision process: if the estimated enterprise cost ratio remains below a crossover point, the current winner is preferable; if it exceeds that point, an alternative configuration may reduce expected operational cost without retraining.

6.2. Winner Analysis Across Cost Scenarios

Figure 5 summarizes the evaluated configuration with the lowest WOCI for each combination of Top-N and cost ratio in 5G-NIDD. When two strategies produce very similar

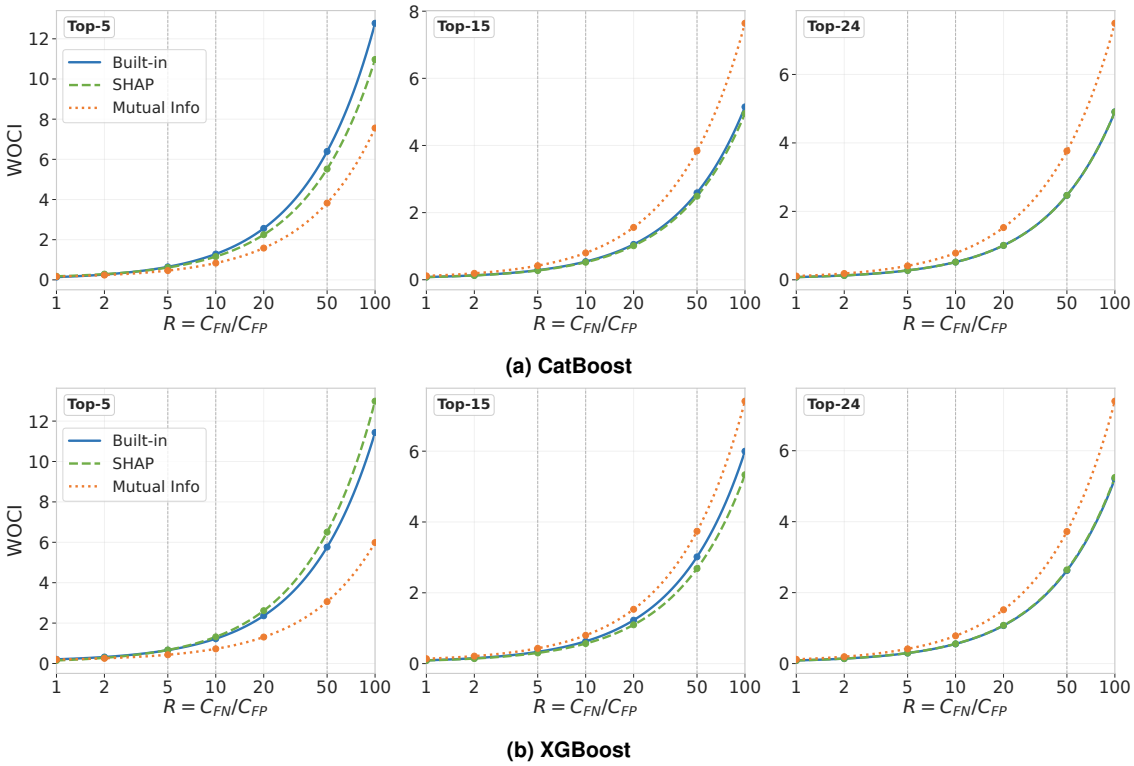

Figure 4. WOCI sensitivity curves for ERENO comparing Built-in, SHAP, and Mutual Information across representative feature subsets.

WOCI values, the crossover analysis in Section 6.3 provides a more informative comparison, as it identifies the exact R^* at which the ranking changes. Since the heatmaps report the numerically lowest WOCI without a tie threshold, configurations with very small differences should be interpreted together with the sensitivity curves and crossover analysis. The heatmap confirms that SHAP is the most competitive strategy under stronger feature reduction, winning consistently for Top-5 and Top-10 in both CatBoost and XGBoost. Built-in importance becomes more competitive in intermediate and larger subsets, especially from Top-15 onward.

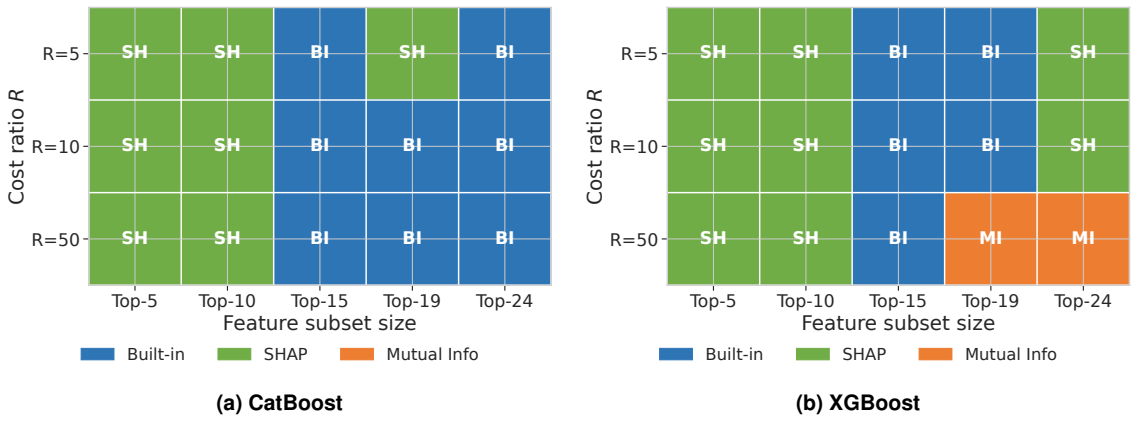

Figure 5. Winner heatmaps for 5G-NIDD, showing the evaluated configuration with the lowest WOCI for each combination of Top-N and cost ratio R.

The most relevant change in 5G-NIDD occurs for XGBoost with larger subsets. At Top-19 and Top-24, Mutual Information produces the lowest WOCI when $R = 50$, while SHAP or Built-in importance are preferable at lower cost ratios. In other words, an organization assuming a higher cost for missed intrusions would choose a different feature-ranking strategy than one operating under moderate cost assumptions.

Figure 6 shows a different pattern for ERENO. In this dataset, the winning method is the same for all evaluated cost ratios for each fixed model and Top-N. Mutual Information wins only when using the smallest feature subset, Top-5, for both CatBoost and XGBoost. For larger subsets, the best method is either Built-in importance and SHAP, depending on the classifier and feature subset size.

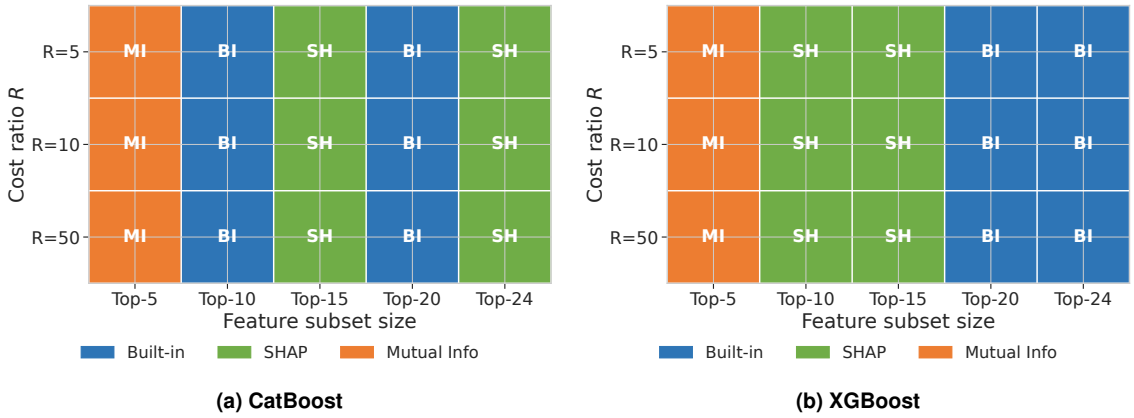

Figure 6. Winner heatmaps for ERENO, showing the evaluated configuration with the lowest WOCI for each combination of Top-N and cost ratio R.

This stability across R is informative. In ERENO, the competing methods produce similar average false-negative counts for most configurations, resulting in WOCI curves with comparable slopes that do not diverge as R increases. This contrasts with 5G-NIDD, where greater divergence in false-negative counts across strategies leads to curves with different slopes, making the ranking more sensitive to the cost ratio.

Overall, the winner heatmaps show that WOCI supports two types of decision. First, it identifies robust choices that remain preferable across cost scenarios, such as SHAP in compact 5G-NIDD subsets and Mutual Information in compact ERENO subsets. Second, it identifies cost-dependent choices, such as the emergence of Mutual Information in 5G-NIDD with XGBoost under $R = 50$. The crossover behavior between competing strategies is examined in detail in Section 6.3. This is useful for enterprise IDS deployment because the selected configuration can be aligned with the organization's tolerance for missed intrusions.

6.3. Crossover and Stability Analysis

The previous analyses showed which configuration minimizes WOCI for selected values of R. While the heatmaps in Section 6.2 already suggest stability at those discrete cost ratios, they do not reveal whether the ranking holds across the full continuous range $R \in [1, 100]$. For each dataset, model, and Top-N configuration, we compared the method with the lowest WOCI at low cost ratio ($R = 1$) and high cost ratio ($R = 100$). When

the winner changed, we estimated the crossover point R^* using Equation (8) at which the two WOCI curves intersect.

Table 2. Configurations in which the WOCI winner changes within $R \in [1, 100]$.

Dataset	Model	Top-N	Transition	$\mathbf{R}^*$
5G-NIDD	CatBoost	15	SH $\rightarrow$ BI	1.45
5G-NIDD	CatBoost	19	SH $\rightarrow$ BI	5.30
5G-NIDD	XGBoost	10	BI $\rightarrow$ SH	2.46
5G-NIDD	XGBoost	19	BI $\rightarrow$ MI	10.85
5G-NIDD	XGBoost	24	SH $\rightarrow$ MI	10.63
ERENO	CatBoost	5	BI $\rightarrow$ MI	1.50
ERENO	XGBoost	5	SH $\rightarrow$ MI	1.67

Table 2 reports only the configurations in which a crossover occurs within the evaluated range. Across the 20 evaluated dataset–model–Top-N configurations, 13 remain stable from $R = 1$ to $R = 100$, while seven present a change in the WOCI winner. This indicates that most choices are robust to the assumed cost ratio, but some configurations are sensitive to how strongly false negatives are penalized.

Most crossover cases occur in 5G-NIDD. For CatBoost, the winner changes from SHAP to Built-in at Top-15 and Top-19, with $R^* = 1.45$ and $R^* = 5.30$, respectively. For XGBoost, the most relevant changes occur at larger subsets: Built-in changes to Mutual Information at Top-19 with $R^* = 10.85$, and SHAP changes to Mutual Information at Top-24 with $R^* = 10.63$. These two cases are particularly important because the transition occurs around $R \approx 10$, which falls within the range of plausible operational scenarios discussed in Section 4, including regulated infrastructure deployments.

ERENO shows a more stable pattern. Only two crossovers are observed, both at Top-5 and both at very low values of R: Built-in changes to Mutual Information for CatBoost at $R^* = 1.50$, and SHAP changes to Mutual Information for XGBoost at $R^* = 1.67$. For the remaining ERENO configurations, the winner remains unchanged across the entire cost range. This suggests that, in ERENO, the choice of configuration is more strongly determined by the model and subset size than by the cost ratio itself.

Overall, the crossover analysis complements the sensitivity curves and heatmaps by distinguishing robust decisions from cost-dependent ones. Stable configurations indicate that the same method remains preferable under different operational assumptions. Crossover configurations, in contrast, identify cases where the preferred method depends on the organization's tolerance for false negatives. This is one of the main practical benefits of WOCI: it does not only rank evaluated configurations, but also shows when that ranking is sensitive to enterprise cost assumptions.

These results indicate that WOCI is most useful when predictive alternatives are close but their FP/FN profiles react differently to enterprise cost assumptions.

7. Limitations and Applicability

Although WOCI provides an interpretable operational view of IDS evaluation, some limitations must be considered. The first limitation concerns the decision threshold, that is, the probability value above which a classifier assigns the positive label to an instance.

In this study, WOCI is computed from confusion matrices obtained at a fixed decision threshold of 0.5, the default value applied by the classifiers during evaluation. This keeps the comparison consistent across configurations, but the conclusions remain threshold-specific. Calibrated or cost-sensitive thresholds could change the FP/FN profile of each configuration and, consequently, shift or eliminate some crossover points.

A second consideration concerns cost calibration. WOCI depends on the cost ratio $R = C_{FN}/C_{FP}$, which is an organization-specific parameter rather than a limitation of the index itself. The sensitivity range used in this study was chosen to represent plausible operational scenarios, and Section 4 discusses how R can be grounded in analyst effort, incident response costs, downtime, and regulatory mechanisms. In real deployments, organizations should calibrate R using their own incident response costs, service-level agreements, regulatory exposure, and business continuity requirements.

A third limitation is related to aggregation and class distribution. WOCI is computed from averaged confusion-matrix outcomes across cross-validation runs, which provides a compact and comparable summary of each configuration but may hide fold-level variability. In addition, when $\overline{N}_{positive}$ is very small, a small absolute change in $\overline{FN}$ produces a large change in WOCI because the denominator is small, especially at high values of R. The present evaluation focuses on ERENO and 5G-NIDD, which provide sufficiently large positive-class populations for the reported aggregate analysis. Future work should compute WOCI at the fold level, report confidence intervals, and apply paired statistical tests directly on WOCI distributions.

Finally, the evaluation is restricted to two IDS datasets, two gradient-boosting classifiers, and three feature-selection strategies used to generate the evaluated configurations. This design supports a focused analysis of WOCI, but it does not cover all IDS settings, classifiers, feature selectors, thresholds, preprocessing strategies, or deployment environments. Additional datasets, including cyber-physical, IoT, enterprise, and cloud traffic scenarios, are needed to assess how broadly the observed patterns generalize.

8. Conclusion

This paper introduced the *Weighted Operational Cost Index* (WOCI), a post-hoc cost-sensitive index for binary detection tasks. WOCI combines averaged counts of false positives and false negatives with a configurable cost ratio $R = C_{FN}/C_{FP}$, then normalizes the result by the number of positive instances. The resulting score can be compared across already evaluated binary detection configurations.

In the IDS case study, WOCI was instantiated to compare configurations generated by Built-in importance, SHAP, and Mutual Information on the ERENO and 5G-NIDD datasets. A key finding is that 13 out of 20 evaluated configurations maintain the same preferred method across the entire cost range $R \in [1, 100]$, confirming that their choice is robust to enterprise cost uncertainty. The other seven configurations, mainly in 5G-NIDD, transition at $R^* \approx 10$, a range consistent with realistic security cost asymmetry.

WOCI is most informative when evaluated configurations have comparable predictive performance but different FP/FN trade-offs under changing cost assumptions. When no crossover exists in a practical R range, the same configuration remains preferable across cost scenarios. When a crossover exists at a plausible value, WOCI reveals a decision boundary that conventional predictive metrics cannot capture.

Acknowledgements

This work was carried out with the support of CAPES, CNPq, FAPERJ, and the United States Department of Energy, under Grant Number DE-CR0000039.

References

ANEEL (2025). Estabelecida a receita anual permitida (RAP) de transmissoras para o ciclo 2025–2026. gov.br/aneel/noticias/2025/rap-transmissoras-2025-2026. Accessed: 2026-04-27.

ANEEL (2026). Receita anual permitida de transmissão. https://portalrelatorios.aneel.gov.br/luznatarifa/relrap. Accessed: 2026-04-27.

Cha, J., Jang, J., Shin, D., and Shin, D. (2026). Cost-sensitive threshold optimization for network intrusion detection: A per-class approach with xgboost. *Electronics*, 15(7):1542.

Chen, Z., Xia, W., Li, Z., Xiong, G., and Gou, G. (2024). RecoSelector: Cost-Sensitive Feature Selection for Network Intrusion Detection in Resource-Constrained Internet of Things. In *International Performance, Computing, and Communications Conference*, pages 1–10. IEEE.

Drummond, C. and Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine learning*, 65(1):95–130.

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978.

Fatima, M., Rehman, O., Rahman, I. M., Ajmal, A., and Park, S. J. (2024). Towards ensemble feature selection for lightweight intrusion detection in resource-constrained IoT devices. *Future Internet*, 16(10):368.

Gombar, M., Topalović, A., and Pejić Bach, M. (2026). Cost-Aware Lightweight Deep Learning for Intrusion Detection: A Comparative Study on UNSW-NB15 and CIC-IDS2017. *Electronics*, 15(8):1603.

Gupta, N., Jindal, V., and Bedi, P. (2022). CSE-IDS: Using cost-sensitive deep learning and ensemble algorithms to handle class imbalance in network-based intrusion detection systems. *Computers & Security*, 112:102499.

Hozouri, A., Mirzaei, A., and Effatparvar, M. (2025). A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. *Discover Artificial Intelligence*, 5(1):314.

IBM (2025). Cost of a data breach report 2025. https://www.ibm.com/reports/data-breach. Accessed: 2026-04-27.

Lee, W., Fan, W., Miller, M., Stolfo, S. J., and Zadok, E. (2002). Toward cost-sensitive modeling for intrusion detection and response. *Journal of computer security*, 10(1-2):5–22.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Nelson, A., Rekhi, S., Souppaya, M., and Scarfone, K. (2025). Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile. Technical Report NIST Special Publication 800-61 Revision 3, National Institute of Standards and Technology.

Operador Nacional do Sistema Elétrico (2019). Submódulo 15.12: Apuração mensal das parcelas variáveis referentes à disponibilidade de instalações da Rede Básica e das Interligações Internacionais. `ons.org.br/procedimentos-de-rede/submodulo-15.12`. Accessed: 2026-04-27.

Papaioannou, N., Myllis, G., Tsimpiris, A., and Vrana, V. (2025). The Role of Mutual Information Estimator Choice in Feature Selection: An Empirical Study on mRMR. *Information*, 16(9):724.

Quincozes, S. E., Albuquerque, C., Passos, D., and Mossé, D. (2023). ERENO: A Framework for Generating Realistic IEC–61850 Intrusion Detection Datasets for Smart Grids. *IEEE Transactions on Dependable and Secure Computing*, 21(4):3851–3865.

Siriwardhana, Y., Samarakoon, S., Porambage, P., Liyanage, M., Chang, S.-Y., Kim, J., Kim, J., and Ylianttila, M. (2025). Descriptor: 5G Wireless Network Intrusion Detection Dataset (5G-NIDD). *IEEE Data Descriptions*.

Skrodelis, H. K. and Romanovs, A. (2026). Explainable Hybrid Intrusion Detection for SCADA/ICS: A Review and Research Agenda. *Frontiers in Computer Science*.

Stolfo, J., Fan, W., Lee, W., Prodromidis, A., and Chan, P. K. (2000). Cost-based modeling and evaluation for data mining with application to fraud and intrusion detection. *Results from the JAM Project by Salvatore*, pages 1–15.

Telikani, A. and Gandomi, A. H. (2021). Cost-sensitive stacked auto-encoders for intrusion detection in the internet of things. *Internet of Things*, 14:100122.

U.S. Bureau of Labor Statistics (2025). Information Security Analysts: Occupational Outlook Handbook. `bls.gov/ooh/information-security-analysts`. Accessed: 2026-04-27.

Wang, H., Liang, Q., Hancock, J. T., and Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of shap-value and importance-based methods. *Journal of Big Data*, 11(1):44.

Westphal, C., Hailes, S., and Musolesi, M. (2025). Feature selection for network intrusion detection. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, page 1599–1610, New York, NY, USA.