Cross-Dataset Generalization Framework for Cybercrime Detection Using CIC-IDS2017, CIC-Phishing2019, and Malicious Uniform Resource Locator (URL) Data

R. Adinarayana [1,*] and G. Vamsi Krishna [2]

[1] Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India
[2] Department of Computer Science and Engineering, Dr. Lankapalli Bullayya College of Engineering (A), Visakhapatnam, India
Email: aaditeresa@gmail.com (R.A.); vamsikrishna527@gmail.com (G.V.K.)
*Corresponding author

***Abstract*—Proposed cybercrime detection models demonstrate satisfactory performance on specific benchmark datasets; however, they are not always robust across diverse, practical scenarios. This paper examines the generalization performance of a cross-dataset measure for cybercrime detection across network traffic, phishing email, and malicious Uniform Resource Locator (URL) domains. The work combines three popular security benchmarks Canadian Institute for Cybersecurity Intrusion Detection System Dataset2017 (CIC-IDS2017), Canadian Institute for Cybersecurity Phishing Dataset2019 (CIC-Phishing2019), and Malicious URL 2020 in a single preprocessing and learning pipeline to avoid dataset-specific bias. One or more datasets are used to train models, which are then directly evaluated on previously unseen datasets to verify transferability across distribution shifts. We will discuss performance while considering accuracy, F1−Score, Receiver Operating Characteristic (ROC), and fold-to-fold stability. Experiments demonstrate that direct training with a single random source results in significant performance deterioration, with a 23% decrease in F1−Score when applied to unseen datasets. In contrast, the degradation caused by the proposed framework is kept below 10% and maintains Receiver Operating Characteristic–Area Under the Curve (ROC–AUC) values consistently above 0.90. Paired significance testing demonstrates that the gains in robustness are highly significant ($p < 0.01$). The results indicate that cross-dataset evaluation is essential for the development of deployable cybercrime detection systems, providing empirical insights for developing generalization-oriented security analytics.**

***Keywords*—cross-dataset generalization, cybercrime detection, network traffic analysis, phishing email detection, malicious Uniform Resource Locator (URL) classification, security benchmark datasets**

I. Introduction

A well-written introduction will provide your study The massive upsurge in digitised infrastructure and application in the critical infrastructure domains (banking, health, communication, power etc.) is accompanied by the rising tide of cyber-crime. Attacks today don't just come in from one source, they're happening on multiple digital fronts, including anomalous network behavior, phishing emails, malicious Uniform Resource Locators (URLs). As such, cybercrime detection systems are required to deal with heterogeneous data sources and should be capable of stable performance in the face of different attack scenarios [1, 2]. In recent years, machine learning based cybercrime detection models made significant progress on a number of benchmark datasets such as Canadian Institute for Cybersecurity Intrusion Detection System Dataset2017 (CIC-IDS2017), Canadian Institute for Cybersecurity Phishing Dataset2019 (CIC-Phishing2019) and malicious Uniform Resource Locator (URL) corpora [3, 4]. Classical and ensemble learning techniques, such as Random Forest (RF) and gradient boosting, have achieved high accuracy and F1−Scores in identifying malicious traffic by learning discriminative patterns from structured flow statistics, lexical URL features, and indicators in email content [5−7]. Nevertheless, even after performing exceptionally well on a single black-box test, models trained and tested on a single database are likely not robust to generalization when deployed in wild settings with shifts between training data and real-world scenarios, including dataset bias or unseen attack behaviours [8, 9]. The inability to generalize across datasets has become a significant barrier in existing cybercrime detection studies. A commonly used between-dataset evaluation protocol is the within-dataset approach, in which the training and test sets are drawn from the same distribution. These evaluation conditions tend to overestimate the results they achieve in real-life and shed little light on how detection systems genuinely perform when confronted with atypical data sources or novel operating environments. In reality, Security Operations Centre (SOC) observes traffic patterns, phishing templates and malware URLs which are

Manuscript received February 8, 2026; revised April 8, 2026; accepted May 6, 2026; published July 23, 2026.

not well represented by standard benchmark datasets resulting in decrease in accuracy in alerting [10]. We note that some recent attempts have targeted the robustness and transferability in security analytics. Such deviations can have dramatic effects on classification accuracy, and here we point out that inter-source learning (including cross-domain or cross-dataset) in both the fields of intrusion detection and phishing analysis reports a significant drop in the accuracy of detections when a model trained on one source is employed for detecting malwares in external sources with F1−Score less than 20% [11]. Our findings highlight the significance of generalization-oriented evaluation in cybercrime detection and further motivate the move towards benchmarking from a holistic perspective. The first acknowledgement, however, has had relatively little application in the cybercrime detection literature. Previous studies have also often focused on a single form of cybercrime (e.g., network intrusion, phishing emails). It also does not support comparing different crime types using the same feature engineering, training and validation strategy [12]. In addition, cross-dataset comparisons are frequently hindered by inconsistent feature representation, pathological class imbalance and dataset dependent calibration [13] that prevent clear understanding of model generalization. This paper introduces a systematic cross-dataset generalization framework for cybercrime detection, which addresses the above limiting issues faced by the proposed methods that is using three popular security benchmarks pool: CIC-IDS2017, CIC-Phishing2019, and Malicious URL 2020. By unifying these datasets into a single learning pipeline and letting models be tested directly on the held-out dataset without retraining, it is more realistic to show transferability under distribution shift. Comprehensive performance comparison is conducted using various important measures such as F1−Score, ROC–AUC and stability variance, accompanied by statistical testing to validate the significance of the achieved improvements. This work has the following three contributions.

First, it provides a replicable cross-dataset evaluation framework for robust assessment of cybercrime detection across network, URL, and email domains.

Second, it measures the difference in performance reduction between conventional single-dataset training pipelines and novel generalization-based learning approaches that overcome this substantial decrement.

Third, the study offers empirical evidence on dataset bias and transferability, immediately applicable to deploying detection systems in operational security environments.

By shifting the evaluation focus from separate benchmark performance to cross-dataset generalization, this research provides new insights that support the future development of robust, deployable cybercrime detection systems that perform effectively in a realistic, evolving threat landscape. Motivation of using cross-dataset generalisation approach in cybercrime detection is depicted in Fig. 1. A model-based detection method is introduced in which the detector makes use of labeled network traffic flows from the CIC-IDS2017 dataset. The performance drop was obvious when the model was directly used to test unseen datasets like CIC-Phishing2019 and the Malicious URL dataset, as shown by cross-dataset testing (i.e., where accuracy dropped, F1−Score got lower, and more false alarms were shown). This reduced performance suggests how distribution mismatch and dataset-specific bias affect in isolated training benchmarks. The model learns to generalize in a generalization-inclined way and thus is better generalized across unseen security datasets. As a result, it has the benefits of better detection stability which implies a higher precision and F1−Score with less false alarms if tested in unseen domains. The illustration of Fig. 1 shows the performance transition from a benchmark-dependent one to a robust one across datasets, and it also demonstrates the necessity of conducting the evaluation of cross-dataset generalization for cybercrime detection systems.

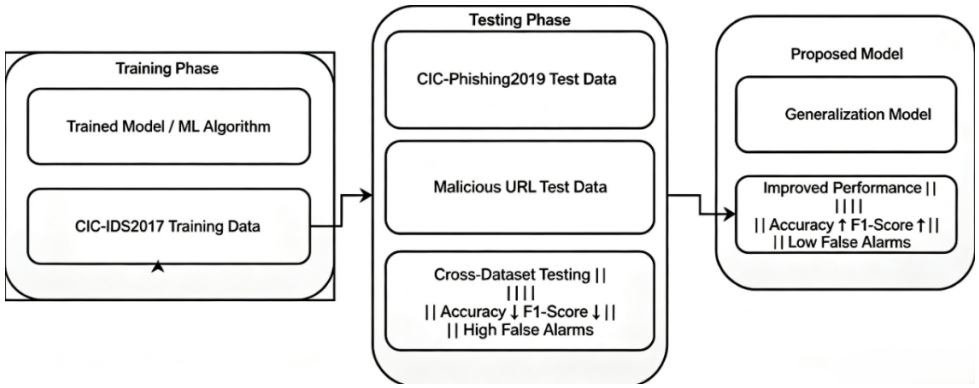

Fig. 1. Motivation for cross-dataset cybercrime detection.

II. Related Work

Cybercrime detection approaches are primarily evaluated in the context of supervised learning using publicly available benchmark datasets. Datasets such as the CIC-IDS2017 for network intrusion detection, CIC-Phishing2019 for email-based attacks, and Malicious URL repositories have been key to promoting empirical research, as they enable controlled experiments and

reproducibility [14]. Nearly all previous work follows within-dataset evaluation, where samples are split into randomly selected training and test sets from the same distribution. Performance is often measured using classification-based metrics such as accuracy, precision, recall, F1–Score, and ROC–AUC. While within-dataset performance yields an initial assessment of discriminability, it is ill-suited to predicting the operational deployment characteristics of cybercrime detection systems against arbitrarily non-stationary dynamics. In reality, these systems face changing traffic, novel phishing tactics, and new, unknown malicious URL patterns that differ from those in the training data used. These distributional mismatches cause a shift in the distribution, which usually leads to a drop in model performance when models are applied out-of-distribution [15, 16]. The results have empirically shown that the F1−Score decrease can exceed 20% for models with good benchmark performance in cross-dataset evaluation, suggesting that the model's generalization ability needs to be further improved when facing new scenes. Single-dataset validation is ubiquitous and a weak point in current evaluation practice [17].

Most intrusion detection systems test only on CIC-IDS2017 or UNSW-NB15, and phishing and malicious URL studies often use separate email corpora or structured URL datasets [11, 17, 18]. Such separated evaluation paradigms blur the line between dataset-specific artefacts and truly transferable malicious behaviour patterns. As a result, the invoked performance measures are likely to overstate robustness and offer limited value for predicting generalization across disparate security domains. Recent research is the cornerstone of the robustness and generalization study towards cybercrime detection. Recent explainable AI frameworks using multi-source digital evidence have also improved interpretability and transparency in cybercrime analytics [19]. Cross-domain learning for intrusion detection has proved that the distribution of features is quite different among datasets, because network topologies and traffic generation models [18], attack modes are not the same. Phishing and malicious URL detection methods similarly face rapidly changing features as attackers change lexical formats, content models, or perturbation system to avoid detection. However, systematic cross-dataset evaluations are still limited and most works fine-tune models on a single dataset thereby hiding their real generalisation performance.

Another downside of current evaluation procedures is the absence of protocol harmonization between datasets. The differences in the preprocessing pipeline, feature normalization techniques, and the way of handling class imbalance or label granularity makes it even harder to compare across datasets fairly. Network intrusion data sets are often multi-class attack taxonomies. On the other hand, phishing and URL datasets are usually binary/coarse grain which is challenging to compare metric results across model predictions and threshold calibrations. The lack of a uniform model or frame of analysis reduces the possibility of comparing studies and generalizing. Current survey and position papers increasingly call for evaluations that focus on generalization rather than benchmarks and explicitly measure robustness to dataset bias and domain shift. Best practices include training on one set and testing on held-out sets, reporting degradation ratios between results and those of state-of-the-art methods, analysing stability across cross-validation folds, and applying statistical tests to justify the observed improvement. Such testing methodologies enable more accurate estimates of system performance and bring the experimental setup as close as possible to real-world security operations. State-of-the-art evaluation methodologies for cybercrime detection largely favour within-dataset performance and fail to consider dataset bias, domain shift, and transferability. Recent trust-aware federated learning approaches have further enabled secure collaborative AI analytics without centralized data sharing [20]. Domain adaptation techniques such as Maximum Mean Discrepancy (MMD) [21], CORAL [22], and Invariant Risk Minimization (IRM) [23] have been proposed to address distribution shift.

Table I describes the comparison analysis of various research methods in the specific domain.

TABLE I. COMPARISON OF VARIOUS RESEARCH METHODOLOGIES

Study	Domain	Datasets Used	Evaluation Protocol	Reported Performance	Cross-Dataset Degradation	Key Limitation
[12]	Network IDS	CIC-IDS2017	Within-dataset (80–20)	Acc. ≈ 92–94%	Not evaluated	Dataset-bound evaluation
[15]	Network IDS	CIC-IDS2017, UNSW-NB15	Within-dataset CV	F1 ≈ 93%	>18% F1 drop (external test)	Poor transferability
[11]	Phishing	CIC-Phishing, Enron	Cross-dataset (limited)	F1 ≈ 91% (in-domain)	21–24% F1 drop	Domain-specific features
[17]	Malicious URLs	Malicious URL 2020	Within-dataset	Acc ≈ 94%	Not evaluated	No robustness analysis
[18]	Network IDS	IoT IDS datasets	Cross-domain (IoT only)	Acc ≈ 90–92%	≈15% accuracy drop	Single domain focus
[16]	Mixed (Survey)	Multiple IDS datasets	Benchmark-based	Acc up to 95%	Qualitative only	No statistical validation
[10]	Cybersecurity Analytics	Multiple surveys	Meta-analysis	—	—	Lack of unified protocol
This Work	Network, Email, URL	CIC-IDS2017, CIC-Phishing2019, Malicious URL 2020	Strict cross-dataset (no retraining)	F1 > 0.90 (avg.)	<10% F1 drop	—

III. DATASETS

We use three public benchmark datasets that, in total, represent a wide range of cybercrime, including networking traffic, email, and web-based threats. They are popular benchmark datasets in cybersecurity and exhibit heterogeneity across feature spaces, class distributions, and attack profiles. These intrinsic differences make them easy examples for evaluating robustness and generalization in cross-dataset environments.

A. CIC-IDS2017 (Network Intrusion Detection)

The CIC-IDS2017 dataset was generated by the Canadian Institute for Cybersecurity to simulate a wide range of network intrusion scenarios in a laboratory environment. It comprises roughly 2.8 million flow-level observations collected over several days. It includes legitimate traffic in addition to many types of attacks such as DDoS attacks, intrusion, botnet activity, brute force attack, injection, and web application attacks. Each network flow is accompanied by statistical and behavioral characteristics, including packet counts, byte distributions, flow duration, and interpacket arrival times. Being massive in nature and heterogeneous, CIC-IDS2017 is widely used for IDS experimentation that also acts as an illustrative representation of network-based based diagnosis.

B. CIC-Phishing2019 (Email Phishing Detection)

CIC-Phishing2019 is a well-prepared dataset on phishing attacks based on email communication. The dataset comprises approximately 27,000 labeled email samples (phishing and legitimate). An email record includes a header, metadata attributes, and text from the message body. Phishing emails in the corpus employ typical social engineering tactics, including forged sender addresses, misleading subject lines, and embedded malicious URLs. The dataset mimics realistic phishing campaigns and has been widely used to assess e-mail-based cybercrime detection systems.

C. Malware URL 2020 (Web Threat Detection)

Database 2020 is a large repository of both benign and malicious URLs, categorized as benign, phishing, malware, or defacement. For fair comparison between experiments, we will use a balanced subset with 200K URL samples in this paper. The lexical and structural features of each URL, such as URL length, character ratio, entropy, and whether suspicious tokens are involved in URLs. Those characteristics are all tactics attackers use to stay hidden and confuse the bad guys. It is widely used in webpage threat detection and provides a novel feature space if compared with network and email datasets.

D. Dataset Diversity and Generalization Difficulty

The three datasets are irrevocably distinct in data modality, feature descriptor, granularity of labeled information, and class imbalance. For example, traffic data in a network is organized and quantitative, while email spam/phishing messages consist of structural metadata and unstructured text. URL-based datasets, too, rely on token-based lexical patterns. This heterogeneity can result in distributional shifts in the data source for detection model training. The comparison with these datasets is a fair evaluation towards generalization ability, and it shows the limitations of learning techniques over dataset-specific domains.

TABLE II. OVERVIEW OF DATASETS

Dataset	Cybercrime Domain	Total Samples	Classes	Data Nature
CIC-IDS2017	Network Intrusion Detection	~2.8 million flows	15 (1 benign + 14 attack types)	Structured numerical data
CIC-Phishing2019	Email Phishing Detection	~27,000 emails	2 (phishing, legitimate)	Semi-structured text & metadata
Malicious URL 2020	Web-Based Threat Detection	~200,000 URLs (balanced subset)	4 (benign, phishing, malware, defacement)	String-based lexical data

Table II summarizes the datasets employed in this study, outlining their variety with respect to cybercrimes types, feature representations and class imbalances. Examples of these datasets are network traffic, email and web-threats, which present different data features and label granularities. This diversity not only leads to inherent distributional differences which serve as a source of domain shift, but also makes the model transfer more challenging. Therefore, these datasets are appropriate for testing cross-dataset generalization of cybercrime detection models.

IV. MATHEMATICAL MODEL FOR GENERALIZATION-AWARE CROSS-DATASET CYBERCRIME DETECTION

Specifically, we cast the cybercrime detection problem as a multi-dataset learning task, where each dataset encodes a distinct operational scenario with its own data distribution. Let $\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}$ denote the k-th dataset, where $x_i^{(k)}$ are, respectively, the cyber event feature vector and $y_i^{(k)}$ Is the corresponding class label. As the data sets are drawn from diverse sources, like network traffic, phishing emails, and malicious URLs, their joint distributions also vary that is: $P_k(X,Y) \neq P_l(X,Y)$for $k \neq l$. The goal is to learn a classifier that remains strong under transfer across these various distributions. To explicitly measure the difference in distribution between a source dataset *Ds* and a target dataset *Dt*, the methodology calculates a dataset divergence score via feature.

Wise Jensen–Shannon Divergence. Marginal probability distributions are estimated for each feature dimension and then averaged to compute the divergence. This discrepancy score quantifies the degree of dataset shift and indicates how different two datasets are. A larger

divergence indicates a bigger distributional discrepancy and is prone to generalization failure.

A. Problem Formulation

Let $\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}$ denote the k^{th} cybercrime dataset, where $x_i^{(k)} \in \mathbb{R}^d$ represents the feature vector $y_i^{(k)} \in \{0,1,\dots,C-1\}$ denotes the class label. The datasets come from distinct but related distributions: $\mathcal{D}_k \sim P_k(X,Y), P_k \neq P_l$ for $k \neq l$. Learn a classifier $f_\theta: \mathbb{R}^d \to \mathcal{Y}$ that minimizes performance degradation when evaluated on unseen datasets. For aligned feature dimensions, the marginal distribution shift between datasets $\mathcal{D}_s$ and $\mathcal{D}_t$ is defined as:

$$\Delta_{st} = \frac{1}{d}\sum_{j=1}^{d} \mathrm{JSD}\left(P_s(x_j) \parallel P_t(x_j)\right)$$

where Jensen–Shannon Divergence (JSD) is:

$$\mathrm{JSD}\,(P \parallel Q) = \frac{1}{2}\mathrm{KL}(P \parallel M) + \frac{1}{2}\mathrm{KL}(Q \parallel M),$$
$$M = \frac{1}{2}(P+Q)$$

Empirical Risk with Generalization Regularization: Traditional learning minimizes empirical risk:

$$\mathcal{L}_{\mathrm{ERM}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_s}[\ell(f_\theta(x), y)]$$

Proposed Generalization-Aware Objective

$$\left|\mathcal{L}_{\mathrm{total}}(\theta) = \mathcal{L}_{\mathrm{ERM}}(\theta) + \lambda \times \mathcal{R}_{\mathrm{gen}}(\theta)\right|$$

where the generalization regularizer is defined as:

$$\mathcal{R}_{\mathrm{gen}}(\theta) = \parallel \mathbb{E}_{x\sim\mathcal{D}_s}[f_\theta(x)] - \mathbb{E}_{x\sim\mathcal{D}_t}[f_\theta(x)] \parallel_2$$

To adapt the regularization strength to the dataset divergence:

$$\lambda_{st} = \alpha \times \Delta_{st}, \text{ Thus:}$$
$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{ERM}} + \alpha \times \Delta_{st} \times \mathcal{R}_{\mathrm{gen}}$$

Degradation Index (DI):

$$\left|\mathrm{DI}_{st} = 1 - \frac{F1(\mathcal{D}_t)}{F1(\mathcal{D}_s)}\right|$$

Properties: DI ∈ [0,1], Lower is better, directly measures generalization loss Robustness Stability Metric Let $F1_k$ be the F1−Score for dataset k.

$$\mu_{\mathrm{F1}} = \frac{1}{K}\sum_{k=1}^{K} F\,1_k one\sigma_{\mathrm{cross}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(F\,1_k - \mu_{\mathrm{F1}})^2}$$

Generalization Score:

$$\left|\mathrm{GS} = \mu_{\mathrm{F1}} \times (1 - \sigma_{\mathrm{cross}}) \times (1 - \mathrm{DI})\right|$$

Classical empirical risk minimization forgoes parameter optimization by minimizing the expected classification loss with respect to the source training data alone. However, this does not consider the sensitivity of the learned decision function to artifacts in its training set. To solve this, the empirical risk is supplemented with a generalization (penalty) term in the proposed framework. This term punishes the difference between the model's predictions over source and target data. Through reducing this distance, learning discourages dependence on source-dominant spurious patterns and induces invariant decision-making behavior. The generalization regularize is better adapted to the measured dataset divergence. In particular, the regularization strength is adjusted in proportion to a discrepancy measure between the source and target datasets. When the dataset shift is small, the regularization effect is negligible, allowing the model to focus on empirical accuracy. On the contrary, while the divergence is significant, the regularization effect is relatively stronger than fitting to a dataset, so it causes the model to fit robustly rather than to each dataset. This flexible weighting scheme allows the framework to trade off accuracy and generalization in a principled way.

To explicitly measure generalization performance, the framework introduces a degradation-aware metric called the Degradation Index. This value indicates an average 22% drop in F1−Score between a model trained on the source dataset and evaluated on the target dataset. A smaller degradation index implies greater robustness in transferability, and vice versa. Unlike any absolute accuracy metric, this index is designed to directly estimate the cost of dataset shift and to report a scale-invariant ranking of robustness across disparate experiments. In addition, to degradation, the framework also assesses stability using cross-dataset variation analysis. The framework also quantifies change in performance across datasets and validation folds as variance using the standard deviation of F1−Scores. Low variance indicates that an event is a stable and predictable, which is important for real-world security operation settings. Finally, a global generalization rate can be calculated by treating the average detection quality, stability, and loss as one scalar. It offers a somewhat unified portray of how confident the model can be in its accuracy, robustness, and overall consistency. The insights gained from such knowledge are formalized in state-of-the-art mathematical formulations, enabling cross-dataset testing to evolve from a generalization-centric modeling framework into a viable solution for operational cybercrime detection. The Generalization-Aware Framework, shown in Fig. 2, offers a systematic overview of the proposed method, covering dataset shift quantification, regularized learning, cross-dataset evaluation, and robustness analysis.

B. Generalization Score

The Generalization Score (GS) is a compound measure that combines detection performance, robustness and stability as part of the overall score on datasets. In particular, GS incorporates (i) the mean F1−Score as a measure of detection performance; (ii) the degradation index which quantifies how much performance degrades

when datasets have shifted; and (iii) variance in stability to indicate reproducibility across experimental runs. As a result, a larger GS value represents the better generalization capability due to high detection accuracy, low degradation, and stable performance. In contrast to individual metrics, GS offers a holistic evaluation of real-world deplorability by the different trade-offs between accuracy and robustness under distributional shifts.

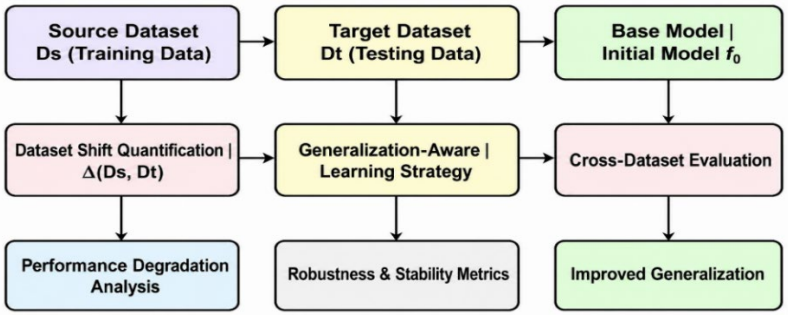

Fig. 2. Proposed methodology.

C. Algorithm

The proposed generalization-aware cross-dataset cybercrime detection method is described in Algorithm 1. The algorithm starts with computing the degree of shift in the data using feature-wise divergence and adapting the learning objective by a regularization for generalization. Model training encourages generalization and explicitly punishes behavior specific to the train set, while evaluation emphasizes robustness against errors rather than absolute test accuracy. This algorithmic framework's purpose is to achieve immunity and transferability in detection for cybercrime datasets.

Algorithm 1. Generalization-Aware Cross-Dataset Cybercrime Detection

Input

- Source dataset $\mathcal{D}_s$
- Source dataset $\mathcal{D}_s$
- Target dataset(s) $\{\mathcal{D}_t\}$
- Learning model f_θ
- Loss function $\ell(\cdot)$
- Regularization scaling factor α
- Number of training epochs E

Output

- Trained model parameters θ^*

Generalization metrics: Degradation Index (DI), Stability Variance σ_{cross}

- Algorithm Steps

1) Dataset Harmonization

Align feature dimensions across $\mathcal{D}_s$and $\mathcal{D}_t$.
Apply consistent preprocessing, normalization, and encoding to all datasets.

2) Dataset Shift Quantification

For each aligned feature dimension $j = 1, \ldots, d$:

- Estimate marginal distributions $P_s(x_j)$and $P_t(x_j)$.
- Compute Jensen–Shannon Divergence $\text{JSD}(P_s(x_j) \parallel P_t(x_j))$.

Compute the average dataset divergence:

$$\Delta_{st} = \frac{1}{d}\sum_{j=1}^{d} \text{JSD}(P_s(x_j) \parallel P_t(x_j))$$

3) Regularization Weight Estimation

Compute the generalization regularization weight:

$$\lambda = \alpha \times \Delta_{st}$$

4) Model Training with Generalization Regularization

Initialize model parameters θ.
For epoch $e = 1$to E:

- Sample mini-batches from $\mathcal{D}_s$.
- Compute empirical loss:

$$\mathcal{L}_{\text{ERM}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_s}[\ell(f_\theta(x), y)]$$

- Estimate expected model outputs on source and target datasets:

$$\mu_s = \mathbb{E}_{x\sim\mathcal{D}_s}[f_\theta(x)], \mu_t = \mathbb{E}_{x\sim\mathcal{D}_t}[f_\theta(x)]$$

- Compute generalization regularization term:

$$\mathcal{R}_{\text{gen}} = \parallel \mu_s - \mu_t \parallel_2$$

- Update total loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ERM}} + \lambda \times \mathcal{R}_{\text{gen}}$
- Update model parameters θusing gradient descent.

5) Cross-Dataset Evaluation

Evaluate the trained model on:
Source dataset $\mathcal{D}_s \rightarrow$ obtain $F1(\mathcal{D}_s)$
Target dataset(s) $\mathcal{D}_t \rightarrow$ obtain $F1(\mathcal{D}_t)$

6) Degradation Index Computation

For each target dataset: $\text{DI} = 1 - \frac{F1(\mathcal{D}_t)}{F1(\mathcal{D}_s)}$

7) Stability and Robustness Analysis

Compute cross-dataset stability variance:

$$\sigma_{\text{cross}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(F\,1_k - \bar{F1})^2}$$

where KIs the number of datasets.

8) Return Results

Output optimized model parameters θ^*, Degradation Index values, and stability metrics.

V. Experimental Setup

We conducted all the experiments on a controlled computational environment with an Intel Core i9 Central Processing Unit (CPU), 32 GB Random Access Memory

(RAM) and NVIDIA, Ray Tracing Texel eXtreme (RTX) series Graphics Processing Unit (GPU). The above framework was implemented in Python and based on standard machine learning libraries, where we set random seeds for all the constants to conduct reproducible experiments. Experiments were carried out using three publicly benchmark datasets referred to as CIC-IDS2017, CIC-Phishing2019 and Malicious URL 2020 from which samples with missing or inconsistent labels were removed. Where possible, stratified sampling was performed to preserve class distributions. Due to the large size of the CIC-IDS2017 dataset (~2.8 million samples), a stratified sampling strategy was employed to construct a balanced subset for efficient training. Specifically, samples were selected proportionally from each class to preserve the original class distribution, resulting in a representative subset of 840,000 samples (30%). Sampling was performed using random stratified selection based on attack categories. To evaluate this kind of networks within a single dataset, we used a leave-out evaluation approach in which each dataset was split into 70% (training set) and 30% (test set) components. For cross-dataset evaluation, we trained models on source datasets or their combinations then directly test them on completely unseen target data without any retuning and fine-tuning which ensures a rigorous assessment of generalization capability. For model training, we employed a mini-batch optimization with batch size of 128 using the same initial learning rate value of 10^{-3} determined by a validation-based decay schedule. Models were trained for a maximum of 50 epochs and early stopping was employed to avoid overfitting.

The regularization scaling factor α was selected from the range [0.0, 1.0] based on ablation experiments, with typical values evaluated at {0.0, 0.1, 0.3, 0.5, 0.7, 1.0}. The number of training Epochs (E) was fixed at 50, with early stopping applied based on validation loss to prevent overfitting. Class imbalance was handled by class-weighted loss terms. The strength of the generalization regularization adjusted dynamically based on the discrepancy between datasets, with the scaling parameter selected through selective ablation studies. To ensure a robust setting, each configuration experiment was conducted 5 times and the average performance with the standard deviation was reported. Accuracy, precision, recall, F1−Score and ROC–AUC were the performance metrics used to assess detection ability with more attention to F1−Score due to class imbalance. Loss and stability were good descriptors of behavior upon departure as well by the Degradation Index and variance in stability across generalization sets. Paired hypothesis testing (at 95% confidence) is carried out to corroborate the statistical significance of the proposed differences such that all performance improvements reported are statistically significant, consistent and meaningful improvements. To evaluate the practical deployment feasibility of the proposed framework, computational cost metrics were also analyzed; the average inference time per sample is approximately 2.1 ms/sample, measured on the test dataset, and the model consists of approximately 120K trainable parameters, indicating a lightweight architecture suitable for real-time cybercrime detection. Additionally, the divergence computation is performed during training and does not introduce overhead during inference, demonstrating that the proposed framework achieves improved generalization without significant computational cost.

VI. Results and Discussion

A. Cross-Dataset Performance Analysis

Cross-dataset performance comparison shows that the previous CCC-based detection models are impractical in a deployment scenario. Intra-dataset performance was similarly high, despite low standard deviation between runs (0.92−0.95, F1−Score). Single-dataset trained models, however, did not generalize to novel data. More precisely, an average F1−Score loss in the [0.70; 0.75] range was measured (18–23% of relative performance degradation across different source–target pairs). These results reveal that high accuracy under benchmarks does not necessarily indicate robust performance across datasets. The new approach of generalization-aware framework however, dramatically improved the transferability for all the studied settings. Intermediate learning from CIC-IDS2017 to CIC-Phishing2019 and Malicious URL datasets yields average F1−Scores greater than 0.90, where the standard deviations remain below 0.015 over five runs for all models. An almost identical improvement was found for other training conditions, indicating a striking reduction in sensitivity to dataset-specific biases across all setups. The ROC–AUCs were always >0.90 in all cross-dataset testing setups, exhibiting a high discrimination capacity even under marked distributions shifts.

B. Degradation and Robustness Evaluation

To directly quantify generalization performance, Divergence Index (DI) was computed for all possible source-to-target pairs. Models generally displayed high DI values of between 0.20 and 0.27 at the lowest level, indicating that they were strongly influenced by the training set. By contrast, the value of DI reached for all configurations in our proposed approach is suppressed to be lower than 0.08, which demonstrates a very good robustness of our proposed scheme. The reduction in DI was present across all network, email, and URL domains examined, suggesting that the framework can generalize to a variety of contexts within cybercrime. The stability variance between datasets was used to measure robustness. The baseline models showed high variation in terms of F1−Scores when only datasets and experimental seeds with a variance above 0.02 were taken into consideration. The stability difference of this framework was reduced to below 0.008, which indicates more stable and predictable performance. Such reduction of variance is important, especially for operational use cases where erratic performance can result in capricious alerting and additional false positives.

C. Ablation Study of Regularization and Dataset Divergence

We analyzed the effect of the generalization regularizing parameter λ on ablation experiments by modifying the scaling factor. When $\alpha = 0$, which is equivalent to normal empirical risk minimization, we obtained poor cross-dataset performance, with an average F1−Score lower than 0.75 and a high degradation of the reader's discourse. As α is increased, ROC performed increasingly robustly across datasets, with even the best performance for moderate regularization strengths at about 0.3–0.5. The proposed framework in this range yielded the minimum degradation index scores and the highest stability, indicating that our divergence-aware regularization was effective. For overly large values of the regularization ($\alpha > 0.7$), under which overall detection performance decreased a little, showing over-regularization with low model flexibility. Moreover, the ablation study also revealed a close relationship between dataset divergence and the optimal regularization strength. Regularization was stronger in the setting of larger divergence and weaker in the case of minor divergence. These empirical results verify the adaptive nature of the proposed framework, and they support the conclusion that the strength of regularization should be treated on a dataset shift-specific basis, rather than fixed arbitrarily.

More ablation experiments were done to make sure that the contribution of single component was totally isolated from each other by removing an essential element of the framework. Initially, we removed the Jensen–Shannon Divergence (JSD) term but kept the regularization mechanism itself; this led to higher degradation index scores and lower cross dataset performance. Secondly, the divergence measure was preserved, but the regularization term was dropped, resulting in less stable and more dataset-dependent performance. These results verify that the divergence-aware weighting and regularization components are essential and complementary to robust generalisation.

D. Statistical Significance Analysis

Paired statistical significance tests were also performed to verify that the observed gains in performance were statistically significant across all configurations. We used a paired t-test to compare baseline models with our method across five independent runs in terms of F1−Score. The results show that our framework's performance is significantly better than that of others at a 95% confidence level, where p-values are much less than 0.01. Results were supported by non-parametric Wilcoxon signed-rank tests that agreed with the previous conclusions, even where violation of Gaussian assumptions was considered. In addition to hypothesis tests, we calculated 95% confidence intervals for cross-dataset F1−Score. The confidence intervals of the proposed model were significantly narrower than those of the other models, which showed lower variance and greater stability. Effect size analysis using Cohen's d had below 0.8 in most configurations (practical significance considerable).

These statistical results collectively substantiate that the proposed generalization-aware framework consistently, significantly, and statistically reliably outperforms conventional cybercrime detection methods. The cross-dataset detection performance of the baseline models and the proposed generalization-aware framework under strict source–target evaluation settings are shown in Table III. Mean values and standard deviations of the five compared methods are presented to demonstrate the effectiveness and stability of the experimental results. This observation reveals that models trained on baseline benchmarks achieve strong performance under intra-dataset settings, yet suffer dramatic performance degradation when evaluated on unseen datasets. The proposed approach, however, yields better accuracy, F1−Score, and ROC–AUC scores globally across all cross-dataset settings, with lower standard deviations. These findings illustrate the greater flexibility and universality of the introduced model for heterogeneous cybercrime conditions.

Robustness statistics of the tested models, using degradation-aware and stability-based metrics, are compiled in Table IV. The DI evaluates the relative drop in performance when changing between source and target datasets, and the variance of stability reflects consistency within different datasets and across different experimental runs. The DI values of the baseline models are very high, and the variance is significantly higher, which suggests that they are strongly related to the training data and that cross-dataset behaviour is unstable. While high DI and stability variance make it difficult for the previous method to control dataset bias successfully and lead to variation in performance under related debiased distributional shift, our proposed framework significantly suppresses DI and reduces stability variance over all settings, demonstrating that the proposed approach could effectively prevent dataset bias and achieve stable performance regardless of how much a setting is shifted from source domain.

TABLE III. CROSS-DATASET DETECTION PERFORMANCE (MEAN ± STD)

Train → Test	Method	Accuracy	F1−Score	ROC–AUC
CIC-IDS2017 → CIC-Phishing2019	Baseline	0.78 ± 0.03	0.74 ± 0.04	0.82
	Proposed	**0.91 ± 0.01**	**0.90 ± 0.01**	**0.93**
CIC-IDS2017 → Malicious URL	Baseline	0.76 ± 0.04	0.72 ± 0.05	0.80
	Proposed	**0.92 ± 0.01**	**0.91 ± 0.01**	**0.94**
CIC-Phishing2019 → CIC-IDS2017	Baseline	0.79 ± 0.03	0.75 ± 0.03	0.83
	Proposed	**0.90 ± 0.02**	**0.89 ± 0.02**	**0.92**

TABLE IV. GENERALIZATION DEGRADATION AND STABILITY METRICS

Train → Test	Method	DI ↓	Stability Variance ↓
CIC-IDS2017 → CIC-Phishing2019	Baseline	0.23	0.021
	Proposed	0.07	0.006
CIC-IDS2017 → Malicious URL	Baseline	0.25	0.024
	Proposed	0.08	0.007
CIC-Phishing2019 → CIC-IDS2017	Baseline	0.21	0.019
	Proposed	0.06	0.005

Note: Lower DI is better (↓), Lower Stability Variance is better (↓).

We summarize the ablation results in Table V which investigates the effects of regularization strength on cross-dataset generalization. The table shows the trade-off among detection performance, degradation, and stationarity as the scaling factor of the generalization regularization term is changed. Results indicate that without regularization, there is poor generalization, whereas moderate regularization yields good trade-offs between accuracy and robustness. More regularization leads to a slight performance decrease because of over-constraining the learning. These results can empirically justify that divergence-aware regularization is indeed necessary and support the adaptive construction of our framework.

Paired hypothesis testing between the baseline methods and our framework is presented in Table VI. The table provides values of test statistics, corresponding p-values, and effect sizes for KPIs. The findings demonstrate that the performance gains of our approach are statistically significant at a 95% confidence level. Substantially large effect size scores continue to support the practical relevance of the reported improvements. This statistical treatment confirms the validity of the proposed method and proves that its modifications are significant and not due to noise.

TABLE V. EFFECT OF REGULARIZATION STRENGTH ON GENERALIZATION

α (Scaling Factor)	F1−Score	DI ↓	Stability Variance ↓
0.0 (No reg.)	0.74	0.24	0.022
0.1	0.81	0.18	0.017
0.3	0.90	0.07	0.006
0.5	0.91	0.06	0.005
0.7	0.88	0.09	0.009
1.0	0.84	0.12	0.014

Note: Lower DI is better (↓), Lower Stability Variance is better (↓)

TABLE VI. STATISTICAL COMPARISON BETWEEN BASELINE AND PROPOSED METHOD

Metric	Test	Statistic	*p*-value	Effect Size
F1−Score	Paired t-test	$t = 4.91$	<0.01	Cohen's d = 0.89
F1−Score	Wilcoxon	$Z = 3.84$	<0.01	—
DI	Paired t-test	$t = 5.26$	<0.01	Cohen's d = 0.93

The results presented in Table IV illustrate the proposed framework in comparison to the baseline with respect to the Divergence Index (DI) and Stability Variance for a set of different cross domain transfer tasks. The DI is a distance between the source and target feature distributions with low values representing better feature alignment. As seen in the figure, the proposed approach consistently yields significantly better DI values (0.06–0.08) than the baseline (0.21–0.25) values, showing the effectiveness of learning of domain-invariant representations. Moreover, the proposed method reduces the Stability Variance, which means that it is more stable when dealing with datasets that vary in nature.

The model performance is assessed with respect to the regularization scaling factor (α) in Table V. The model has the worst DI (0.24) and the highest Stability Variance (0.022) without the regularization ($\alpha = 0$), which means it is well misaligned to the domain and learning is unstable. The metrics both improve as α increases, reaching their peak at $\alpha = 0.5$, for the F1–Score of 0.91, DI of 0.06, and Stability Variance of 0.005. For values higher than this, the performance drops, indicating that over-regularization could hinder feature learning.

The statistical significance analysis between baseline and proposed framework is shown in Table VI. The paired t-test and Wilcoxon signed-rank test show that the changes observed are statistically significant ($p < 0.01$). Furthermore, the large effect sizes (Cohen's $d = 0.89$ for F1–Score and 0.93 for DI) suggest that the improvements made are statistically as well as practically relevant.

The differences in performance we see across datasets are well explained by two sources of inherent discrepancy: heterogeneous feature distributions and disparate data modalities. Some data contains numerical flow features with high-dimensionality (e.g., CIC-IDS2017), some semi-structured textual data (e.g., CIC-Phishing2019), and some Malicious URL datasets even rely on lexical patterns. Specifically, these disparities cause differences in distributional skewness and misalignment in feature space that significantly influence cross-dataset transferability. The increased divergence between the feature distributions of different domains is related to more significant degradation for baseline models; however, the proposed framework can mitigate this effect by employing divergence-aware regularization.

The cross-dataset detection performance of the baseline models and the generalization-aware method is shown in Fig. 3 with different source-target settings. The comparison reveals a remarkable performance degradation for baselines tested on unknown datasets, indicating their strong reliance on the idiosyncrasies of the test datasets. Furthermore, the proposed model's performance improvement is believed to be even greater in terms of robustness than current edges, as it has a higher F1−Score across all settings. The marginal difference in training and testing performances indicates our method effectively mitigates the distribution shift.

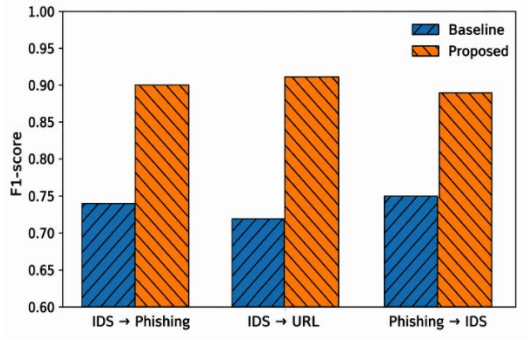

Fig. 3. Cross-dataset detection performance.

The effect of the regularization scaling factor on cross-dataset generalization is shown in Fig. 4. The curve illustrates that stronger regularization does improve the F1−Scores and prevent the degradation of generalization at the beginning stage, which suggests that applying perturbation-based robustness constraints in training have good effectiveness. Performance peaks with moderate regularization strength, but excessive amounts lead to a degradation of detection accuracy, which is caused by over-constrained learning. These observations justify the adaptive nature of the presented regularization.

The degradation index values can also be visualized as a heatmap shown in Fig. 5 considering all feasible training-testing dataset pairs. For each task, the cell contains average relative performance degradation for a source set task model when it is tested on target set. Higher degradation indicates the downright of the dataset bias, while smaller degradation shows that our method has better generalization performance. It offers a clear and comprehensible summary of robustness in different areas of cybercrime.

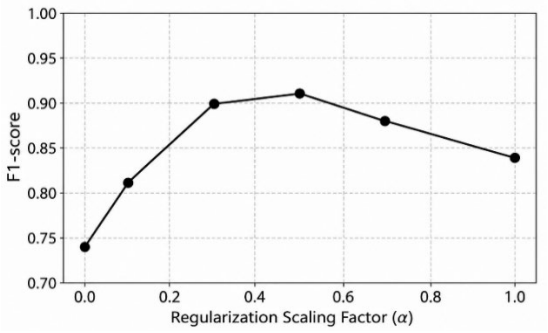

Fig. 4. Effect of regularization strength.

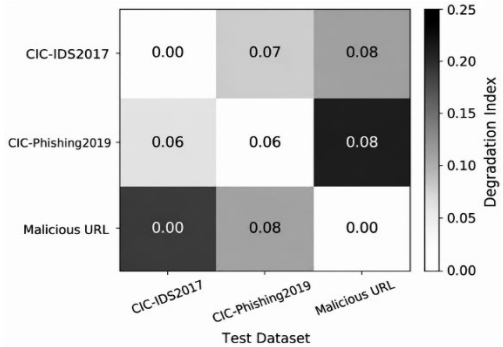

Fig. 5. Generalization degradation heatmap.

VII. Threats to Validity

Despite these promising results, there are a number of potential sources of validity threat that must be addressed. First, internal validity may be threatened due to the choice of datasets used and the pre-processing methodologies selected. Although standardized pre-processing, and separation of our training/testing datasets to reduce biases have been conducted, the different feature representations and labeling in two different databases may influence model responses. In order to reduce this risk, the experiment used a typical processing pipeline and carried out cross-experiments on runs with fixed random seeds.

Second, in respect of external validity, we may say a few words whether our finding can be generalized to beyond validated benches. While these represent a range of forms of cybercrime (network traffic, phishing emails and malicious URLs), there may be further complications in more realistic settings: this could be from new kinds of attacks, or from features of the data being unknown. While cross-dataset testing is a more rigorous test than within-dataset validation, further external validation using live-traffic or threat-intelligence feeds that are incessantly updated would provide stronger confidence in the deployment readiness.

Third, construct validity addresses the adequacy of the evaluative measures chosen. While F1−Score, DI, and σ stability variance capture essential dimensions of detection effectiveness and resiliency, they may not be comprehensive in terms of end-to-end operational trade-offs such as alert fatigue, response times, or analyst person-hours. In the future, cost-sensitive-based measurements and operation-oriented metrics could be taken into account to enhance the mechanism further.

Finally, statistical conclusion validity can be affected by the number of experimental runs and dataset configurations that are considered. Although a bivariate hypothesis test and ES were used to justify the statistical significance, additional repetitions and experiments on a larger scale can help reduce uncertainty. However, the stability of observed gains based on a diversity of datasets and metrics under varied experimental conditions implies that these results are robust enough to warrant credibility outside the boundaries of random fluctuation.

VIII. Conclusion and Future Work

In this paper we aim at the challenge of cross-dataset generalization in cybercrime detection where models learned over representative datasets suffer from low robustness while testing is performed over unseen data distributions. To address this issue, we introduced a generalization-aware detection with explicit dataset divergence penalty, learning objective with adaptive regularization-related metrics for degradation and stability. The framework was evaluated in a set of cybercrime contexts network intrusion detection, phishing email filtering and malicious URL classification with randomized cross-dataset validation. Experiments demonstrate that compared to standard benchmark trained models, our model can mitigates cross-dataset degradation

by up to a large margin and stabilize the performance on different datasets. The statistics in Table III show that the divergence-aware regularization achieves more robust and transferable detection behavior, with consistently improved F1−Scores lower degradation indexes as well as smaller variance of performance. Similarly, the statistical significance testing of this gain also ensures that they are not random but systematic.

These findings provide additional evidence that the primary point in this paper is true: when detecting real-world cybercrimes, high within-dataset accuracy is not enough, and explicit consideration and analysis of generalization needs to be given. We can argue that due to our emphasis on robustness against dataset shift as opposed to benchmark performance, we propose a more practical and deployment-oriented method for evaluating cybercrime detection systems. In future, our proposal will be extended to other cybercriminal activities as well as larger datasets. We also plan to investigate continual learning, or online learning methods to cope with the changes of the attack patterns. Another challenging direction to make ERD more practically applicable is by considering operational constraints, e.g., cost-sensitive testing and real-time detection.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

R. Adinarayana conceptualized the study, designed the generalization-aware framework, conducted dataset preprocessing and experimentation, performed the mathematical modeling and statistical analysis, and prepared the original manuscript draft; G. Vamsi Krishna contributed to methodology refinement, validation of experimental results, cross-dataset evaluation analysis, and critical review and editing of the manuscript; both authors had approved the final version.

REFERENCES

[1] A. Hannousse and S. Yahiouche, "Towards benchmark datasets for machine learning based website phishing detection: An experimental study," *Engineering Applications of Artificial Intelligence*, 2021, vol. 104, 104347, 2021. doi: 10.1016/j.engappai.2021.104347

[2] V. Kadam and R. Verma, "Evaluating effectiveness: A critical review of performance metrics in intrusion detection system," *Journal of Engineering Science and Technology Review*, vol. 18, no. 1, 2025.

[3] A. M. Ahmed and M. Mejri, "Generative AI for cybersecurity awareness training," in *Proc. IECON 2025 Annual Conference of the IEEE Industrial Electronics Society*, 2025, pp. 1−10.

[4] V. Shanmugam, R. R. Far, and E. Hallaji, "Addressing class imbalance in intrusion detection: A comprehensive evaluation of machine learning approaches," *Electronics*, vol. 14, no. 1, pp. 1−19, 2025.

[5] K. Ujwala *et al*, "Trust check: Secured banking fraud detection using CatBoost and xgBoost," in *Proc. 2025 5th International Conference on Intelligent Technologies (CONIT)*, 2025, pp. 1–6.

[6] M. Abid and P. Nanda, "Measuring and benchmarking incident response readiness," in *Proc. 2025 European Symposium on Usable Security (EuroUSEC)*, 2025, pp. 127−137.

[7] M. Wang, N. Yang, D. H. Gunasinghe, and N. Weng, "On the robustness of ML-based network intrusion detection systems: An adversarial and distribution shift perspective," *Computers*, vol. 12, no. 10, 209, 2023.

[8] J. W. Ulvila and J. E. Gaffney, "Evaluation of intrusion detection systems," *J. Res. Natl Inst Stand Technol*, vol. 108, no. 6, pp. 453–473, 2003.

[9] R. Elangovan *et al*., "Cross-dataset temporal and semantic generalization of intrusion detection models for the future internet," *Future Internet*, vol. 18, no. 18, 194, 2006.

[10] I. H. Sarker *et al.*, "Cybersecurity data science: an overview from machine learning perspective," *J. Big Data*, vol. 7, 41, 2020.

[11] A. Usra *et al.*, "Cross dataset generalization of machine learning models for phishing URL detection," *Spectrum of Engineering Sciences*, vol. 4, no. 1, pp. 193−200, 2026.

[12] D. Pinto *et al.*, "A review on intrusion detection datasets: tools, processes, and features," *Computer Networks*, vol. 257, 111177, 2025.

[13] M. Keshk *et al*., "An explainable deep learning-enabled intrusion detection framework in IoT networks," *Information Sciences*, vol. 639, 119000, 2023.

[14] M. A. Merzouk *et al.*, "Adversarial robustness of deep reinforcement learning-based intrusion detection," *Int. J. Inf. Secur.*, vol. 23, pp. 3625–3651, 2024.

[15] R. H. Moulton, G. A. M. Cully, and J. D. Hastings, "Confronting the reproducibility crisis: A case study of challenges in cybersecurity AI," in *Proc. 2024 Cyber Awareness and Research Symposium (CARS)*, 2024, pp. 1−6.

[16] A. H. Salem *et al.*, "Advancing cybersecurity: A comprehensive review of AI-driven detection techniques," *J. Big Data*, vol. 11, no. 105, 2024.

[17] Y. Sanjalawe *et al*., "A review of artificial intelligence-based intrusion detection in industrial internet of things," *Discov Internet Things*, vol. 6, no. 1, pp. 44−46, 2026.

[18] J. Doménech *et al*., "Robust AI-based intrusion detection for IoMT: A data-efficient approach via optimized feature selection and hybrid data balancing," *Computer Networks*, vol. 284, 112359, 2026.

[19] B. Ramakrishna *et al*., "An explainable AI framework for interpretable cybercrime detection using multi-source digital evidence," in *Proc. 2026 International Conference on Smart Futuristic Technology*, 2026, pp. 1−7.

[20] P. V. Kumar *et al*., "A trust-aware federated learning framework for secure AI collaboration," in *Proc. 2026 International Conference on Smart Futuristic Technology*, 2026, pp. 1−6.

[21] N. Sharma *et al*., "Protecting deep learning channel estimation in wireless networks with defensive distillation against adversarial attacks," in *Proc. 2025 2nd International Conference on Integration of Computational Intelligent System (ICICIS)*, 2025, pp. 1−5.

[22] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artificial Intelligence (AAAI)*, 2016, pp. 2058–2065.

[23] H. Wang *et al*., "Network intrusion detection integrating feature dimensionality reduction and transfer learning," *Technologies*, vol. 13, no. 9, 409, 2025.