

Article

Numerically Stabilized Regularized Learning for Intrusion Detection: Conditioning, Scaling, and Cross-Dataset Transfer Analysis

Miguel Arcos-Argudo ^{*,†} , Rodolfo Bojorque [†]  and Mauricio Ortiz [†] 

Department of Advanced Computing and Data Research Group, Universidad Politécnica Salesiana, Campus El Vecino, Turuhuayco Ave. 3-69, Cuenca 010102, Ecuador; rbojorque@ups.edu.ec (R.B.); mortizo@ups.edu.ec (M.O.)

* Correspondence: marcos@ups.edu.ec

† These authors contributed equally to this work.

Abstract

This paper presents a numerical-computational analysis of ℓ_2 -regularized logistic learning for binary intrusion detection under heterogeneous datasets, class imbalance, and cross-dataset shift. Rather than proposing a new intrusion detection architecture, the study examines how numerical conditioning, feature scaling, feature set design, threshold selection, false negative behavior, false alarm behavior, and distribution shift affect operational detection performance. Experiments were conducted on CICIDS2017, UNSW-NB15, and CIRA-CIC-DoHBrw-2020 using reproducible train-validation-test protocols over five fixed random seeds. The numerical audit showed that standard scaling reduced the spectral condition number of traffic feature matrices by several orders of magnitude across datasets and feature configurations. However, scaling did not produce uniformly monotonic predictive gains: in some cases, raw feature optimization achieved comparable or higher F1-score, whereas scaled preprocessing produced more controlled false alarm behavior. In-domain experiments showed that dataset-specific features may improve ranking metrics such as area under the receiver-operating-characteristic curve (AUROC) or area under the precision-recall curve (AUPR) without necessarily improving thresholded operational metrics. Cross-dataset transfer experiments revealed strong source-target asymmetry, with transferred thresholds producing either near-zero positive detection or excessive false alarms. Additional robustness experiments with Random Forest and XGBoost improved in-domain F1-score and false negative rate (FNR), but did not eliminate off-domain degradation, with high FNR persisting under direct cross-dataset transfer. Finally, a Kolmogorov–Smirnov-based distribution shift analysis showed that in-domain discrepancies were small, whereas cross-dataset discrepancies were consistently large under common standardized traffic features. These findings suggest that numerical stability, ranking quality, thresholded detection performance, false negative and false alarm behavior, and distribution shift should be analyzed jointly when evaluating intrusion detection models.



Academic Editors: Juan Ramón Torregrosa Sánchez and Theodore E. Simos

Received: 17 June 2026

Revised: 9 July 2026

Accepted: 22 July 2026

Published: 25 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: numerical conditioning; regularized learning; logistic regression; intrusion detection; network security; class imbalance; false alarm rate; cross-dataset transfer; dataset shift; Kolmogorov–Smirnov statistic

MSC: 65Z05

1. Introduction

Network intrusion detection systems (IDSs) are commonly evaluated using benchmark datasets that contain heterogeneous traffic patterns, class imbalance, and diverse attack behaviors. Public datasets such as CICIDS2017, UNSW-NB15, and CIRA-CIC-DoHBrw-2020 have enabled reproducible experimental comparisons in intrusion detection research [1–3]. However, the use of such datasets also introduces methodological challenges. Traffic features often combine heterogeneous physical units, including durations, byte counts, rates, packet counts, and port-related variables. In addition, class distributions are frequently skewed, and models trained on one dataset may not preserve their operating behavior when transferred to another dataset. The datasets used in this study were obtained from their official public repositories to ensure traceability and reproducibility. CICIDS2017 was downloaded from the Canadian Institute for Cybersecurity repository at the University of New Brunswick (<https://www.unb.ca/cic/datasets/ids-2017.html>, accessed on 12 November 2025); UNSW-NB15 was obtained from the official University of New South Wales dataset page (<https://research.unsw.edu.au/projects/unsw-nb15-dataset>, accessed on 12 November 2025); and CIRA-CIC-DoHBrw-2020, from which the DoH-Combined subsets were derived, was downloaded from the Canadian Institute for Cybersecurity repository (<https://www.unb.ca/cic/datasets/dohbrw-2020.html>, accessed on 16 November 2025).

Most IDS studies emphasize predictive performance, typically through accuracy, F1-score, AUROC, or AUPR. These metrics are necessary but not sufficient to understand whether a model is numerically stable, operationally useful, and transferable across datasets. In particular, accuracy can be misleading under class imbalance, while AUROC and AUPR describe ranking behavior but do not fully determine the quality of a deployed binary decision rule [4–6]. For operational intrusion detection, the false alarm rate (FAR) is especially important because excessive false positives can make a detector impractical even when recall is high.

A second limitation is that preprocessing is often treated as a routine step rather than as a numerical component of the learning pipeline. Feature scaling is commonly applied before linear or distance-based learning algorithms, but its effect is rarely quantified in IDS studies through explicit conditioning analysis. From a numerical linear algebra perspective, ill-conditioned feature matrices can increase the sensitivity of optimization and estimation procedures to perturbations in the data [7]. Therefore, analyzing the condition number of traffic feature matrices can provide a more rigorous justification for preprocessing decisions such as imputation and standard scaling.

A third challenge concerns cross-dataset generalization. A detector trained and validated on a single dataset may not preserve its decision behavior under distribution shift. Dataset shift occurs when the data-generating distribution differs between training and deployment conditions [8]. In domain adaptation theory, performance on a target domain depends not only on source domain performance but also on the discrepancy between source and target distributions [9]. This issue is especially relevant for IDS evaluation because traffic-generation processes, feature extraction pipelines, attack taxonomies, and label semantics may differ substantially across datasets.

This paper addresses these issues through a numerical-computational analysis of ℓ_2 -regularized logistic learning for binary intrusion detection. The objective is not to introduce a new IDS architecture. Instead, the study uses a simple convex model as an interpretable probe to examine how numerical conditioning, scaling, feature set design, class imbalance, threshold selection, and cross-dataset transfer interact across heterogeneous IDS datasets. The selected data sources were chosen to cover complementary IDS conditions, including large-scale flow-based attack traffic, a distinct modern network testbed with different

feature distributions, and encrypted Domain Name System over HTTPS (DNS-over-HTTPS, DoH) traffic with protocol-specific label semantics.

Unlike prior IDS benchmarking studies that primarily emphasize predictive comparison among classifiers, this work focuses on the numerical and operational mechanisms underlying regularized intrusion detection. The novelty of the study lies in the joint analysis of dimensions that are usually examined separately: feature matrix conditioning, standard scaling, ℓ_2 -regularized optimization, class imbalance, validation-based threshold selection, false alarm behavior, cross-dataset transfer degradation, and feature-level distribution shift. Therefore, the contribution is not a new IDS architecture, but a numerical-computational framework for understanding when and why regularized IDS models remain stable, fail operationally, or degrade under heterogeneous source–target conditions.

This joint perspective differs from three common lines of previous work. First, conventional IDS benchmarking studies mainly compare classifiers using in-domain predictive metrics, often without explicitly analyzing the numerical conditioning of the input matrices or the stability of thresholded decisions. Second, preprocessing-oriented studies commonly evaluate whether scaling or feature engineering improves accuracy, but they rarely connect these transformations to condition number reduction, operational false alarms, and threshold behavior. Third, cross-dataset transfer studies usually report performance degradation across source–target pairs, but they do not always link transfer outcomes to numerical preprocessing, class prior changes, validation threshold transfer, and feature-level distribution shift within the same experimental protocol.

This positioning also defines an important limitation and, at the same time, the intended methodological contribution of the study. The individual components used in the framework—standard scaling, ℓ_2 -regularized logistic regression, validation-based threshold selection, and KS-based distribution shift analysis—are established techniques. The manuscript does not claim novelty for these components in isolation, nor does it claim superiority over nonlinear classifiers, ensemble methods, or deep intrusion detection architectures. Instead, the contribution lies in combining these established elements into a single controlled and reproducible numerical-computational protocol that links matrix conditioning, preprocessing, regularized optimization, class imbalance, thresholded IDS decisions, false negative and false alarm behavior, cross-dataset transfer degradation, and feature-level distribution shift. Consequently, the novelty of the work lies in the integrated analytical framework and the evidence it provides about the interaction among these factors, rather than in the development of a new learning algorithm.

The paper makes the following contributions:

- It introduces a reproducible experimental pipeline for binary IDS evaluation across CICIDS2017, UNSW-NB15, and CIRA-CIC-DoHBrw-2020, using five fixed random seeds and explicit train–validation–test split validation.
- It quantifies the numerical conditioning of raw and scaled feature matrices, showing that standard scaling reduces condition numbers by several orders of magnitude across datasets and feature configurations.
- It evaluates ℓ_2 -regularized logistic regression under common and dataset-specific feature sets, demonstrating that richer feature sets may improve ranking metrics while degrading thresholded operational behavior.
- It performs a scaling ablation showing that improved conditioning does not always imply monotonic gains in F1-score, but may reduce false alarm instability in highly imbalanced settings.
- It analyzes cross-dataset transfer using common standardized traffic features, revealing strong source–target asymmetry and threshold transfer failures.

- It quantifies feature-level distribution shift using a bounded Kolmogorov–Smirnov-based score and relates distributional mismatch to transfer degradation.
- It supplements the regularized logistic analysis with model family robustness baselines based on Random Forest and XGBoost, showing that improved in-domain performance does not eliminate cross-dataset transfer degradation.

The remainder of the paper is organized as follows. Section 2 reviews related work on IDS datasets, traffic analysis and Internet of Things (IoT) and Industrial Internet of Things (IIoT) detection methods, imbalanced evaluation, numerical conditioning, regularized linear learning, and dataset shift. Section 3 describes the datasets, preprocessing, splitting protocol, regularized model, threshold selection, transfer protocol, and distribution shift analysis. Section 4 presents the numerical and predictive results. Section 5 discusses implications, limitations, and future work. Finally, Section 6 summarizes the main findings.

2. Related Work

2.1. Benchmark Datasets for Intrusion Detection

Benchmark datasets are essential for reproducible intrusion detection research. CICIDS2017 was designed to include benign traffic and several attack families, addressing limitations of older IDS datasets such as outdated traffic patterns and insufficient attack diversity [1]. UNSW-NB15 was generated using the IXIA PerfectStorm tool and contains both normal and attack traffic with a mixture of flow and content features [2]. CIRA-CIC-DoHBrw-2020 focuses on encrypted DNS-over-HTTPS traffic and supports layered detection tasks, including DoH identification and malicious DoH detection [3].

These datasets differ not only in size and class balance but also in feature definitions, traffic-generation procedures, and label semantics. Such heterogeneity is useful for testing robustness, but it complicates direct cross-dataset comparison. For this reason, the present study separates in-domain experiments from cross-dataset transfer experiments and restricts transfer analysis to features that are available and semantically comparable across datasets.

2.2. Traffic Analysis and IoT/IIoT Detection Methods

Traffic analysis for intrusion detection has increasingly moved from shallow classifiers over hand-crafted flow statistics toward deep and representation learning models capable of modeling complex feature interactions. FlowTransformer provides a modular transformer framework for flow-based network intrusion detection systems, emphasizing architectural choices, model size, inference speed, and accuracy across public benchmark datasets [10]. Similarly, IDS-INT uses transformer-based transfer learning, Synthetic Minority Over-sampling Technique (SMOTE)-based balancing, and convolutional neural network–long short-term memory (CNN–LSTM) detection to handle imbalanced network traffic across UNSW-NB15, CICIDS2017, and NSL-KDD [11]. These studies illustrate an important direction in traffic analysis research: improving detection performance by increasing model expressiveness and learning richer network traffic representations.

In IoT and IIoT environments, recent studies have also focused on rare-attack detection under severe class imbalance. S2CGAN-IDS expands minority IoT attack classes in both data and feature spaces to improve detection under highly imbalanced conditions [12]. TMG-GAN uses a multigenerator generative adversarial network (GAN) architecture to generate diverse attack samples and mitigate imbalanced learning in network intrusion detection [13]. More recently, HybridGF was proposed for extremely imbalanced malicious traffic detection in IIoT; it combines hybrid sample generation, boundary filtering, and density-regulation modules to improve rare-class detection on NSL-KDD, IoT-23, and TON-IoT [14].

The present study is complementary to these detection-oriented and data-augmentation approaches. It does not propose a new traffic detector, deep architecture, or synthetic-sample generation algorithm. Instead, it uses ℓ_2 -regularized logistic regression as a controlled convex probe to analyze numerical conditioning, scaling, validation threshold selection, false alarm behavior, and direct cross-dataset transfer. In this sense, recent traffic analysis methods provide stronger predictive architectures, whereas the present work contributes a numerical-computational diagnostic framework for understanding how preprocessing and transfer decisions affect operational IDS behavior.

2.3. Evaluation Under Class Imbalance

Class imbalance is a central issue in intrusion detection because benign and malicious samples are often unevenly represented. Imbalanced learning has been extensively studied, and prior work has shown that conventional accuracy may provide an incomplete view of performance when one class dominates the dataset [15]. This is particularly relevant in IDS settings, where false negatives and false positives have different operational consequences.

Recent traffic analysis studies address imbalance through several mechanisms. Data-level methods synthesize or resample minority traffic, algorithm-level methods modify the loss or weighting scheme, and generative approaches use adversarial models to produce additional minority samples. HybridGF is representative of a recent IIoT-oriented line of work: it identifies extreme imbalance as a central obstacle and controls generated-sample boundaries through a filtering stage [14]. In contrast, this study deliberately avoids changing the empirical training distribution through sample generation. Class imbalance is handled through class-balanced logistic weights and validation-based threshold selection so that the numerical effect of scaling and the operational effect of threshold transfer can be analyzed without the confounding influence of synthetic traffic.

ROC analysis is widely used for classifier evaluation and provides a threshold-independent view of the tradeoff between true positive and false positive rates [5]. However, precision–recall analysis is often more informative in imbalanced binary classification because it directly reflects the behavior of the positive class under skewed class distributions [4,6]. For this reason, the present study reports AUROC and AUPR together with thresholded metrics such as F1-score, balanced accuracy, recall, precision, and FAR.

2.4. Numerical Conditioning and Feature Scaling

Feature scaling is standard in many machine learning workflows, especially for linear models and optimization-based algorithms. Nevertheless, its numerical effect is often not explicitly reported in IDS studies. From a numerical analysis perspective, the condition number of a matrix measures the sensitivity of computations to perturbations and finite-precision effects [7]. A large condition number indicates that small changes in the input may be amplified in downstream calculations.

IDS feature matrices can be naturally ill-conditioned because they combine features with very different scales and units. Examples include duration, packet counts, byte counts, rates, and port numbers. In such settings, standardization can improve numerical conditioning by centering and rescaling each feature. The present work explicitly measures this effect through the spectral condition number before and after scaling.

2.5. Regularized Linear Models

Regularized logistic regression remains a useful baseline for binary classification because it is convex, interpretable, and computationally efficient. Large-scale linear classification libraries such as LIBLINEAR support logistic regression and linear support vector machines and have demonstrated the practical relevance of regularized linear models [16]. In this study, logistic regression is not used as a state-of-the-art IDS model, but as a con-

trolled mathematical probe for analyzing the interaction between feature conditioning, regularization, imbalance-aware weighting, and threshold selection.

The use of ℓ_2 regularization is especially appropriate in this context because it constrains the coefficient norm and improves the stability of the fitted linear decision function. This allows the analysis to focus on numerical and operational effects without introducing the additional complexity of nonlinear architectures.

2.6. Dataset Shift and Cross-Dataset Transfer

Dataset shift arises when the training and test distributions differ. The problem has been formalized in machine learning as a mismatch between the joint distributions of inputs and outputs across stages or domains [8]. Domain adaptation theory further shows that target domain performance depends on both source domain performance and the discrepancy between source and target domains [9]. These ideas are directly relevant to IDS evaluation because a model trained on one traffic corpus may encounter substantially different feature distributions and label priors in another.

The present study evaluates direct cross-dataset transfer without target domain recalibration. This is a deliberately strict setting: the source-fitted imputer, scaler, model coefficients, and decision threshold are applied unchanged to the target dataset. This protocol makes it possible to isolate threshold transfer failures and operational degradation under distribution shift.

Recent cross-dataset IDS studies have shown that models trained on one traffic dataset may generalize poorly to unseen datasets, even when in-domain performance is high. However, these studies usually emphasize predictive degradation across classifiers, whereas the present work focuses on the numerical and operational factors behind such degradation, including conditioning, scaling, transferred thresholds, FAR variation, and KS-based feature shift.

2.7. Calibration and Threshold Transfer

Binary classifiers produce scores or probabilities, but the conversion of scores into decisions depends on a threshold. Under distribution shift, a threshold selected on the source validation set may not remain appropriate for the target domain. Calibration research has shown that predicted probabilities may become unreliable, and that confidence estimates can degrade under shifted data conditions [17,18]. Although the present study does not perform post-transfer recalibration, it explicitly evaluates how source-selected thresholds behave when transferred to a different dataset.

This threshold-centered perspective is important for IDS deployment. A transferred model may retain partial ranking ability while failing operationally because its decision threshold produces either almost no positive detections or an excessive number of false alarms. Therefore, cross-dataset evaluation should report both ranking metrics and thresholded operational metrics.

2.8. Positioning of the Present Study

The literature reviewed above shows that IDS research has advanced along several complementary directions, including benchmark dataset construction, traffic analysis architectures, imbalance-aware learning, numerical preprocessing, and cross-dataset evaluation. However, these dimensions are often treated as separate methodological concerns. Benchmarking studies typically focus on classifier comparison; preprocessing studies often emphasize predictive gains; and transfer studies mainly report source–target degradation.

The present work is positioned differently. It uses a simple and interpretable regularized model not as a proposed state-of-the-art detector, but as a diagnostic instrument for jointly examining numerical stability and operational IDS behavior. In particular, the study

connects condition number analysis, standard scaling, feature set design, validation-based threshold selection, false alarm rate, cross-dataset transfer, and KS-based distribution shift within a single reproducible protocol. This joint design allows the paper to identify situations in which numerical improvement, ranking quality, thresholded F1-score, and operational false alarm behavior do not necessarily move in the same direction.

Accordingly, the main contribution is methodological rather than algorithmic. The paper does not introduce a new preprocessing method, classifier, thresholding rule, or distribution shift metric. Instead, it provides a controlled numerical-computational framework for evaluating how these established components interact when IDS models are trained, thresholded, and transferred across heterogeneous datasets. This complements, rather than replaces, recent work on advanced IDS architectures, data-augmentation methods, and cross-dataset benchmarking.

3. Materials and Methods

3.1. Study Design and Scope

This study evaluates numerically stabilized regularized learning for binary network traffic detection under class imbalance and dataset shift. The central objective is not to propose a new intrusion detection architecture, but to analyze how numerical conditioning, feature scaling, ℓ_2 -regularized optimization, threshold selection, and cross-dataset transfer interact in heterogeneous intrusion detection datasets.

The experimental design focuses exclusively on binary targets. For CICIDS2017 and UNSW-NB15, the negative class corresponds to benign or normal traffic and the positive class corresponds to attack traffic. For CIRA-CIC-DoHBrw-2020 Layer 2, the negative class corresponds to benign DoH traffic and the positive class to malicious DoH traffic. CIRA-CIC-DoHBrw-2020 Layer 1 was retained only for in-domain binary experiments, since its task is protocol-level discrimination, i.e., DoH versus non-DoH, rather than benign-versus-attack detection. Consequently, Layer 1 was excluded from cross-dataset transfer experiments to avoid mixing incompatible label semantics.

CIRA-CIC-DoHBrw-2020 Layer 3 was not included because the present study was restricted to binary detection tasks with validated label semantics. Layer 3 introduces a finer-grained label structure that would require a separate multiclass or hierarchical evaluation protocol, which is beyond the scope of this work.

Figure 1 summarizes the overall numerical-computational workflow followed in this study.

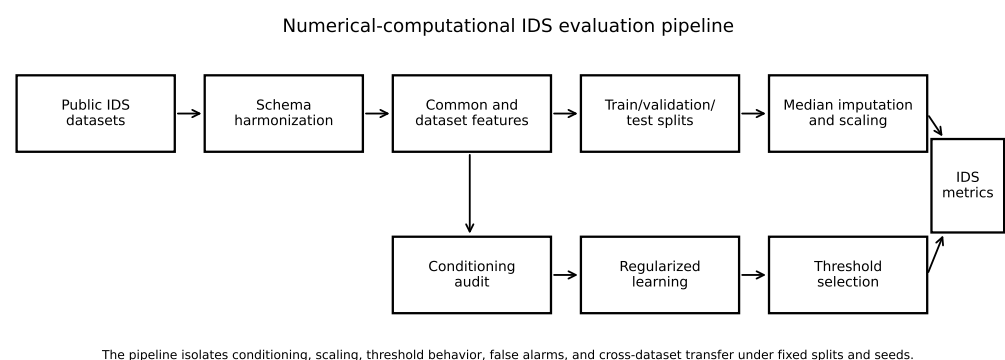


Figure 1. Overview of the numerical-computational IDS evaluation framework used in this study. The workflow emphasizes dataset harmonization, feature set construction, reproducible splits, numerical preprocessing, conditioning analysis, regularized learning, threshold selection, and operational IDS metrics.

As shown in Figure 1, the analysis was designed to separate numerical effects from operational evaluation effects. The conditioning audit and scaling analysis quantify the numerical properties of the feature matrices, whereas the validation-based threshold selection stage links probabilistic scores to binary IDS decisions. This structure supports the main objective of the paper: to analyze not only whether a model performs well in-domain, but also whether its thresholded behavior remains stable under heterogeneous source–target conditions.

3.2. Datasets

Four processed binary datasets were considered: CICIDS2017, UNSW-NB15, CIRA-CIC-DoHBrw-2020 Layer 1, and CIRA-CIC-DoHBrw-2020 Layer 2. CICIDS2017 is a modern intrusion detection dataset generated to include benign traffic and multiple attack families [1]. UNSW-NB15 is a widely used network intrusion detection benchmark containing normal and attack traffic generated using the IXIA PerfectStorm tool [2]. CIRA-CIC-DoHBrw-2020 focuses on DNS-over-HTTPS traffic and includes both protocol-level and malicious DoH detection scenarios [3].

Table 1 summarizes the processed datasets used in the experiments. The binary class imbalance is substantial in several cases, particularly in DoH-Combined-L2, where malicious DoH traffic accounts for more than 92% of the samples. This imbalance motivated the use of balanced accuracy, AUPR, F1-score, and false alarm rate in addition to standard accuracy.

Table 1. Processed dataset summary and binary class distribution.

Dataset	Rows	Columns	Negative Class	Positive Class	Positive Rate
CICIDS2017	2,830,743	18	2,273,097	557,646	0.1970
UNSW-NB15	257,673	18	93,000	164,673	0.6391
DoH-Combined-L1	1,167,136	18	897,493	269,643	0.2310
DoH-Combined-L2	269,643	18	19,807	249,836	0.9265

3.3. Dataset Selection Rationale

The three public data sources were selected to provide complementary coverage of modern intrusion detection challenges rather than to represent a single homogeneous traffic environment. CICIDS2017 was included because it provides large-scale benign and malicious traffic with multiple attack families and flow-level characteristics commonly used in IDS benchmarking [1]. UNSW-NB15 was included because it was generated using a different traffic-generation infrastructure and contains a distinct mixture of normal and attack traffic, providing a complementary benchmark with different feature distributions and label priors [2]. CIRA-CIC-DoHBrw-2020 was included because it represents a more recent encrypted-traffic scenario centered on DNS-over-HTTPS, where traffic semantics differ from conventional benign-versus-attack flow datasets [3].

Together, these datasets cover three relevant dimensions of IDS evaluation: heterogeneous network environments, different attack and label semantics, and substantial variation in class imbalance and feature distributions. This diversity is important for the objectives of the present study because the paper analyzes numerical conditioning, feature scaling, operational threshold behavior, and cross-dataset transfer rather than only in-domain predictive accuracy. Table 2 summarizes the rationale for selecting these datasets.

This selection is therefore appropriate for evaluating whether a numerically stabilized regularized model preserves its operational behavior across heterogeneous IDS settings. At the same time, the datasets are not claimed to exhaustively cover all possible deployment scenarios, such as wireless IoT telemetry, industrial control protocols, or real-time

streaming environments. Rather, they provide complementary and widely used benchmark conditions for testing the interaction between conditioning, scaling, class imbalance, threshold selection, and dataset shift.

Table 2. Rationale for dataset selection and coverage of IDS evaluation challenges.

Data Source	Network Environment	Detection Scope	Main Challenge Represented
CICIDS2017	Enterprise-like traffic with benign and attack flows	Binary benign-vs-attack detection derived from multiple attack families	Large-scale traffic, heterogeneous flow features, and strong feature-scale variation
UNSW-NB15	Synthetic modern network testbed with normal and attack traffic	Binary normal-vs-attack detection with different feature distributions and priors	Cross-dataset heterogeneity and distribution mismatch relative to CICIDS2017
CIRA-CIC-DoHBrw-2020	Encrypted DNS-over-HTTPS traffic	Layer 1 protocol detection and Layer 2 benign-vs-malicious DoH detection	Encrypted-traffic semantics, protocol-specific behavior, and severe class imbalance

3.4. Feature Sets

Two feature configurations were evaluated. The first one, denoted as common features, contains only variables that are available and reasonably comparable across all processed datasets:

$$\mathcal{X}_{common} = \{\text{duration}, \text{bytes_fwd}, \text{bytes_bwd}, \text{bps}\}.$$

These features were used in all in-domain experiments and were the only features used for cross-dataset transfer analysis.

The second configuration, denoted as dataset features, uses all available numerical variables that were meaningful within each dataset after schema harmonization. These features were used only for in-domain comparison, because variables such as ports, packet counts, packets per second, and protocol-specific fields are not consistently available across all datasets.

The feature-availability audit showed that the four common features were present in all datasets with negligible missingness, except for bps in UNSW-NB15, which had a missing rate of approximately 1.4%. By contrast, several candidate features were structurally absent from some datasets: `src_port` was absent from CICIDS2017 and UNSW-NB15, `dst_port` was absent from UNSW-NB15, and packet count features were absent from the DoH datasets. This justified the separation between in-domain feature sets and the common cross-dataset feature set.

3.5. Splitting Protocol and Reproducibility

All experiments were repeated over five fixed random seeds:

$$S = \{0, 42, 1337, 2026, 9999\}.$$

For CICIDS2017 and the DoH datasets, stratified train–validation–test splits were generated with the proportions 70%, 15%, and 15%, respectively. For UNSW-NB15, the official train–test split was preserved; the validation set was obtained by stratified sampling from the official training partition. Split-integrity checks confirmed zero overlap between train, validation, and test identifiers for every dataset and seed.

All preprocessing transformations were fitted exclusively on the training partition. For cross-dataset transfer, the imputer and scaler were fitted on the source training set and then applied unchanged to the target test set. This prevents information leakage and reflects a realistic transfer setting.

3.6. Numerical Preprocessing and Conditioning Audit

Missing numerical values and non-finite values were handled through median imputation. The default numerically stabilized preprocessing pipeline consisted of median imputation followed by standard scaling:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j},$$

where μ_j and σ_j were estimated only from the corresponding training split.

The effect of scaling was quantified using the spectral condition number of the design matrix,

$$\kappa_2(X) = \frac{\sigma_{\max}(X)}{\sigma_{\min}(X)},$$

where $\sigma_{\max}(X)$ and $\sigma_{\min}(X)$ denote the largest and smallest nonzero singular values of X . The condition number is a standard indicator of sensitivity in numerical linear algebra and is closely related to the stability of least-squares and optimization computations [7]. For each dataset, seed, and feature set, condition numbers were computed for the imputed raw matrix and for the imputed-scaled matrix.

3.7. Regularized Logistic Learning

The main predictive model was ℓ_2 -regularized logistic regression. Given training pairs $\{(x_i, y_i)\}_{i=1}^n$, with $y_i \in \{0, 1\}$, the model estimates a linear score

$$s_i = \beta^\top x_i + b,$$

and posterior probability

$$p_i = \frac{1}{1 + \exp(-s_i)}.$$

The optimized objective can be written as

$$\min_{\beta, b} \frac{1}{n} \sum_{i=1}^n w_{y_i} [-y_i \log(p_i) - (1 - y_i) \log(1 - p_i)] + \frac{\lambda}{2} \|\beta\|_2^2,$$

where w_{y_i} denotes class-balanced weights and λ controls the strength of regularization. The inverse regularization parameter $C = 1/\lambda$ was selected from

$$C \in \{0.01, 0.1, 1.0, 10.0\}.$$

Regularized logistic regression was selected because it is convex, interpretable, computationally efficient, and directly connected to numerical conditioning and feature scaling [16,19].

All models were implemented in Python 3.11.9 using scikit-learn [20]. The solver was lbfgs, the maximum number of iterations was set sufficiently high for convergence in all experiments, with 1500 iterations used in the ablation and transfer analyses, and class weights were balanced automatically from the training class frequencies.

3.8. Hyperparameter and Threshold Selection

For each dataset, seed, feature set, and preprocessing configuration, the regularization parameter C was selected on the validation set. Since class imbalance affects decision quality, the decision threshold was not fixed at 0.5. Instead, for each candidate C , thresholds were searched over

$$\tau \in \{0.01, 0.02, \dots, 0.99\}.$$

The threshold selection rule was designed to reflect both class imbalance and operational intrusion detection requirements. The primary criterion was validation balanced accuracy because it gives equal weight to positive class detection and negative class rejection. Equivalently, balanced accuracy jointly accounts for missed detections and false alarms, since sensitivity is $1 - \text{FNR}$ and specificity is $1 - \text{FAR}$. This makes it more appropriate than standard accuracy when class priors differ substantially across datasets.

F1-score was used as the second criterion because it emphasizes positive class detection quality and penalizes thresholds that achieve high specificity by detecting very few attacks. This is important in security applications, where a low false alarm rate is insufficient if the detector misses a large proportion of positive samples. FAR was then used as a third criterion to select, among thresholds with comparable balanced accuracy and F1-score, the operating point with fewer false alarms. Therefore, FAR was not prioritized over missed detection; rather, it was used as an operational tie-breaker after criteria that already incorporate positive class detection.

The final value of C was selected using the same validation-ordering principle: balanced accuracy, F1-score, FAR, and then AUPR. AUPR was placed last because it is a threshold-independent ranking metric. It is useful for comparing the quality of score ordering under class imbalance, but it does not by itself define the final binary operating point. Therefore, AUPR was used only as a final tie-breaker among otherwise similar validation configurations. The selected C and threshold were then applied unchanged to the test set and, in transfer experiments, to the target dataset without recalibration.

3.9. Evaluation Metrics

Performance was evaluated using accuracy, balanced accuracy (BA), precision, recall, F1-score, AUROC, AUPR, false alarm rate (FAR), and false negative rate (FNR). Because several datasets are highly imbalanced, AUPR was included as a ranking metric that is more informative than AUROC in skewed binary classification problems [4,6].

The false alarm rate was defined as

$$\text{FAR} = \frac{FP}{FP + TN},$$

where FP denotes false positives and TN denotes true negatives. The false negative rate was defined as

$$\text{FNR} = \frac{FN}{FN + TP} = 1 - \text{Recall},$$

where FN denotes false negatives and TP denotes true positives. FNR is especially relevant in security applications because it measures the proportion of attacks or positive events that remain undetected.

Balanced accuracy was defined as

$$\text{BA} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right).$$

Across the five seeds, results were reported as mean $\pm$ 95% confidence interval (CI). Since $n = 5$, CIs were computed using the Student- t multiplier $t_{0.975,4} = 2.776$.

3.10. Scaling Ablation

To determine whether improved numerical conditioning translated into predictive gains, two preprocessing variants were compared:

imputed_raw : median imputation only,

and

`imputed_scaled` : median imputation followed by standard scaling.

Both variants used the same logistic model, regularization grid, validation threshold search, and test protocol.

3.11. Cross-Dataset Transfer Protocol

Cross-dataset transfer was evaluated using only the common features. A model was trained on the source dataset training split, hyperparameters and thresholds were selected on the source validation split, and the resulting pipeline was evaluated on the target dataset test split. The source-fitted median imputer, standard scaler, model coefficients, and decision threshold were transferred without adaptation.

Cross-dataset transfer was performed only among CICIDS2017, UNSW-NB15, and DoH-Combined-L2. DoH-Combined-L1 was excluded from transfer because its positive class denotes DoH traffic rather than malicious or attack traffic. Therefore, including it in benign-versus-attack transfer matrices would confound protocol identification with intrusion detection.

3.12. Transfer-Degradation Metrics

To quantify out-of-distribution degradation, each transfer result was compared against the in-domain performance of the target dataset. For a metric M where larger values are better, the retention ratio was defined as

$$R_M(A \rightarrow B) = \frac{M(A \rightarrow B)}{M(B \rightarrow B)}.$$

For FAR, where lower values are better, the transfer delta was defined as

$$\Delta_{\text{FAR}}(A \rightarrow B) = \text{FAR}(A \rightarrow B) - \text{FAR}(B \rightarrow B).$$

Thus, positive FAR deltas indicate worse operational behavior than the target in-domain baseline.

3.13. Model Family Robustness Baselines

To verify whether the main findings were specific to ℓ_2 -regularized logistic regression, the cross-dataset transfer protocol was repeated using two classical nonlinear ensemble baselines: Random Forest and XGBoost [21,22]. Random Forest and XGBoost were selected because they represent widely used tree-based model families for tabular classification and provide stronger nonlinear baselines than a single decision tree.

The comparison was performed in the common feature space using the same fixed seeds, train-validation-test partitions, median imputation, standard scaling, validation-based threshold selection criterion, and evaluation metrics used in the main experiment. The experiment was restricted to common features because this is the only feature representation that supports direct cross-dataset transfer among CICIDS2017, UNSW-NB15, and DoH-Combined-L2 without changing the input schema.

For computational tractability and to keep the comparison balanced across model families, each source training split was capped at 100,000 stratified samples when the available training partition exceeded this size. For Random Forest, class imbalance was handled using class-balanced bootstrap weights. For XGBoost, imbalance was addressed through the positive class weighting parameter computed from the source training partition. In all cases, the decision threshold was selected on the validation set using the same ordering rule applied to the regularized logistic model: maximum balanced accuracy,

followed by F1-score, minimum FAR, and AUPR as a final ranking tie-breaker. The selected model and threshold were then evaluated unchanged on the corresponding in-domain or transfer target.

3.14. Distribution Shift Analysis

To quantify feature-level distribution shift, source training and target test distributions were compared after applying source-fitted median imputation and standard scaling. For each common feature, the two-sample Kolmogorov–Smirnov (KS) statistic was computed [23]. Wasserstein distance was also computed as a secondary descriptive metric grounded in optimal transport [24]. However, because Wasserstein distance and location differences are unbounded and may be dominated by heavy-tailed traffic features, the primary bounded shift score was defined as the mean KS statistic across the four common features:

$$D_{\text{KS}}(A \rightarrow B) = \frac{1}{4} \sum_{j=1}^4 D_{\text{KS}}^{(j)}(A_{\text{train}}, B_{\text{test}}).$$

This score lies in $[0, 1]$ and is therefore more suitable for cross-pair comparison than unbounded distance averages.

4. Results

4.1. Numerical Conditioning

Table 3 summarizes the condition number audit. Standard scaling reduced the condition number by several orders of magnitude in all datasets and feature configurations. The largest effect was observed for UNSW-NB15 common features, where the condition number decreased from approximately 7.93×10^6 to 1.30, corresponding to a reduction factor of approximately 6.09×10^6 . CICIDS2017 dataset-specific features also showed strong ill-conditioning before scaling, with a raw condition number of approximately 2.08×10^6 , reduced to approximately 85.9 after scaling.

These results support the numerical motivation for median imputation followed by standard scaling. The raw feature matrices combine heterogeneous physical units, including durations, byte counts, rates, packet counts, and port numbers. Such heterogeneous scales can produce ill-conditioned design matrices even when the underlying classification task is not intrinsically unstable.

Table 3. Condition number reduction after standard scaling. Values are mean $\pm$ 95% CI over five seeds.

Dataset	Feature Set	Raw $\kappa_2(X)$	Scaled $\kappa_2(X)$	Reduction Factor
CICIDS2017	Common	$4.93 \times 10^3 \pm 2.4 \times 10^3$	1.62 ± 0.39	$2.95 \times 10^3 \pm 7.3 \times 10^2$
CICIDS2017	Dataset-specific	$2.08 \times 10^6 \pm 1.4 \times 10^5$	85.9 ± 11.0	$2.44 \times 10^4 \pm 2.7 \times 10^3$
DoH-L1	Common	$1.09 \times 10^4 \pm 3.0 \times 10^2$	1.62 ± 0.008	$6.76 \times 10^3 \pm 2.1 \times 10^2$
DoH-L1	Dataset-specific	$1.27 \times 10^4 \pm 3.5 \times 10^2$	3.64 ± 0.005	$3.47 \times 10^3 \pm 92$
DoH-L2	Common	$4.52 \times 10^3 \pm 97$	3.29 ± 0.006	$1.37 \times 10^3 \pm 31$
DoH-L2	Dataset-specific	$7.19 \times 10^3 \pm 1.5 \times 10^2$	5.73 ± 0.009	$1.25 \times 10^3 \pm 26$
UNSW-NB15	Common	$7.93 \times 10^6 \pm 3.9 \times 10^4$	1.30 ± 0.005	$6.09 \times 10^6 \pm 3.2 \times 10^4$
UNSW-NB15	Dataset-specific	$8.27 \times 10^6 \pm 3.5 \times 10^4$	20.7 ± 0.49	$3.99 \times 10^5 \pm 9.9 \times 10^3$

Figure 2 visualizes the same numerical-conditioning result from Table 3 on a logarithmic scale. The figure highlights that standard scaling did not merely produce small numerical changes, but reduced the design matrix condition numbers by several orders of magnitude across all datasets and feature configurations.

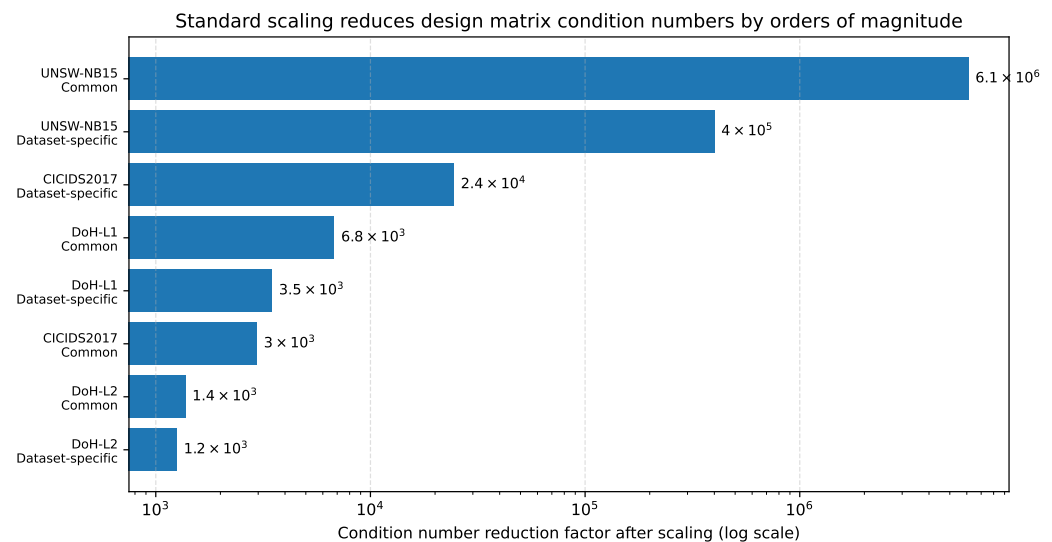


Figure 2. Condition number reduction factors after standard scaling. The horizontal axis is shown on a logarithmic scale to emphasize the orders-of-magnitude reduction in $\kappa_2(X)$ across datasets and feature configurations.

4.2. In-Domain Regularized Logistic Baseline

Table 4 reports the in-domain performance of ℓ_2 -regularized logistic regression using scaled preprocessing. Results show that the model provides stable and interpretable baselines across all datasets, but its operational behavior varies substantially by dataset and feature set.

For UNSW-NB15, both feature sets produced stable F1-scores around 0.745, while dataset-specific features improved AUPR from 0.7633 to 0.7959. This indicates that additional features improved ranking quality, but did not materially improve the final thresholded F1-score.

For CICIDS2017, common features produced higher recall but also a higher FAR. Dataset-specific features reduced FAR from 0.3297 to 0.2766, but also reduced recall from 0.9103 to 0.8072. Thus, the operational trade-off is not captured by a single ranking metric.

For DoH-Combined-L1, common features outperformed dataset-specific features in accuracy, F1-score, and FAR. This is an important result: adding more available variables did not improve operational decision quality. For DoH-Combined-L2, dataset-specific features achieved a slightly higher AUPR, but strongly reduced recall and F1-score while lowering FAR. This again shows that AUPR and thresholded operational metrics may disagree under imbalance.

Table 4. In-domain ℓ_2 -regularized logistic baseline with median imputation and standard scaling. Values are mean $\pm$ 95% CI over five seeds.

Dataset	Feature Set	Balanced Acc.	F1-Score	FAR	AUPR
CICIDS2017	Common	0.7903 ± 0.0062	0.5594 ± 0.0047	0.3297 ± 0.0034	0.4488 ± 0.0060
CICIDS2017	Dataset-specific	0.7653 ± 0.0027	0.5502 ± 0.0047	0.2766 ± 0.0077	0.5945 ± 0.0064
DoH-L1	Common	0.8301 ± 0.0011	0.7312 ± 0.0011	0.0915 ± 0.0011	0.7029 ± 0.0018
DoH-L1	Dataset-specific	0.8062 ± 0.0012	0.6366 ± 0.0014	0.2627 ± 0.0008	0.7335 ± 0.0021
DoH-L2	Common	0.7467 ± 0.0167	0.8708 ± 0.0038	0.2959 ± 0.0406	0.9678 ± 0.0021
DoH-L2	Dataset-specific	0.6719 ± 0.0044	0.6115 ± 0.0221	0.1003 ± 0.0315	0.9712 ± 0.0004
UNSW-NB15	Common	0.7556 ± 0.0003	0.7456 ± 0.0004	0.1616 ± 0.0019	0.7633 ± 0.0003
UNSW-NB15	Dataset-specific	0.7567 ± 0.0036	0.7444 ± 0.0027	0.1539 ± 0.0212	0.7959 ± 0.0001

4.3. Effect of Scaling on Predictive Performance

Table 5 reports the scaling ablation results for ℓ_2 -regularized logistic regression. The comparison evaluates two preprocessing variants: median imputation without scaling and median imputation followed by standard scaling. This experiment is important because the numerical conditioning audit showed that scaling substantially reduces the condition number of the feature matrices; however, Table 5 shows that this numerical improvement does not translate into uniformly superior predictive performance.

Although standard scaling consistently improved numerical conditioning, its predictive effect was not uniformly monotonic. In CICIDS2017 with dataset-specific features, the raw-imputed model obtained a slightly higher F1-score (0.5579 ± 0.0019) than the scaled model (0.5502 ± 0.0047), whereas the scaled model reduced FAR from 0.3052 ± 0.0011 to 0.2766 ± 0.0077 . This suggests that scaling made the model less aggressive in terms of false alarms, but did not improve thresholded F1-score in this configuration.

For DoH-Combined-L1, Table 5 shows that raw and scaled preprocessing produced nearly identical results in both common and dataset-specific feature configurations. This indicates that, for this dataset, scaling improved numerical conditioning without substantially changing the learned decision behavior. In contrast, DoH-Combined-L2 with common features shows a clearer operational trade-off: the raw model achieved a higher F1-score (0.9036 ± 0.0391), but with a much larger and more unstable FAR (0.4394 ± 0.1752). Scaling reduced FAR to 0.2959 ± 0.0406 and improved balanced accuracy from 0.7075 ± 0.0491 to 0.7467 ± 0.0167 , although F1-score decreased to 0.8708 ± 0.0038 .

Table 5. Effect of scaling on ℓ_2 -regularized logistic regression. Values are mean $\pm$ 95% CI over five seeds.

Dataset	Feature Set	Preprocessing	Balanced Acc.	F1-Score	FAR	AUPR
CICIDS2017	Common	Imputed raw	0.7954 ± 0.0086	0.5621 ± 0.0113	0.3374 ± 0.0155	0.4537 ± 0.0052
CICIDS2017	Common	Imputed + scaled	0.7903 ± 0.0062	0.5594 ± 0.0047	0.3297 ± 0.0034	0.4488 ± 0.0060
CICIDS2017	Dataset-specific	Imputed raw	0.7815 ± 0.0022	0.5579 ± 0.0019	0.3052 ± 0.0011	0.6091 ± 0.0015
CICIDS2017	Dataset-specific	Imputed + scaled	0.7653 ± 0.0027	0.5502 ± 0.0047	0.2766 ± 0.0077	0.5945 ± 0.0064
DoH-L1	Common	Imputed raw	0.8301 ± 0.0010	0.7310 ± 0.0009	0.0917 ± 0.0012	0.7030 ± 0.0019
DoH-L1	Common	Imputed + scaled	0.8301 ± 0.0011	0.7312 ± 0.0011	0.0915 ± 0.0011	0.7029 ± 0.0018
DoH-L1	Dataset-specific	Imputed raw	0.8062 ± 0.0011	0.6366 ± 0.0014	0.2627 ± 0.0009	0.7336 ± 0.0021
DoH-L1	Dataset-specific	Imputed + scaled	0.8062 ± 0.0012	0.6366 ± 0.0014	0.2627 ± 0.0008	0.7335 ± 0.0021
DoH-L2	Common	Imputed raw	0.7075 ± 0.0491	0.9036 ± 0.0391	0.4394 ± 0.1752	0.9712 ± 0.0039
DoH-L2	Common	Imputed + scaled	0.7467 ± 0.0167	0.8708 ± 0.0038	0.2959 ± 0.0406	0.9678 ± 0.0021
DoH-L2	Dataset-specific	Imputed raw	0.6729 ± 0.0034	0.6057 ± 0.0079	0.0918 ± 0.0132	0.9712 ± 0.0003
DoH-L2	Dataset-specific	Imputed + scaled	0.6719 ± 0.0044	0.6115 ± 0.0221	0.1003 ± 0.0315	0.9712 ± 0.0004
UNSW-NB15	Common	Imputed raw	0.7568 ± 0.0015	0.7401 ± 0.0057	0.1419 ± 0.0201	0.7563 ± 0.0085
UNSW-NB15	Common	Imputed + scaled	0.7556 ± 0.0003	0.7456 ± 0.0004	0.1616 ± 0.0019	0.7633 ± 0.0003
UNSW-NB15	Dataset-specific	Imputed raw	0.7602 ± 0.0029	0.7384 ± 0.0010	0.1243 ± 0.0111	0.7954 ± 0.0001
UNSW-NB15	Dataset-specific	Imputed + scaled	0.7567 ± 0.0036	0.7444 ± 0.0027	0.1539 ± 0.0212	0.7959 ± 0.0001

The results in Table 5 therefore support a nuanced conclusion: scaling is strongly justified from a numerical-conditioning perspective, but its predictive effect depends on the dataset, feature set, class imbalance, and threshold behavior. In particular, scaling may improve operational stability by reducing false alarm variability, as observed in DoH-Combined-L2, even when it does not maximize F1-score. Consequently, numerical conditioning, ranking quality, thresholded F1-score, and false alarm behavior must be evaluated jointly rather than interpreted as equivalent indicators.

4.4. Cross-Dataset Transfer

Cross-dataset transfer was evaluated using only the four common features and scaled preprocessing. Table 6 shows the F1-score transfer matrix. The diagonal entries correspond to in-domain performance, while off-diagonal entries correspond to transfer from a source dataset to a different target dataset.

The results reveal strong asymmetry. CICIDS2017 did not transfer effectively to either UNSW-NB15 or DoH-Combined-L2, yielding near-zero F1-scores. Conversely, UNSW-NB15 transferred partially to DoH-Combined-L2, achieving 0.8921 ± 0.0011 F1-score, slightly above the DoH-L2 in-domain F1-score. However, this transfer also produced a high FAR, as shown in Table 7. Therefore, this transfer direction should be interpreted as partial detection transfer rather than operationally robust deployment.

Table 6. Cross-dataset transfer matrix in terms of F1-score. Rows indicate source datasets and columns indicate target datasets. Values are mean $\pm$ 95% CI.

Source $\rightarrow$ Target	CICIDS2017	DoH-L2	UNSW-NB15
CICIDS2017	0.5594 ± 0.0047	0.0000 ± 0.0000	0.0004 ± 0.0003
DoH-L2	0.2869 ± 0.0014	0.8708 ± 0.0038	0.0235 ± 0.0065
UNSW-NB15	0.3111 ± 0.0006	0.8921 ± 0.0011	0.7456 ± 0.0004

Table 7. Cross-dataset transfer matrix in terms of false alarm rate. Rows indicate source datasets and columns indicate target datasets. Values are mean $\pm$ 95% CI.

Source $\rightarrow$ Target	CICIDS2017	DoH-L2	UNSW-NB15
CICIDS2017	0.3297 ± 0.0034	0.0000 ± 0.0000	0.0000 ± 0.0000
DoH-L2	0.6492 ± 0.0182	0.2959 ± 0.0406	0.0067 ± 0.0040
UNSW-NB15	0.9796 ± 0.0012	0.3950 ± 0.0065	0.1616 ± 0.0019

The transfer from UNSW-NB15 to CICIDS2017 is especially illustrative. It achieved high recall, but FAR reached 0.9796 ± 0.0012 , meaning that nearly all benign CICIDS2017 samples were incorrectly marked as positive. This confirms that recall alone is insufficient for operational IDS evaluation. The transferred threshold can collapse under distribution shift, producing either nearly no attack detections or excessive false alarms.

4.5. Relative Transfer Degradation

To compare transfer directions relative to target in-domain baselines, retention ratios were computed. Table 8 reports F1 retention and FAR delta for off-diagonal transfer directions. CICIDS2017 retained virtually no F1-score when transferred to UNSW-NB15 or DoH-L2. DoH-L2 retained about 51% of the CICIDS2017 in-domain F1-score, but increased FAR by 0.3195. UNSW-NB15 retained 102.46% of the DoH-L2 in-domain F1-score, but increased FAR by 0.0991.

Table 8. Relative transfer degradation for off-diagonal source–target pairs. F1 retention is computed relative to the target in-domain F1-score; FAR delta is computed relative to the target in-domain FAR.

Source	Target	F1 Retention Ratio	FAR Delta
CICIDS2017	DoH-L2	0.0000 ± 0.0000	-0.2959 ± 0.0406
CICIDS2017	UNSW-NB15	0.0005 ± 0.0004	-0.1616 ± 0.0019
DoH-L2	CICIDS2017	0.5129 ± 0.0051	0.3195 ± 0.0187
DoH-L2	UNSW-NB15	0.0315 ± 0.0087	-0.1550 ± 0.0055
UNSW-NB15	CICIDS2017	0.5561 ± 0.0048	0.6499 ± 0.0030
UNSW-NB15	DoH-L2	1.0246 ± 0.0047	0.0991 ± 0.0437

These findings show that transfer is highly asymmetric. Some directions preserve part of the detection capability, but at unacceptable operational cost. Other directions reduce FAR only because the model almost never predicts the positive class. Therefore, transfer degradation must be analyzed jointly through F1 retention and FAR change.

4.6. Model Family Robustness

To verify whether the main findings were specific to regularized logistic regression, the cross-dataset transfer protocol was repeated using Random Forest and XGBoost as classical nonlinear ensemble baselines. The comparison was performed in the common feature space using the same seeds, data splits, preprocessing, threshold selection criterion, and evaluation metrics used in the main experiment. Table 9 summarizes the resulting in-domain and off-domain performance in terms of F1-score, false negative rate, and false alarm rate.

Table 9. Model family robustness analysis including false negative rate. Values are mean $\pm$ 95% CI over the corresponding source–target evaluations.

Model	In-Domain F1	In-Domain FNR	In-Domain FAR	Off-Domain F1	Off-Domain FNR	Off-Domain FAR
ℓ_2 logistic	0.7240 $\pm$ 0.0812	0.1975 $\pm$ 0.0611	0.3017 $\pm$ 0.0656	0.2519 $\pm$ 0.1182	0.6311 $\pm$ 0.1517	0.3133 $\pm$ 0.1389
Random Forest	0.9107 $\pm$ 0.0359	0.0327 $\pm$ 0.0151	0.0912 $\pm$ 0.0373	0.2086 $\pm$ 0.0858	0.8218 $\pm$ 0.0757	0.1632 $\pm$ 0.0740
XGBoost	0.9437 $\pm$ 0.0186	0.0381 $\pm$ 0.0111	0.0655 $\pm$ 0.0343	0.2092 $\pm$ 0.0850	0.8235 $\pm$ 0.0674	0.1502 $\pm$ 0.0655

Figure 3 visualizes the main result of the model family robustness experiment. Although Random Forest and XGBoost substantially improved in-domain F1-score, their off-domain F1-score remained low, confirming that higher model expressiveness did not eliminate cross-dataset transfer degradation.

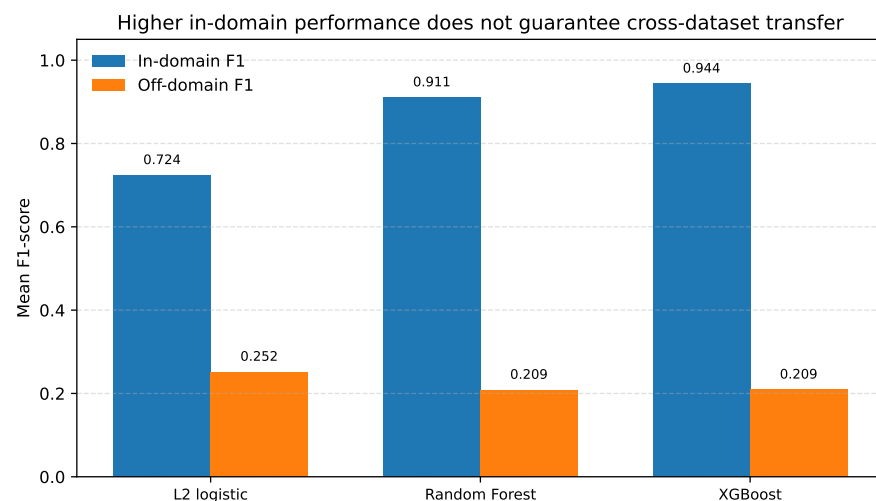


Figure 3. Model family robustness analysis comparing mean in-domain and off-domain F1-score across ℓ_2 -regularized logistic regression, Random Forest, and XGBoost. The figure shows that stronger in-domain performance does not guarantee reliable cross-dataset transfer.

The robustness experiment shows that nonlinear ensemble models substantially improved mean in-domain F1-score relative to the regularized logistic model. Random Forest increased mean in-domain F1-score from 0.7240 to 0.9107, while XGBoost further increased it to 0.9437. In-domain FNR also decreased substantially, from 0.1975 for the logistic model to 0.0327 for Random Forest and 0.0381 for XGBoost. Thus, the nonlinear baselines reduced missed detections under in-domain evaluation.

However, this improvement did not translate into reliable cross-dataset transfer. The mean off-domain F1-score remained low for all model families, with values of 0.2519, 0.2086, and 0.2092 for logistic regression, Random Forest, and XGBoost, respectively. More importantly for security-oriented evaluation, the mean off-domain FNR was high across all models: 0.6311 for logistic regression, 0.8218 for Random Forest, and 0.8235 for XGBoost. This

indicates that, under direct cross-dataset transfer, the stronger nonlinear baselines often missed a large fraction of positive target samples despite achieving strong in-domain performance.

These results support the main operational conclusion of the paper. Increasing model expressiveness can improve in-domain discrimination and reduce in-domain missed detections, but it does not remove the degradation induced by heterogeneous source–target distributions. In the robustness experiment, off-domain evaluation remained substantially weaker than in-domain evaluation across all model families. Therefore, IDS evaluation should jointly report FNR and FAR, since a model with fewer false alarms may still be unacceptable if it misses many attacks.

4.7. Distribution Shift Across Datasets

Table 10 reports the KS-based distribution shift matrix computed over the four common features. In-domain shift scores were low: 0.0033 ± 0.0007 for CICIDS2017, 0.0050 ± 0.0017 for DoH-L2, and 0.0646 ± 0.0012 for UNSW-NB15. By contrast, all cross-dataset shift scores were substantially larger, ranging from 0.5411 ± 0.0023 to 0.6725 ± 0.0010 .

Table 10. Distribution shift matrix based on the mean KS statistic over common standardized traffic features. Rows indicate source datasets and columns indicate target datasets. Values are mean $\pm$ 95% CI.

Source $\rightarrow$ Target	CICIDS2017	DoH-L2	UNSW-NB15
CICIDS2017	0.0033 ± 0.0007	0.6409 ± 0.0010	0.5411 ± 0.0023
DoH-L2	0.6408 ± 0.0010	0.0050 ± 0.0017	0.6647 ± 0.0004
UNSW-NB15	0.5624 ± 0.0007	0.6725 ± 0.0010	0.0646 ± 0.0012

Figure 4 provides a compact visual summary of the relationship among cross-dataset F1-score, false alarm rate, and KS-based distribution shift. The figure is intended to complement the transfer and shift tables by making source–target asymmetries easier to identify.

The visualization highlights that in-domain evaluations lie on the diagonal and generally show stronger operational behavior than off-diagonal transfer evaluations. It also illustrates that transfer degradation is directional rather than symmetric. For example, the UNSW-NB15-to-DoH-L2 direction preserves a high F1-score, but this favorable result must be interpreted together with the corresponding false alarm behavior and the large KS-based distribution shift. Thus, the figure reinforces the main interpretation that F1-score, FAR, and distribution shift must be analyzed jointly rather than independently.

The distribution shift results support the interpretation that cross-dataset transfer failures are associated with substantial feature distribution mismatch. However, the relationship is not strictly monotonic. The transfer from UNSW-NB15 to DoH-L2 achieved high F1 retention despite a high KS shift score. This indicates that distribution shift is an important explanatory factor, but not the only one. Threshold transferability, class imbalance, source–target label semantics, and directional effects also influence operational behavior.

Exploratory correlations between shift scores and transfer-retention metrics were computed only over six off-diagonal source–target pairs. Because of the small number of pairs, these correlations should be interpreted cautiously. The analysis did not provide sufficient evidence to claim a simple monotonic relationship between larger distribution shift and lower F1 retention. Therefore, the main conclusion is that distribution shift helps explain transfer degradation, but does not fully determine it.

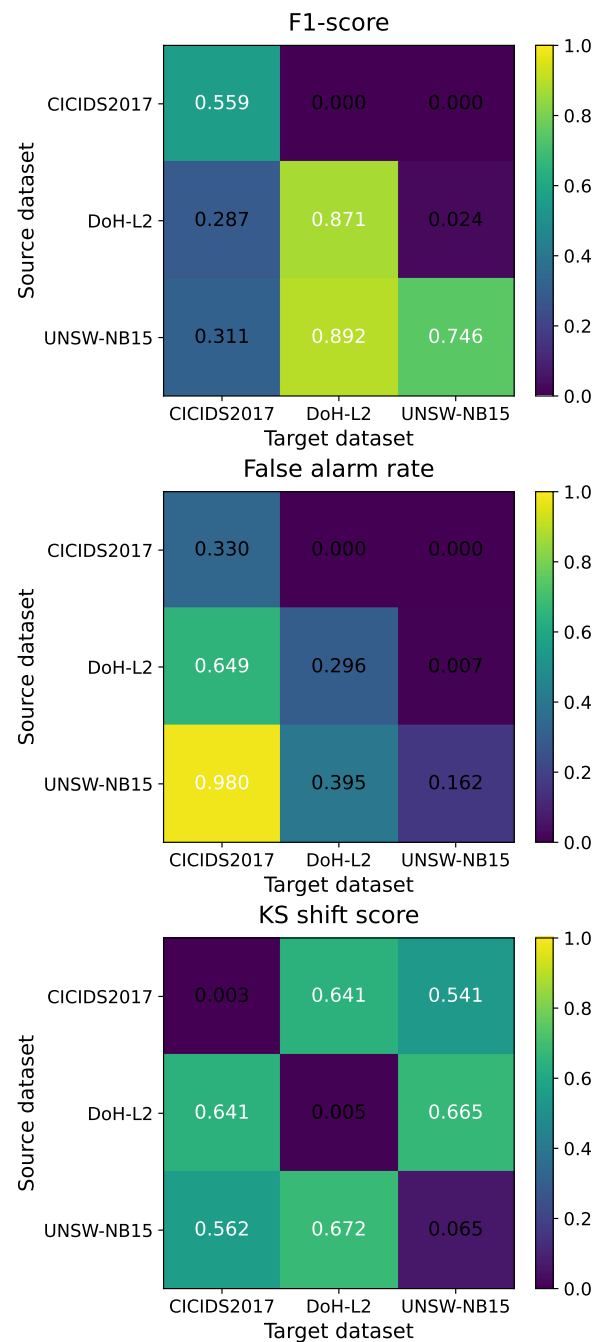


Figure 4. Cross-dataset transfer behavior and feature-level distribution shift. The three panels summarize F1-score transfer, false alarm rate transfer, and KS-based distribution shift over the common feature space. Rows indicate source datasets and columns indicate target datasets.

5. Discussion

The results provide evidence that numerical conditioning, feature scaling, class imbalance, threshold selection, and cross-dataset transfer cannot be treated as independent design decisions in intrusion detection pipelines. Although the predictive model used in this work is intentionally simple, the experiments reveal several non-trivial effects that are directly relevant to applied numerical analysis and computational learning.

First, the numerical audit showed that the raw feature matrices were often severely ill-conditioned before scaling. This was especially visible in UNSW-NB15, where the condition number of the common feature matrix was reduced from approximately 7.93×10^6 to approximately 1.30 after standard scaling. Similar reductions were observed in CICIDS2017

and the DoH datasets. This supports the use of scaling not merely as a conventional preprocessing step, but as a numerically motivated transformation that improves the conditioning of the optimization problem. Such behavior is consistent with classical numerical linear algebra, where ill-conditioning is associated with increased sensitivity of computations to perturbations in the data [7].

Second, the scaling ablation showed that improved conditioning does not automatically imply monotonic improvements in all predictive metrics. In some cases, raw feature optimization achieved comparable or higher F1-score, whereas scaled preprocessing reduced false alarm instability. The clearest example was DoH-Combined-L2 with common features: the raw model obtained higher F1-score, but with a substantially larger and more variable false alarm rate; the scaled model produced lower F1-score but a more controlled FAR. This suggests that the practical value of numerical stabilization should be evaluated jointly with operational metrics, not only with ranking metrics or threshold-independent measures.

Third, the comparison between common features and dataset-specific features showed that adding more variables does not necessarily improve operational performance. In DoH-Combined-L1, the common feature configuration achieved better F1-score and lower FAR than the dataset-specific configuration, despite the latter achieving higher AUROC and AUPR. In DoH-Combined-L2, dataset-specific features slightly improved AUPR but strongly reduced recall and F1-score while lowering FAR. This confirms that ranking quality and thresholded operational quality can diverge under class imbalance. This observation is aligned with prior work showing that ROC and PR analyses capture different aspects of classifier behavior, especially in skewed binary problems [4–6].

Fourth, the cross-dataset transfer experiments showed that numerically stabilized common features were not sufficient to guarantee robust out-of-distribution generalization. In-domain performance was relatively stable, but off-diagonal transfer was highly asymmetric. Models trained on CICIDS2017 almost completely failed to detect positive samples in UNSW-NB15 and DoH-Combined-L2, producing near-zero F1-scores. Conversely, the model trained on UNSW-NB15 transferred partially to DoH-Combined-L2 in terms of F1-score, but at the cost of a higher false alarm rate. This behavior is consistent with the broader dataset shift literature, where performance can degrade when the joint distribution of features and labels differs between training and deployment conditions [8,18].

Fifth, threshold transferability emerged as a major limitation. The transferred source domain thresholds frequently collapsed on the target domain, leading to two opposite operational failures: almost no positive detections, as in CICIDS2017-to-UNSW-NB15 and CICIDS2017-to-DoH-L2, or excessive false alarms, as in UNSW-NB15-to-CICIDS2017. This indicates that a model may retain some ranking ability while still failing as a decision system if the operating threshold is not recalibrated. This point is related to calibration research, which has shown that predicted probabilities and decision confidence may become unreliable under distributional changes [17,18]. The validation threshold selection order also reflects this operational perspective. Balanced accuracy was used first to avoid favoring the majority class and to account jointly for sensitivity and specificity. F1-score was used next to preserve positive class detection quality, which is critical in security applications where missed attacks are costly. FAR was used only as an operational tie-breaker among thresholds with comparable detection performance, and AUPR was reserved as a final threshold-independent ranking tie-breaker during model selection. This ordering is particularly relevant under dataset shift: even a threshold selected with an imbalance-aware rule may fail when transferred to a target dataset with different class priors, score distributions, or label semantics.

Sixth, the distribution shift analysis confirmed that the in-domain train–test discrepancies were small, whereas cross-dataset discrepancies were consistently large under the bounded KS-based score. The in-domain KS scores were close to zero for CICIDS2017 and DoH-Combined-L2, while all cross-dataset scores were above 0.54. This supports the interpretation that transfer degradation is associated with substantial distributional mismatch in common traffic features. However, the relationship was not strictly monotonic. The transfer from UNSW-NB15 to DoH-Combined-L2 achieved high F1 retention despite a large KS shift score. Therefore, distribution shift is an important explanatory factor, but not the only determinant of transfer behavior.

A plausible explanation for this apparently favorable transfer direction is that the KS-based score measures marginal feature distribution discrepancy, not the preservation of the source decision boundary, the conditional label distribution, or the calibration of transferred probabilities. Thus, two datasets may exhibit large marginal differences in common traffic features while still allowing a source-trained linear separator to assign many target positive samples to the correct side of the decision boundary. This effect is particularly relevant for DoH-Combined-L2, where the positive class is highly dominant. In such a setting, a transferred classifier that predicts many samples as positive can obtain a high F1-score even if its false alarm behavior deteriorates. This explains why the UNSW-NB15-to-DoH-L2 transfer direction preserved F1-score but still increased FAR.

Consequently, high F1 retention under large KS shift should not be interpreted as robust transfer by itself. It may instead indicate that the transferred decision rule remains partially aligned with the target positive class while being poorly calibrated for operational deployment. This reinforces the need to evaluate cross-dataset transfer through multiple complementary indicators, including F1 retention, FAR delta, class prior changes, threshold behavior, and distribution shift scores.

The results also show why DoH-Combined-L1 should not be included in cross-dataset benign-versus-attack transfer. Its positive class represents DoH traffic rather than malicious traffic, and therefore it corresponds to protocol identification rather than intrusion detection. Including it in the transfer matrix would mix different binary semantics and could produce misleading conclusions. Similarly, DoH-Combined-L3 was excluded because the present study focuses on binary detection and the validated processed data used in the experiments covered Layer 1 and Layer 2. Adding Layer 3 would require a separate label-mapping and validation strategy.

Overall, the main finding is not that a simple regularized logistic model is sufficient for intrusion detection, but that such a model provides a useful mathematical probe for studying numerical and operational behavior. Its convex objective, low computational cost, and sensitivity to feature scaling make it suitable for analyzing the interaction between conditioning, regularization, class imbalance, and transfer degradation. More complex models may improve absolute predictive performance, but they may also obscure the numerical mechanisms that this study aims to isolate.

The model family robustness experiment further indicates that the main transfer-related conclusions are not limited to the linear logistic model. Random Forest and XGBoost achieved higher in-domain F1-scores and lower in-domain false alarm rates than the regularized logistic model, confirming that nonlinear ensembles can improve in-domain discrimination. They also reduced in-domain FNR substantially, which is important for security-sensitive applications where missed attacks are especially costly. However, their mean off-domain FNR remained above 0.82, despite in-domain F1-scores above 0.91. This shows that model expressiveness alone does not eliminate the operational risks induced by heterogeneous traffic distributions. Explicit threshold validation, missed detection

analysis through FNR, false alarm analysis through FAR, and cross-dataset evaluation remain necessary regardless of the classifier family.

5.1. Implications

The findings have three practical implications. First, preprocessing should be justified numerically when models are trained on heterogeneous traffic features with different physical units. Second, intrusion detection evaluation should report missed detection and false alarm metrics, especially FNR and FAR, together with F1-score, AUROC, and AUPR. Third, cross-dataset validation should not rely only on in-domain test performance, because transferred thresholds may fail under distribution shift even when the model performs acceptably on the source dataset.

5.2. Limitations

This study has several limitations. First, it focuses on binary classification only. Multiclass intrusion detection was not evaluated because several CICIDS2017 attack classes contain very few samples, which would require a different evaluation strategy. Second, the main predictive model is restricted to ℓ_2 -regularized logistic regression. This is intentional for numerical interpretability, but it limits claims about deep or ensemble models. Third, the distribution shift correlation analysis was exploratory because only six off-diagonal source–target pairs were available. Therefore, the results should not be interpreted as proving a universal monotonic relationship between KS shift and transfer degradation. Fourth, cross-dataset transfer was evaluated without target domain recalibration. This was appropriate for testing direct transferability, but future work should evaluate source-free calibration, unsupervised adaptation, and domain-alignment strategies.

5.3. Future Work

Future work should extend the framework in four directions. First, multiclass detection should be analyzed with class-aware metrics and rare-class handling. Second, the model family robustness analysis should be expanded to include kernel methods, calibrated probabilistic models, lightweight neural architectures, and broader hyperparameter searches. Third, post-transfer threshold recalibration should be evaluated to determine whether the observed collapse is mainly due to score miscalibration or deeper representation mismatch. Fourth, the distribution shift analysis should be extended with multivariate distances and domain-adaptation diagnostics, since univariate KS statistics may not capture feature interactions.

6. Conclusions

This paper presented a numerical-computational analysis of ℓ_2 -regularized logistic learning for binary intrusion detection under heterogeneous datasets, class imbalance, and cross-dataset shift. The goal was not to propose a new IDS architecture, but to examine how numerical conditioning, feature scaling, feature set design, threshold selection, and distribution shift affect operational detection behavior.

Accordingly, the conclusions should be interpreted within this scope. This work does not introduce a new machine learning algorithm, preprocessing technique, thresholding rule, or distribution shift metric, and it does not aim to establish state-of-the-art intrusion detection accuracy. Its contribution is instead methodological and analytical: it provides a controlled numerical-computational framework for studying how established components—standard scaling, ℓ_2 -regularized learning, validation-based threshold selection, operational IDS metrics, and KS-based shift analysis—interact under ill-conditioning, class imbalance, threshold transfer, and cross-dataset shift.

The numerical audit showed that median imputation followed by standard scaling reduced the spectral condition number of the feature matrices by several orders of magnitude across CICIDS2017, UNSW-NB15, and CIRA-CIC-DoHBrw-2020. This confirms that preprocessing is not merely a routine engineering step, but a numerically relevant component of the learning pipeline when traffic features combine heterogeneous physical units such as durations, byte counts, packet rates, and ports.

The in-domain experiments showed that ℓ_2 -regularized logistic regression provides a stable and interpretable baseline across heterogeneous IDS datasets. However, the relationship between feature richness and operational performance was non-monotonic. Dataset-specific features sometimes improved ranking metrics such as AUROC or AUPR, but did not necessarily improve F1-score or false alarm behavior. This finding is especially relevant for IDS evaluation, where a model with strong ranking performance may still be operationally problematic if its thresholded decisions generate excessive false alarms.

The scaling ablation further showed that better numerical conditioning does not automatically imply uniformly better predictive performance. In some cases, raw feature optimization achieved comparable or higher F1-score, whereas scaled preprocessing produced more controlled false alarm behavior. Therefore, numerical stability, ranking quality, thresholded detection quality, and false alarm rate should be analyzed jointly rather than treated as interchangeable indicators.

The cross-dataset transfer experiments revealed strong source–target asymmetry. Models trained on CICIDS2017 failed to transfer effectively to UNSW-NB15 and DoH-Combined-L2, while UNSW-NB15 transferred partially to DoH-Combined-L2 in terms of F1-score but at the cost of increased false alarms. These results indicate that common numerically stabilized features are not sufficient to guarantee robust out-of-distribution generalization. In particular, source-selected thresholds may collapse on target datasets, producing either near-zero positive detection or excessive false alarms.

The distribution shift analysis confirmed that in-domain discrepancies were small, whereas cross-dataset discrepancies were consistently large under the bounded KS-based shift score. Nevertheless, the relationship between distribution shift and transfer degradation was not strictly monotonic. This suggests that distributional mismatch is an important explanatory factor, but not the only determinant of transfer behavior. Threshold calibration, class priors, label semantics, and source–target directionality also play important roles.

Overall, the main contribution of this work is a reproducible numerical-computational framework that links matrix conditioning, scaling, regularized optimization, class imbalance, operational IDS metrics, cross-dataset transfer degradation, and feature-level distribution shift. Future work should extend this framework to multiclass detection, nonlinear and ensemble models, post-transfer threshold recalibration, and domain adaptation methods.

Author Contributions: Conceptualization, M.A.-A.; methodology, M.A.-A. and R.B.; software, M.A.-A. and R.B.; validation, R.B. and M.O.; formal analysis, M.A.-A.; investigation, M.A.-A.; resources, M.A.-A.; data curation, M.A.-A.; writing—original draft preparation, M.A.-A.; writing—review and editing, M.O.; visualization, R.B.; supervision, M.A.-A.; project administration, M.A.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was not funded by any organization.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in this study are publicly available from their original providers: CICIDS2017, UNSW-NB15, and CIRA-CIC-DoHBrw-2020. To support reproducibility, the author created a public repository containing the experimental scripts, configuration files, split-generation procedures, aggregated result tables, and LaTeX table exports used in this

study in <https://github.com/miguelarcosa/Numerically-Stabilized-Regularized.git> (accessed on 3 June 2026).

Acknowledgments: During manuscript preparation, the authors used generative artificial intelligence tools (ChatGPT, GPT-5.2 Thinking; OpenAI, San Francisco, CA, USA; accessed on 5 May 2026) for language refinement, editorial assistance, and support in generating/formatting figures from the experimental outputs. All study design, data processing, analysis, interpretation of results, and final manuscript review were carried out by the authors, who take full responsibility for the content of the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*; SciTePress: Setúbal, Portugal, 2018; pp. 108–116. [\[CrossRef\]](#)
- Moustafa, N.; Slay, J. UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set). In *Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS)*; IEEE: New York, NY, USA, 2015; pp. 1–6. [\[CrossRef\]](#)
- MontazeriShatoori, M.; Davidson, L.; Kaur, G.; Habibi Lashkari, A. Detection of DoH Tunnels Using Time-Series Classification of Encrypted Traffic. In *Proceedings of the 2020 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCOM/CyberSciTech)*; IEEE: New York, NY, USA, 2020; pp. 63–70. [\[CrossRef\]](#)
- Davis, J.; Goadrich, M. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning*; ACM: New York, NY, USA, 2006; pp. 233–240. [\[CrossRef\]](#)
- Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [\[CrossRef\]](#)
- Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [\[CrossRef\]](#) [\[PubMed\]](#)
- Higham, N.J. *Accuracy and Stability of Numerical Algorithms*, 2nd ed.; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2002. [\[CrossRef\]](#)
- Quiñero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; Lawrence, N.D. (Eds.) *Dataset Shift in Machine Learning*; MIT Press: Cambridge, MA, USA, 2009.
- Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A Theory of Learning from Different Domains. *Mach. Learn.* **2010**, *79*, 151–175. [\[CrossRef\]](#)
- Manocchio, L.D.; Layeghy, S.; Lo, W.W.; Kulatilleke, G.K.; Sarhan, M.; Portmann, M. FlowTransformer: A Transformer Framework for Flow-Based Network Intrusion Detection Systems. *Expert Syst. Appl.* **2024**, *241*, 122564. [\[CrossRef\]](#)
- Ullah, F.; Ullah, S.; Srivastava, G.; Lin, J.C.W. IDS-INT: Intrusion Detection System Using Transformer-Based Transfer Learning for Imbalanced Network Traffic. *Digit. Commun. Netw.* **2024**, *10*, 190–204. [\[CrossRef\]](#)
- Wang, C.; Xu, D.; Li, Z.; Niyato, D. Effective Intrusion Detection in Highly Imbalanced IoT Networks With Lightweight S2CGAN-IDS. *IEEE Internet Things J.* **2024**, *11*, 15140–15151. [\[CrossRef\]](#)
- Ding, H.; Sun, Y.; Huang, N.; Shen, Z.; Cui, X. TMG-GAN: Generative Adversarial Networks-Based Imbalanced Learning for Network Intrusion Detection. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 1156–1167. [\[CrossRef\]](#)
- Wei, N.; Wu, Z.; Yin, L.; Cai, X. HybridGF: A Novel Data Augmentation Algorithm for Detecting the Extremely Imbalanced Malicious Traffic in Industrial Internet of Things. *IEEE Internet Things J.* **2026**, *13*, 7179–7191. [\[CrossRef\]](#)
- He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [\[CrossRef\]](#)
- Fan, R.E.; Chang, K.W.; Hsieh, C.J.; Wang, X.R.; Lin, C.J. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.
- Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1321–1330.
- Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
- Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009. [\[CrossRef\]](#)

20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
21. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
22. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794. [[CrossRef](#)]
23. Massey, F.J. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **1951**, *46*, 68–78. [[CrossRef](#)]
24. Villani, C. *Optimal Transport: Old and New*; Grundlehren der Mathematischen Wissenschaften; Springer: Berlin, Germany, 2009; Volume 338. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.