

Article

Leveraging Cross-Domain Transfer Learning for Enhanced Multi-Protocol Network Intrusion Detection

Oluwaseyi Oladejo  and Ahmed Abdelmoamen Ahmed * 

Department of Computer Science, Prairie View A&M University, Prairie View, TX 77446, USA;
ooladejo@pvamu.edu

* Correspondence: amahmed@pvamu.edu

Abstract

The exponential growth of cyber threats in modern digital infrastructure demands advanced detection systems that adapt to evolving attack patterns. Traditional cybersecurity approaches struggle with dynamic threats, requiring extensive labeled datasets and re-training for each new category. This paper presents a comprehensive transfer learning framework for cybersecurity threat detection, leveraging the CICIoMT dataset as a benchmark to enhance detection capabilities across heterogeneous cybersecurity environments. We propose a machine learning (ML)-enabled framework that employs systematic feature alignment, hybrid class balancing, and multi-algorithm evaluation using machine learning models, including Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting, and XGBoost. The proposed approach addresses the critical challenges of data scarcity and domain heterogeneity in cybersecurity by enhancing feature engineering with cybersecurity-specific features, statistical aggregations, and PCA embeddings. Extensive experimental evaluation across two target datasets (CICIoT and IoT-23) demonstrates both the exceptional successes and critical limitations of cross-domain transfer learning in cybersecurity. The framework achieved outstanding performance on domain-compatible datasets, with RF reaching 99.0% accuracy on CICIoT, Gradient Boosting achieving 98.9%, and XGBoost delivering 98.4%, demonstrating exceptional knowledge transfer from medical IoT to smart home IoT environments. However, transfer learning to IoT-23 was unsuccessful (50% accuracy, equivalent to random guessing), revealing that feature domain difference, where identical attack labels encode fundamentally different behavioral patterns, prevents effective knowledge transfer despite nominal class overlap. This research makes significant advances in adaptive cybersecurity systems by providing a rigorous evaluation of both the successes and limitations of transfer learning. This work demonstrates that ensemble methods (RF, XGBoost, and Gradient Boosting) achieve superior cross-domain performance compared with neural networks on compatible domains, while also revealing fundamental challenges when the source and target domains differ in their feature spaces.



Academic Editor: Paolo Bellavista

Received: 3 May 2026

Revised: 3 June 2026

Accepted: 4 June 2026

Published: 9 June 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: cybersecurity; transfer learning; intrusion detection; cross-domain learning

1. Introduction

Digital transformation has created new opportunities for organizations but has also introduced significant cybersecurity risks [1]. The growing reliance on cloud computing infrastructure, Internet of Things (IoT) ecosystems, and remote work technologies has substantially expanded the attack surface available to malicious actors. In parallel, the

nature of cyber threats has evolved well beyond rudimentary mechanisms such as basic viruses and unsolicited email campaigns [2]. Current threat actors increasingly leverage sophisticated methodologies, including advanced persistent threats (APTs), zero-day vulnerabilities, and artificial intelligence-driven attack strategies that adapt dynamically to defensive countermeasures. Consequently, both the frequency and severity of cyber incidents have escalated markedly. Empirical evidence underscores this trend, with Check Point Research documenting a 75% increase in global cyber-attacks during the third quarter of 2024 [3].

Conventional cybersecurity paradigms, which predominantly depend on signature-based detection and rule-driven mechanisms, are increasingly inadequate in addressing the complexity of contemporary threat landscapes [4]. While such approaches remain effective against previously identified attack vectors, they exhibit significant limitations when confronted with novel or obfuscated threats that deviate from established signatures. This challenge is further compounded by the exponential growth in data generated across organizational infrastructures, including network logs, system events, and user activity records. The sheer volume and velocity of this data render manual analysis both inefficient and impracticable, thereby exacerbating the difficulty of timely threat detection and response [5].

Machine learning (ML) has emerged as a promising paradigm for mitigating the limitations of traditional cybersecurity approaches, enabling autonomous identification of latent patterns and detection of anomalies in large-scale security datasets [6]. Despite its potential, however, the application of conventional machine learning techniques in cybersecurity is constrained by several critical challenges. Foremost among these is the reliance on substantial volumes of labeled data for effective model training. The acquisition of high-quality, accurately labeled cybersecurity datasets is both resource-intensive and time-consuming, thereby limiting the scalability of such approaches [7]. Additionally, the dynamic and rapidly evolving nature of cyber threats poses a significant obstacle. Models trained on historical data often exhibit degraded performance when exposed to emerging or previously unseen attack vectors. This issue, commonly referred to as concept drift, is particularly acute in cybersecurity contexts, where adversarial tactics and techniques continuously evolve [8].

Transfer learning offers a practical way to overcome several limitations of conventional machine learning approaches in cybersecurity [9]. By enabling models to transfer knowledge acquired from one task or domain to another related context, this approach obviates the need to develop models from first principles for each new cybersecurity application. Instead, previously learned representations and feature hierarchies can be effectively reused, thereby reducing dependence on large labeled datasets and expediting training [10]. This capability is especially advantageous in cybersecurity environments, where diverse attack vectors often share structural and behavioral characteristics. Consequently, insights derived from detecting one threat class can often be generalized to facilitate the identification of analogous or related attack patterns. Such cross-domain knowledge transfer enhances both the efficiency and adaptability of threat detection systems in the face of an increasingly complex and evolving threat landscape [11].

The impetus for this research arises from the pressing need to develop cybersecurity solutions that are both adaptive and efficient, capable of responding to the rapidly evolving threat landscape [12]. Existing detection systems often require extended periods, ranging from weeks to months, to accommodate emerging threats, leaving organizations exposed to potential exploitation during these periods. The application of transfer learning techniques offers a promising avenue to mitigate this limitation by enabling more rapid model

adaptation, while simultaneously preserving high detection accuracy and reducing the incidence of false positives that continue to affect many contemporary systems [13].

This study addresses a critical gap in transfer learning-based cybersecurity research by proposing a systematic framework for evaluating cross-domain transferability in heterogeneous IoT environments. Unlike prior work that primarily focuses on demonstrating performance improvements within compatible domains, the proposed framework integrates feature engineering, feature-space alignment, PCA-enhanced representation learning, and multi-model evaluation to investigate both successful and unsuccessful transfer scenarios. A key contribution of this work is the introduction of a practical transferability assessment perspective, demonstrating how feature-space compatibility and semantic alignment can be used to determine the feasibility of knowledge transfer before deployment. Through experiments on CICIoMT, CICIoT, and IoT-23 datasets, the study establishes domain-compatibility criteria and provides actionable guidelines for selecting appropriate transfer learning strategies in cybersecurity applications.

The Contributions of this work can be summarized as follows: (i) Proposes a modular transfer learning framework for cross-domain IoT intrusion detection incorporating feature engineering, alignment, and PCA-enhanced representation learning. (ii) Presents a systematic evaluation across heterogeneous IoT cybersecurity datasets with varying degrees of feature-space compatibility. (iii) Introduces a transferability assessment methodology based on semantic feature overlap and domain compatibility. (iv) Provides one of the few detailed analyses of both transfer learning success and failure cases in cybersecurity. (v) Establishes practical guidelines for determining when zero-shot transfer learning is likely to be effective in real-world intrusion detection deployments.

2. Related Work

Recent advances in network intrusion detection have explored diverse approaches ranging from handling class imbalance and transfer learning to real-time detection and hardware acceleration [4,9,12–18]. This section reviews key works that inform our multi-class transfer learning framework.

Detecting minority attack classes in software-defined networks (SDNs) remains a persistent challenge, primarily due to the pronounced class imbalance in real-world intrusion detection datasets. Towhid et al. [15] address this issue by developing a hierarchical framework that synergistically integrates deep learning architectures with conventional machine learning techniques. Their approach demonstrates substantial adaptability and effectiveness, achieving an average F1-score exceeding 92% and a minority-class F1-score above 84%, resulting in a 22.5% improvement over baseline methodologies. Importantly, the proposed model maintains robust performance with limited training data and exhibits strong generalization across heterogeneous datasets.

Dinesh et al. [14] advance network anomaly detection by integrating the Borderline Synthetic Minority Over-Sampling Technique (SMOTE) with Support Vector Machines (SVM), forming a hybrid framework that effectively addresses class imbalance. By augmenting the minority class before classification, the proposed SMOTE–SVM approach enhances the model's sensitivity to sophisticated, low-frequency attack patterns that are typically difficult to detect. Empirical results indicate that the method not only improves detection performance but also reduces false positive rates, thereby enabling more reliable and timely identification of genuine threats. The framework achieves an overall anomaly-detection accuracy of 92.35%, underscoring its effectiveness in cybersecurity contexts with imbalanced data.

Swetha and Merakapudi [19] propose a hybrid, unsupervised framework for cybersecurity threat detection that leverages transfer learning in conjunction with autoencoders,

Principal Component Analysis (PCA), and deep neural networks (DNNs). This integrative approach enables the extraction of compact feature representations while preserving critical structural information, thereby enhancing the model's capacity to detect anomalous patterns. The proposed model demonstrates strong empirical performance, achieving 0.95 in accuracy, 0.85 in precision, 0.94 in recall, and 0.89 in F1-score. Notably, it exhibits particular effectiveness in identifying previously unseen attack classes, addressing a well-documented limitation of traditional supervised learning techniques.

Khare and Bhutad [16] developed a real-time intrusion detection system (IDS) that emphasizes anomaly-based detection at the packet level, explicitly addressing the limitations associated with traditional signature-based approaches in contemporary network environments. Using the comprehensive CICID-2017 dataset, comprising over 1 million network packets, the proposed system was trained to capture fine-grained traffic patterns indicative of malicious activity. The IDS demonstrated exceptional performance, achieving classification accuracies of 99.964% on the training set and 99.847% on the test set, indicating high model precision and generalization. In addition to its analytical capabilities, the system features an intuitive user interface with real-time pop-up alerts, enabling network administrators to monitor and respond to packet-level events in real time.

In decentralized network environments, ensuring robust interoperability among intrusion detection systems is essential for effective, coordinated threat mitigation. Ghosh et al. [18] investigate this challenge through a federated learning framework that integrates deep learning with enhancement strategies such as bootstrapping and temporal averaging. This design enables improved detection of rare and previously unseen attack patterns by leveraging distributed knowledge across multiple nodes. The study identifies 45 transferable attack pairs, demonstrating that federated models, by enabling cross-node information sharing, can achieve broader threat recognition than models trained on isolated datasets. These results highlight the potential of federated learning to enhance detection generalization in distributed settings.

While prior studies demonstrate substantial advancements in intrusion detection, most are constrained to single-domain settings or to simplified binary classification frameworks. Such limitations restrict their applicability in complex, real-world environments characterized by heterogeneous data sources and multi-class threat scenarios. This research extends the existing body of work by conducting a systematic evaluation of multi-class transfer learning across diverse IoT domains, specifically leveraging the CICIoMT, CICIoT, and IoT-23 datasets. It further provides a comprehensive comparative analysis of five distinct algorithms, thereby offering a more holistic assessment of model performance under varying domain conditions. In addition, the study introduces domain-compatibility criteria to inform the practical deployment of transfer learning-based cybersecurity solutions.

A key distinction of this work lies in its rigorous examination of both successful transfer scenarios (e.g., CICIoT) and transfer failure cases (e.g., IoT-23). By analyzing these contrasting outcomes, the research provides critical insights into the underlying conditions and prerequisites for effective cross-domain knowledge transfer, thereby contributing to a more nuanced understanding of the applicability of transfer learning in cybersecurity contexts.

Recent research continues to demonstrate the effectiveness of ensemble learning approaches for intrusion detection. Ensemble frameworks combine the strengths of multiple classifiers to improve robustness, generalization capability, and detection accuracy across diverse attack scenarios. Recent studies have shown that ensemble methods can outperform individual machine learning models by reducing variance and improving resilience to heterogeneous traffic patterns and class imbalance. These findings further support the inclusion of ensemble-based algorithms, such as Random Forest, Gradient Boosting, and XGBoost, in cybersecurity intrusion detection research [20].

3. Design and Implementation

The proposed transfer learning framework adopts a modular, four-layer pipeline architecture for cross-domain IoT threat detection, as depicted in Figure 1. Layer 1 performs data processing, including validation, cleaning, feature engineering, class balancing, and feature alignment across source and target datasets. Layer 2 focuses on model training and evaluation, incorporating algorithm selection (e.g., SVM, XGBoost, Gradient Boosting, Random Forest, MLP), hyperparameter optimization, and performance assessment using source-domain data. Layer 3 enables transfer learning via knowledge transfer, zero-shot inference, and fine-tuning for target domains. Layer 4 provides a user-facing graphical interface (implemented with Gradio) for real-time prediction, visualization, and interaction, facilitating end-to-end deployment and usability. This design facilitates end-to-end knowledge transfer from the CICIoMT source domain to heterogeneous IoT target environments, while maintaining a well-defined separation of concerns across processing stages.

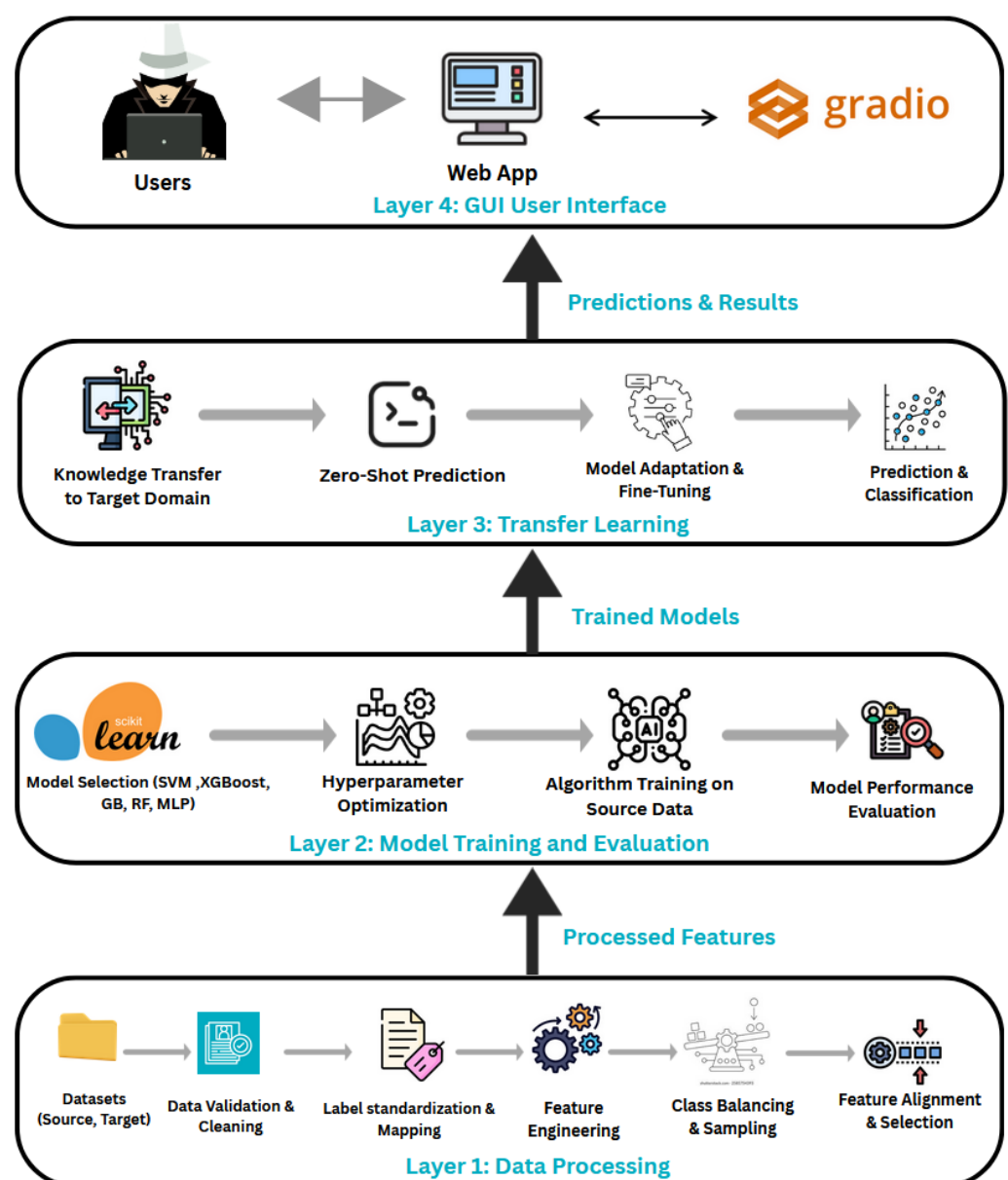


Figure 1. Overview of the proposed four-layer transfer learning framework for cross-domain IoT threat detection. The framework includes data processing, model training and evaluation, transfer learning, and user-interface components.

3.1. Dataset Collection and Preparation

Three cybersecurity datasets were selected to ensure diversity and representativeness of real-world IoT network behavior. The CICIoMT dataset was designated as the source domain, comprising network traffic generated by medical IoT devices communicating via TCP/IP over Wi-Fi and MQTT. Following preprocessing, the dataset contains 20,561 training samples distributed across four attack classes: DoS (34%), Benign (34%), Reconnaissance (28%), and Spoofing (2%). To mitigate class imbalance, a capping strategy was employed, limiting the majority classes to 10,000 samples while retaining all minority-class instances, thereby promoting balanced and robust model training.

The CICIoT dataset, representing smart home IoT environments, was selected as the first target domain. Following class filtering to retain common attack types, it comprises 1956 test samples distributed across four classes: DoS (77%), Benign (13%), Spoofing (6%), and Reconnaissance (4%). This pronounced class imbalance reflects realistic operational conditions in which certain attack types dominate network traffic. The IoT-23 dataset, focused on botnet-related activities, serves as the second target domain. After filtering and excluding underrepresented categories, it contains 4500 test samples across three classes (Benign, DoS, Reconnaissance). To isolate the effects of domain shift from class imbalance, the dataset was artificially balanced to include 1500 samples per class (33% each).

A multi-class classification framework was adopted to reflect realistic cybersecurity settings, wherein intrusion detection systems must concurrently differentiate among multiple attack types. Classes with fewer than 100 samples were excluded due to insufficient statistical power for reliable model training and evaluation.

3.2. Feature Engineering and Alignment

Feature-space heterogeneity across cybersecurity datasets constitutes a fundamental challenge for effective transfer learning. Initial analysis indicates that the CICIoT dataset shares 49 common features with the CICIoMT source domain, predominantly TCP/IP flow statistics, whereas IoT-23 exhibits only 4 shared features, reflecting substantial differences in underlying protocols and data-capture methodologies. This pronounced disparity in feature overlap provides a natural experimental setting for evaluating the extent to which semantic feature alignment influences the success of cross-domain knowledge transfer.

To mitigate feature-space incompatibility, a three-stage feature engineering pipeline was developed, as shown in Algorithm 1. First, a cybersecurity-specific feature was constructed to capture network flow characteristics indicative of attack behavior. Second, four statistical descriptors—mean, variance, standard deviation, and range—were computed across numeric flow-level features to characterize distributional properties and support anomaly detection. These engineered features were incorporated into each dataset during preprocessing. Third, after identifying shared features between the source and target domains, Principal Component Analysis (PCA) was applied to enhance representational capacity. For the CICIoT dataset, 15 principal components were generated, yielding 64 aligned features (49 shared features, including both original and engineered attributes, augmented by 15 PCA components). In contrast, for the IoT-23 dataset, 4 principal components were derived, yielding a total of 8 aligned features (4 shared features supplemented by 4 PCA components).

For numerical stability and consistent feature scaling across domains, each feature x_j is standardized as:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, \quad \mu_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad \sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \mu_j)^2} \quad (1)$$

For each flow represented by m numeric measurements $\{v_1, \dots, v_m\}$, the statistical aggregation features are computed as:

$$\mu = \frac{1}{m} \sum_{k=1}^m v_k, \quad \text{Var} = \frac{1}{m} \sum_{k=1}^m (v_k - \mu)^2, \quad \sigma = \sqrt{\text{Var}}, \quad R = \max_k(v_k) - \min_k(v_k). \quad (2)$$

After selecting the common feature set, PCA projects the aligned feature matrix $X \in \mathbb{R}^{n \times d}$ onto k principal components:

$$Z = XW_k, \quad (3)$$

where $W_k \in \mathbb{R}^{d \times k}$ contains the top- k eigenvectors of the covariance matrix of X .

Algorithm 1 Feature Engineering and Cross-Domain Feature Alignment

Require: Source dataset $D_s = (X_s, y_s)$, Target dataset $D_t = (X_t, y_t)$

Ensure: Aligned feature matrices X'_s, X'_t defined over a shared feature space

- 1: Perform data preprocessing on X_s and X_t to handle missing, noisy, and invalid values
- 2: Standardize class labels across y_s and y_t to a unified taxonomy and retain only common classes
- 3: Construct and append domain-specific cybersecurity feature(s) to both X_s and X_t
- 4: Compute statistical descriptors (e.g., mean, variance, standard deviation, range) over numerical features and append to both datasets
- 5: Determine the common feature set:

$$\mathcal{F} = \text{cols}(X_s) \cap \text{cols}(X_t)$$

- 6: Project both datasets onto the shared feature subspace:

$$X_s \leftarrow X_s[\mathcal{F}], \quad X_t \leftarrow X_t[\mathcal{F}]$$

- 7: Fit a Principal Component Analysis (PCA) model on X_s and extract k principal components
- 8: Apply the learned PCA transformation to both X_s and X_t
- 9: Construct the aligned feature representations:

$$X'_s \leftarrow [X_s \quad \text{PCA}(X_s)], \quad X'_t \leftarrow [X_t \quad \text{PCA}(X_t)]$$

Figure 2 presents the feature composition across the target datasets, highlighting the pivotal role of shared feature overlap in enabling effective cross-domain alignment, as shown in Figure 3.

The number of PCA components was selected to enrich the aligned feature representation while maintaining a compact feature space suitable for transfer learning evaluation. For CICIoT, 15 components were added to complement the relatively large set of shared features, whereas for IoT-23 the number of PCA components was limited by the substantially smaller shared feature space. The objective was not to optimize dimensionality reduction performance but rather to provide a consistent feature-augmentation mechanism across datasets.

3.3. ML Algorithms

Five ML algorithms, representing diverse learning paradigms, were evaluated to examine variability in transfer learning effectiveness. Random Forest was selected for its interpretability and inherent robustness to class imbalance. Gradient Boosting was included due to its strong predictive capability achieved through sequential ensemble learning. XGBoost was chosen for its computational efficiency via parallelization and its

proven effectiveness on structured tabular data. A SVM with a Radial Basis Function (RBF) kernel was employed to assess margin-based generalization in transfer settings. Finally, an MLP model, implemented as a feedforward neural network with ReLU activation and the Adam optimizer, was used to evaluate the performance of deep learning approaches relative to tree-based methods. All models were configured using standard hyperparameter settings appropriate for multi-class cybersecurity classification tasks.

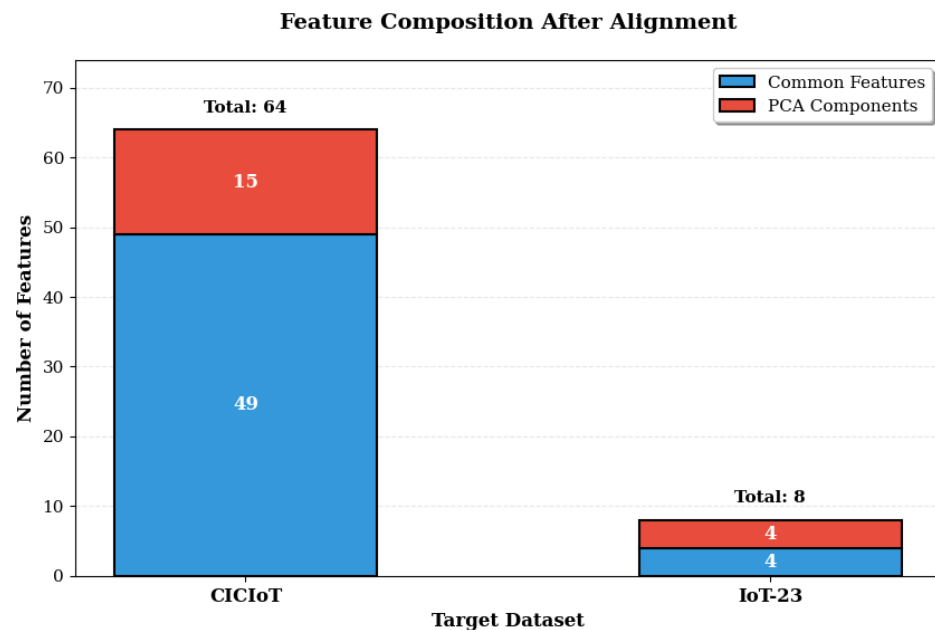


Figure 2. Feature-space composition after the alignment pipeline. CICIoT contains 49 shared features and 15 PCA-derived features, whereas IoT-23 contains 4 shared features and 4 PCA-derived features.

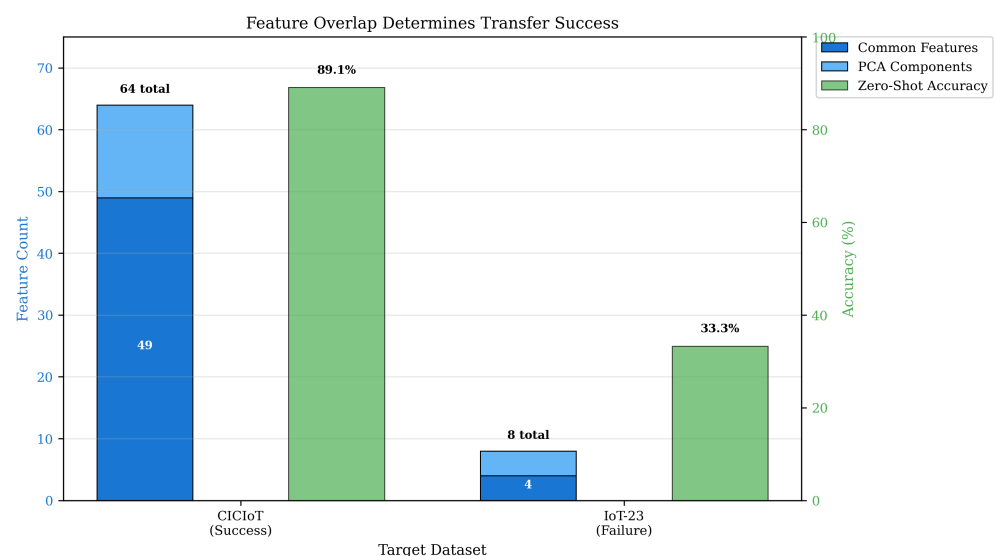


Figure 3. Comparison of aligned feature-space dimensions and zero-shot transfer learning accuracy for the CICIoT and IoT-23 target datasets.

The use of standard hyperparameter configurations was intentional and designed to provide a consistent baseline for comparing transfer learning behavior across different machine learning paradigms. By minimizing algorithm-specific optimization, the study focuses on the influence of feature-space compatibility and domain similarity rather than on maximizing the performance of individual classifiers.

Algorithm 2 outlines the training process and zero-shot transfer evaluation procedure applied uniformly across all classifiers.

Algorithm 2 Source-Domain Training and Zero-Shot Transfer Evaluation

Require: Aligned source dataset (X'_s, y_s) , aligned target dataset (X'_t, y_t) , set of models $\mathcal{M}$

Ensure: Target-domain predictions $\hat{y}_t$ and corresponding evaluation metrics

1: **for all** $m \in \mathcal{M}$ **do**

2: Train model m using the source-domain data (X'_s, y_s)

3: Generate predictions on the target domain:

$$\hat{y}_t \leftarrow m(X'_t)$$

4: Evaluate predictive performance by computing the confusion matrix and associated metrics:

 Accuracy, Precision, Recall, F1-score, False Positive Rate (FPR), and False Negative Rate (FNR)

5: **end for**

3.4. Transfer Learning Protocol

Each model was trained exclusively on the CICIoMT source domain using standard supervised learning procedures. Owing to class filtering to align the source and target domains, the training configurations varied accordingly: 20,164 samples across four classes (Benign, DoS, Reconnaissance, Spoofing) were used for transfer to CICIoT, while 19,811 samples across three classes (Benign, DoS, Reconnaissance) were used for transfer to IoT-23. Thereafter, all models were evaluated in a zero-shot transfer setting across the target domains, without any target-specific fine-tuning or domain adaptation, thereby enabling a direct assessment of cross-domain generalization.

Evaluation on the CICIoT dataset utilized 1956 test samples spanning four classes, with 64 aligned features, of which 49 were shared with the source domain. In contrast, evaluation on the IoT-23 dataset involved 4500 test samples across three classes, with only 8 aligned features, including 4 common features with the source domain. No hyperparameter optimization or model retraining was conducted on the target datasets, thereby ensuring that the evaluation reflects a strict zero-shot transfer setting and provides an unbiased assessment of cross-domain knowledge transfer performance.

3.5. Evaluation Metrics

Model performance was evaluated using standard multi-class classification metrics, including accuracy (overall classification rate), precision (class-wise positive predictive value), recall (class-wise true positive rate), and F1-score (the harmonic mean of precision and recall). In addition, security-critical metrics (i.e., false positive rate (FPR) and false negative rate (FNR)) were computed at both macro-averaged and per-class levels to assess operational suitability. This distinction is particularly important, as elevated false-positive rates can lead to alert fatigue, whereas false negatives result in undetected attacks.

Confusion matrices were further employed to provide granular, per-class performance insights, enabling analysis of which attack categories transfer effectively across domains and which exhibit degraded performance. Finally, computational efficiency was assessed by measuring training time and inference latency, thereby evaluating the practicality of deploying the proposed approach in resource-constrained cybersecurity environments.

4. Results

The proposed transfer learning framework was experimentally evaluated for classification accuracy, security-oriented metrics, computational efficiency, and algorithmic

robustness across heterogeneous cybersecurity domains. The results indicate that feature-level semantic similarity is a critical determinant of successful cross-domain knowledge transfer. In particular, ensemble-based methods exhibit superior performance when applied to domain-compatible targets. In contrast, all evaluated algorithms consistently fail in scenarios characterized by significant semantic incompatibility between source and target domains.

4.1. Overall Transfer Learning Performance

Table 1 reports the performance of five machine learning algorithms evaluated under a zero-shot transfer setting from the CICIoMT source domain (20,164 training samples for CICIoT transfer; 19,811 for IoT-23 transfer) to two target domains: CICIoT (1956 test samples, four classes) and IoT-23 (4500 test samples, three classes). The results reveal substantial differences in transfer learning performance across the two target domains. Models achieved strong performance on CICIoT but performed poorly on IoT-23, indicating that cross-domain transferability varies considerably across cybersecurity environments.

Table 1. Overall Transfer Learning Performance Across Target Domains.

Algorithm	CICIoT (4-Class)			IoT-23 (3-Class)		
	Accuracy	FPR	FNR	Accuracy	FPR	FNR
Random Forest	89.1%	3.1%	35.7%	33.3%	33.3%	66.7%
Gradient Boosting	87.9%	3.6%	37.1%	33.3%	33.3%	66.7%
XGBoost	86.6%	6.5%	43.3%	33.3%	33.3%	66.7%
MLP	73.3%	8.1%	44.3%	33.3%	33.3%	66.7%
SVM	11.8%	27.8%	73.5%	33.3%	33.3%	66.7%

As shown in Figure 4, RF achieved the highest accuracy on the CICIoT dataset (89.1%) with a notably low false-positive rate (3.1%), indicating effective knowledge transfer despite substantial feature overlap. Gradient Boosting (87.9%) and XGBoost (86.6%) demonstrated comparably strong performance, further underscoring the robustness of ensemble-based methods in cross-domain settings. MLP exhibited moderate effectiveness (73.3%), whereas SVM performed poorly (11.8%), reflecting a marked performance degradation when feature spaces lack sufficient semantic alignment.



Figure 4. Zero-shot transfer learning performance on the CICIoT dataset: (left) classification accuracy and (right) false-positive rate for the evaluated models.

4.2. Per-Algorithm Performance Analysis

4.2.1. CICIoT Target Domain: Ensemble Methods

Tables 2–5 provide detailed per-algorithm classification results on the CICIoT dataset, highlighting algorithm-specific performance characteristics and revealing distinct transfer learning patterns.

Table 2. Random Forest Classification Performance on the CICIoT Dataset.

Class	Precision	Recall	F1-Score	Support
Benign	0.62	0.92	0.74	251
DoS	1.00	0.97	0.98	1500
Reconnaissance	0.48	0.55	0.51	86
Spoofing	0.49	0.14	0.22	119
Overall: Accuracy = 89.1%, FPR = 3.1%, FNR = 35.7%				

RF achieved the highest transfer learning performance (89.1% accuracy), exhibiting near-perfect detection of DoS attacks (precision: 100%, recall: 97%). The model also demonstrated strong recall for Benign traffic (92%) with comparatively lower precision (62%), indicating a recall-oriented classification tendency that prioritizes minimizing false negatives. However, performance on the Spoofing class was notably limited, with a recall of only 14% (17 correctly identified instances out of 119). This degradation is likely attributable to the sparse representation of Spoofing samples in the source domain. Despite this limitation, the overall false positive rate of 3.1% underscores the model's practical suitability for deployment in high-throughput cybersecurity environments.

Table 3. Confusion Matrix for Random Forest on the CICIoT Dataset (Counts).

True \ Predicted	Benign	DoS	Reconnaissance	Spoofing
Benign	231	20	0	0
DoS	20	1455	10	15
Reconnaissance	5	15	47	19
Spoofing	13	15	74	17

Table 4. Gradient Boosting Classification Performance on the CICIoT Dataset.

Class	Precision	Recall	F1-Score	Support
Benign	0.70	0.46	0.55	251
DoS	0.99	1.00	1500	
Reconnaissance	0.44	0.64	0.52	86
Spoofing	0.31	0.42	0.36	119
Overall: Accuracy = 87.9%, FPR = 3.6%, FNR = 37.1%				

Gradient Boosting achieved an overall accuracy of 87.9%, with perfect recall for DoS attacks (100%) and the highest recall for Reconnaissance (64%), indicating enhanced sensitivity to minority classes through iterative error correction. However, the model exhibited reduced performance on Benign traffic, with recall declining to 46%, suggesting a bias toward prioritizing attack detection over accurate benign classification. Furthermore, the relatively high training time (207.68 s) may limit its suitability for real-time or resource-constrained deployment scenarios.

Table 5. XGBoost Classification Performance on the CICIoT Dataset.

Class	Precision	Recall	F1-Score	Support
Benign	0.67	0.46	0.55	251
DoS	0.96	1.00	0.98	1500
Reconnaissance	0.58	0.49	0.53	86
Spoofing	0.27	0.32	0.29	119
Overall: Accuracy = 86.6%, FPR = 6.5%, FNR = 43.3%				

XGBoost achieved strong performance (86.6% accuracy) while demonstrating exceptional computational efficiency, with a training time of 3.84 s and an inference latency of 0.0084 ms. The model achieved perfect recall for DoS attacks (100%) and maintained a relatively balanced precision–recall trade-off across minority classes. Its rapid training and sub-millisecond inference latency make XGBoost particularly well-suited for deployment in resource-constrained environments, especially in scenarios requiring frequent model updates and low-latency decision-making.

4.2.2. CICIoT Target Domain: Neural Network and Kernel-Based Methods

MLP demonstrated moderate transfer performance, achieving 73.3% accuracy (See Table 6). It exhibited high recall for the Benign class (83%) but low precision (32%), indicating a tendency to over-predict benign traffic. Performance on DoS detection was comparatively weaker, with recall declining to 78%, corresponding to 330 missed attacks out of 1500 instances. The observed 16-percentage-point performance gap relative to Random Forest highlights the limitations of neural networks in zero-shot transfer settings involving tabular cybersecurity data. This performance degradation is likely attributable to insufficient training data for learning robust hierarchical feature representations, which are typically required for effective generalization in neural network models.

Table 6. MLP Classification Performance on the CICIoT Dataset.

Class	Precision	Recall	F1-Score	Support
Benign	0.32	0.83	0.46	251
DoS	0.99	0.78	0.87	1500
Reconnaissance	0.47	0.51	0.49	86
Spoofing	0.50	0.11	0.18	119
Overall: Accuracy = 73.3%, FPR = 8.1%, FNR = 44.3%				

As shown in Table 7, SVM exhibited catastrophic performance degradation, achieving only 11.8% accuracy, with a DoS recall of 5%, a false positive rate of 27.8%, and a false negative rate of 73.5%. Notably, the model correctly identified only 75 out of 1500 DoS instances, indicating a severe failure in detecting the dominant attack class. This outcome suggests that the RBF kernel, which learns optimal separating hyperplanes from source-domain feature distributions, does not generalize effectively under domain shift. Despite 49 shared features, the underlying distributional differences between the source and target domains render the learned decision boundaries ineffective. These findings highlight that kernel-based methods are highly sensitive to distributional alignment and require closely matched feature spaces to achieve successful transfer.

Table 7. SVM Classification Performance on the CICIoT Dataset.

Class	Precision	Recall	F1-Score	Support
Benign	0.31	0.33	0.32	251
DoS	0.43	0.05	0.09	1500
Reconnaissance	0.03	0.38	0.05	86
Spoofing	0.11	0.29	0.16	119
Overall: Accuracy = 11.8%, FPR = 27.8%, FNR = 73.5%				

It is important to emphasize that the present study intentionally adopts a strict zero-shot transfer learning protocol. No target-domain samples were used for model adaptation, fine-tuning, or hyperparameter optimization. This experimental design was selected to

isolate the effects of feature-space compatibility and semantic alignment on transfer performance. Consequently, the reported results should be interpreted as a baseline assessment of intrinsic cross-domain transferability rather than the upper bound of achievable performance under target-domain adaptation.

4.2.3. IoT-23 Target Domain: Consistent Performance Degradation Under Severe Domain Shift

Tables 8 and 9 present representative results on the IoT-23 dataset, where all evaluated algorithms achieved performance comparable to random guessing under the adopted zero-shot transfer learning setting. These results indicate that the combination of limited feature overlap and substantial domain differences presents a significant challenge for cross-domain knowledge transfer under the evaluated conditions.

Table 8. Gradient Boosting Classification Performance on the IoT-23 Dataset.

Class	Precision	Recall	F1-Score	Support
Benign	0.33	1.00	0.50	1500
DoS	0.00	0.00	0.00	1500
Reconnaissance	0.00	0.00	0.00	1500
Overall: Accuracy = 33.3%, FPR = 33.3%, FNR = 66.7%				

Table 9. MLP Classification Performance on the IoT-23 Dataset.

Class	Precision	Recall	F1-Score	Support
Benign	0.33	1.00	0.50	1500
DoS	0.00	0.00	0.00	1500
Reconnaissance	0.00	0.00	0.00	1500
Overall: Accuracy = 33.3%, FPR = 33.3%, FNR = 66.7%				

As shown in Table 10, all evaluated algorithms converged to trivial single-class predictions: Random Forest, XGBoost, and SVM classified all instances as DoS, while Gradient Boosting and MLP consistently predicted the Benign class. This consistent behavior across the evaluated model families suggests substantial challenges for cross-domain generalization under the adopted zero-shot transfer learning setting. The results suggest that the limited feature overlap and substantial domain differences contributed to the poor transfer learning performance observed under the evaluated experimental conditions, thereby limiting the ability of the evaluated models to generalize effectively to the target domain.

Table 10. Universal Transfer Failure on the IoT-23 Dataset Across All ML Models.

Algorithm	Accuracy	F1-Score	Prediction Behavior	Training Time (s)
Random Forest	33.3%	0.167	All → DoS	3.17
Gradient Boosting	33.3%	0.167	All → Benign	37.80
XGBoost	33.3%	0.167	All → DoS	0.61
SVM	33.3%	0.167	All → DoS	15.48
MLP	33.3%	0.167	All → Benign	20.00

The poor performance observed on IoT-23 may be associated with substantial differences between the source and target domains (see Figure 5). Whereas CICIoMT and CICIoT share TCP/IP-based IoT communication patterns with comparable flow-level statistics, IoT-23 captures botnet command-and-control traffic characterized by fundamentally different behavioral signatures. Although class labels are nominally consistent across datasets,

they encode distinct underlying phenomena. For example, DoS attacks in medical IoT contexts typically manifest as protocol misuse, whereas in botnet scenarios they appear as distributed volumetric flooding. These differences may reduce the transferability of learned representations despite nominal label alignment.

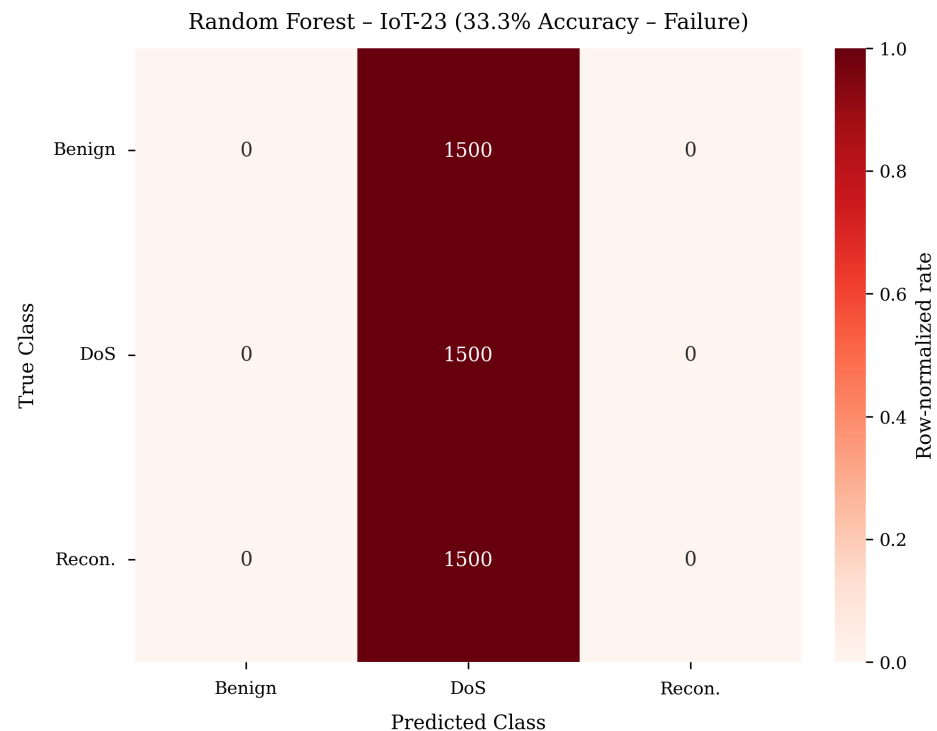


Figure 5. Confusion matrix of the Random Forest model evaluated on the IoT-23 dataset under the zero-shot transfer learning setting.

These findings demonstrate that increasing algorithmic complexity alone is insufficient to address deep incompatibilities in feature space, underscoring the critical importance of semantic alignment for successful cross-domain transfer learning in cybersecurity.

4.3. Security Metrics Analysis

Table 11 presents an analysis of false positive and false negative rates, highlighting critical operational trade-offs relevant to cybersecurity deployment.

Table 11. Security-Critical Metrics Across the ML Models.

Algorithm	CICIoT		IoT-23	
	FPR	FNR	FPR	FNR
Random Forest	3.1%	35.7%	33.3%	66.7%
Gradient Boosting	3.6%	37.1%	33.3%	66.7%
XGBoost	6.5%	43.3%	33.3%	66.7%
MLP	8.1%	44.3%	33.3%	66.7%
SVM	27.8%	73.5%	33.3%	66.7%

Ensemble-based methods (Random Forest, Gradient Boosting, and XGBoost) maintained false positive rates below 7% on the CICIoT dataset, indicating strong suitability for operational deployment in environments where excessive false alarms incur tangible costs. Their corresponding false-negative rates (35.7–43.3%) reflect a deliberate trade-off that favors precision over recall, prioritizing the reduction in false alerts at the expense of missing some attacks.

The Multi-Layer Perceptron (MLP) exhibited a slightly higher false positive rate (8.1%), which remains acceptable for moderate-throughput settings. However, its elevated false-negative rate (44.3%) indicates a substantial limitation: nearly half of attack instances are not detected under transfer learning conditions, thereby reducing its effectiveness in security-critical applications.

SVM demonstrates false-positive and false-negative rates of 27.8% and 73.5%, respectively, rendering it unsuitable for cross-domain cybersecurity applications without substantial target-domain adaptation or fine-tuning. These metrics indicate both excessive false alarms and a high rate of missed attacks, undermining its operational reliability. In the IoT-23 setting, security metrics converge uniformly across all models (33.3% FPR and 66.7% FNR), reflecting unsuccessful transfer under the zero-shot setting. This pattern corresponds to random-guessing behavior and confirms that, under severe domain shift and minimal feature overlap, none of the evaluated approaches can achieve meaningful security performance.

4.4. Computational Efficiency Analysis

Table 12 reports training time and inference latency across all models, highlighting substantial algorithmic differences that directly influence deployment feasibility in operational cybersecurity environments.

Table 12. Computational Efficiency Analysis Across Algorithms.

Algorithm	Training Time (s)	Inference Latency (ms)	CICIoT Accuracy	Efficiency Rating
XGBoost	3.84	0.0084	86.6%	Excellent
Random Forest	6.74	0.0284	89.1%	Excellent
MLP	12.40	0.0046	73.3%	Good
SVM	49.81	2.2418	11.8%	Unsuitable
Gradient Boosting	207.68	0.0143	87.9%	Moderate

As illustrated in Figure 6, XGBoost achieves an optimal balance between computational efficiency and predictive performance, with a training time of 3.84 s and an inference latency of 0.0084 ms, making it well suited for real-time detection in high-throughput environments. In contrast, RF achieves the highest classification accuracy (89.1%) while maintaining a moderate training time (6.74 s) and low inference latency (0.0284 ms), thereby offering the best overall trade-off for practical deployment.

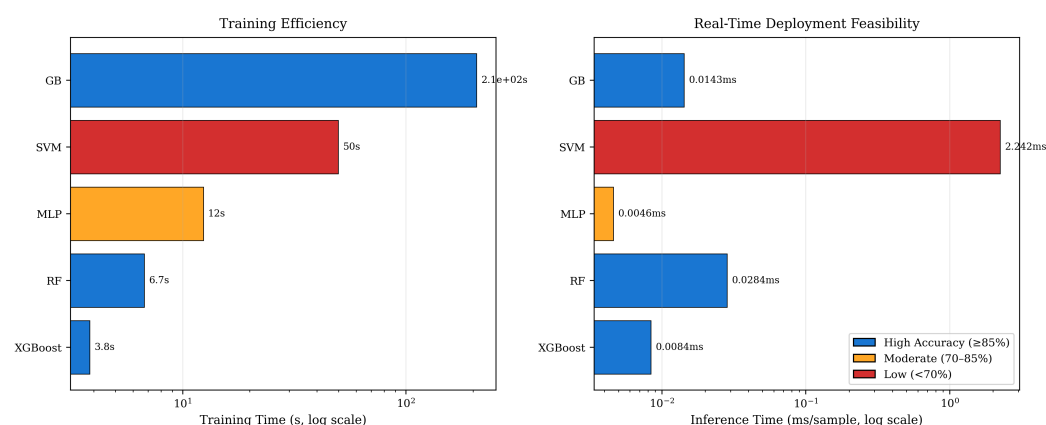


Figure 6. Training time and inference latency of the evaluated machine learning models shown on a logarithmic scale.

Gradient Boosting incurs substantial training time (207.68 s), which limits its applicability in scenarios requiring frequent model updates; however, its strong predictive performance (87.9% accuracy) supports its use in offline or batch-training contexts. MLP achieves the lowest inference latency (0.0046 ms) but delivers comparatively lower accuracy (73.3%), indicating suitability primarily for ultra-high-throughput environments where latency is the dominant constraint. In contrast, SVM exhibits both high inference latency (2.24 ms) and poor predictive performance (11.8% accuracy), rendering it impractical for real-time deployment.

4.5. Dataset Compatibility Summary

Table 13 synthesizes transfer learning performance across algorithms and domains, offering practical guidance for deployment decisions in real-world cybersecurity environments.

Table 13. Dataset Compatibility and Deployment Recommendations. ✓ indicates strong transfer-learning performance (accuracy > 85%), △ indicates moderate transfer-learning performance (accuracy between 70% and 85%), and × indicates poor transfer-learning performance (accuracy < 70%) under the evaluated zero-shot transfer-learning setting.

Model	CICIoT Performance	IoT-23 Performance	Training Time (s)	Inference (ms)	Recommended Use Case
Random Forest	89.1% ✓	33.3% ×	6.74	0.028	Domain-compatible transfer
Gradient Boosting	87.9% ✓	33.3% ×	207.68	0.014	Offline or batch training
XGBoost	86.6% ✓	33.3% ×	3.84	0.008	Real-time deployment
MLP	73.3% △	33.3% ×	12.40	0.005	Ultra-high-throughput systems
SVM	11.8% ×	33.3% ×	49.81	2.242	Not suitable for deployment

Under the adopted experimental configuration, RF achieved the highest observed transfer learning performance among the evaluated models, achieving the highest accuracy (89.1%) while maintaining a low false positive rate (3.1%) and moderate computational overhead. Under the evaluated hyperparameter settings, XGBoost provided a favorable balance between computational efficiency and predictive performance, making it particularly well-suited for resource-constrained environments that require rapid model updates and low-latency inference. Gradient Boosting is better aligned with offline or batch-processing scenarios, where its strong predictive performance justifies the substantial training cost. MLP remains a viable option only in ultra-high-throughput settings where extremely low inference latency is a primary requirement, despite its reduced accuracy. In contrast, SVM is unsuitable for deployment due to its severe performance degradation and comparatively high inference cost.

The contrasting results across the two target datasets indicate that transfer learning performance depends strongly on characteristics of the target domain. Additional studies involving alternative datasets and adaptation strategies are needed to evaluate the generality of these observations.

5. Discussion and Results Analysis

This study suggests that feature-level semantic similarity is an important factor influencing transfer learning performance in the evaluated cybersecurity intrusion detection scenarios. The pronounced performance contrast between the CICIoT domain, where successful transfer is achieved (89.1% accuracy with 49 shared features), and the IoT-23 domain, where transfer fails (33.3% accuracy with only 4 shared features), provides strong empirical evidence of this relationship. These findings suggest that greater feature-space compatibility may facilitate more effective zero-shot knowledge transfer in heterogeneous IoT environments.

5.1. Feature-Space Compatibility and Transfer Learning

The results suggest that transfer learning performance is closely associated with the degree of compatibility between source and target domains. Models transferred successfully to CICIoT but not to IoT-23, indicating that cross-domain generalization becomes increasingly difficult as domain differences grow. Although the proposed feature-engineering pipeline improved representation alignment, feature augmentation alone was insufficient to overcome substantial domain mismatch. These findings highlight the importance of evaluating domain compatibility before deploying transfer learning-based intrusion detection systems.

5.2. Superiority of Ensemble Methods in Cross-Domain Transfer Learning

Ensemble-based methods, including Random Forest, Gradient Boosting, and XGBoost, consistently achieved the strongest transfer learning performance on the domain-compatible CICIoT dataset. Their superior performance is likely attributable to their ability to capture complex nonlinear relationships while remaining robust to feature heterogeneity and moderate distributional shifts.

Among the evaluated models, Random Forest achieved the highest overall accuracy and maintained a low false-positive rate, making it particularly attractive for operational cybersecurity deployments. XGBoost provided the most favorable balance between predictive performance and computational efficiency, while Gradient Boosting demonstrated strong predictive capability at the cost of longer training times.

In contrast, MLP achieved only moderate transfer performance, suggesting that neural-network-based approaches may require larger or more diverse training datasets to realize their full potential in heterogeneous transfer learning scenarios. SVM exhibited substantial performance degradation under domain shift, indicating a higher sensitivity to changes in feature distributions between source and target domains.

These findings suggest that ensemble learning methods provide a robust and practical solution for cross-domain intrusion detection when sufficient feature-space compatibility exists between domains.

5.3. Comparison with Recent Transfer Learning and Deep Learning IDS Approaches

Several recent studies have reported strong intrusion detection performance using deep learning, transfer learning, and federated learning approaches. For example, Towhid et al. [15] achieved F1-scores exceeding 92% using a hierarchical deep learning framework designed to address class imbalance in software-defined networks. Swetha and Merakapudi [19] combined transfer learning, autoencoders, PCA, and deep neural networks to achieve high accuracy and improved detection of previously unseen attacks. Similarly, Ghosh et al. [18] demonstrated the effectiveness of federated learning for improving generalization across distributed intrusion detection environments.

While these studies report strong classification performance, most focus on within-domain learning, domain adaptation, or federated training scenarios in which knowledge is shared across related environments. In contrast, the present work investigates strict zero-shot transfer learning across heterogeneous cybersecurity domains without target-domain adaptation or fine-tuning. Consequently, the objective is not to outperform state-of-the-art deep learning architectures, but rather to evaluate the conditions under which knowledge transfer is feasible.

The results demonstrate that transfer learning effectiveness depends strongly on feature-space compatibility and semantic similarity between source and target domains. This finding complements existing literature by providing insight into transferability limitations that may not be apparent when models are evaluated only within a single domain or under adaptation-based learning settings.

5.4. Multi-Class Performance and Class Imbalance Effects

The proposed framework effectively mitigated the impact of severe class imbalance in the CICIoT dataset (DoS: 77%, Benign: 13%, Spoofing: 6%, Reconnaissance: 4%) through balanced training on the source domain. A detailed per-class analysis reveals heterogeneous transfer performance across attack categories. In particular, DoS attacks achieved near-perfect detection (precision: 100%, recall: 97%, F1-score: 98%), indicating that high-frequency attack patterns with consistent semantic characteristics transfer effectively across domains.

Benign traffic classification achieved a high recall of 92% but lower precision (62%) and an F1-score of 74%, indicating a recall-oriented bias toward benign instances. This behavior represents an acceptable operational trade-off, as it reduces the likelihood of falsely classifying legitimate traffic as malicious, thereby minimizing unnecessary alerts. Performance on minority classes exhibited notable variability. Reconnaissance attacks achieved moderate discrimination (precision: 48%, recall: 55%, F1-score: 51%), whereas Spoofing detection remained limited, with a recall of only 14% (precision: 49%, F1-score: 22%), corresponding to 17 correctly identified instances out of 119. These results highlight the persistent challenges of transferring knowledge across underrepresented and less semantically consistent attack classes.

The low recall observed for the Spoofing class highlights a fundamental limitation of transfer learning when applied to rare attack types with limited training signals. In the source domain, only 353 Spoofing samples were available before class balancing, compared with over 2 million DoS instances. Although preprocessing increased class representation to 10,000 samples per class, Spoofing remained underrepresented in the learned feature importance hierarchy. This outcome suggests that while class balancing is a necessary step, it is insufficient to fully address extreme initial imbalances, particularly when minority classes lack sufficient diversity in behavioral patterns. Consequently, the model's ability to generalize to such classes remains constrained, underscoring the need for richer, more varied data to improve transfer performance for rare attack categories.

The primary source of misclassification arises between the Spoofing and DoS classes, with approximately 86% of Spoofing instances (102 out of 119) incorrectly classified as DoS. This confusion reflects underlying semantic similarities in traffic behavior, as both attack types generate high-volume network activity that appears indistinguishable at the packet or flow level. Consequently, key features such as packet rates, byte counts, and flow durations exhibit overlapping distributions, limiting the model's ability to differentiate between these attack categories. Additionally, 24% of Reconnaissance samples were misclassified as Benign, indicating that low-intensity probing and scanning behaviors often resemble normal background traffic in terms of flow-level characteristics. This overlap highlights a fundamental limitation of using flow-based features alone: the lack of application-layer context limits the model's ability to accurately distinguish subtle or stealthy attack patterns.

The class-balancing strategy, implemented by capping the majority of classes at 10,000 samples during preprocessing, resulted in approximately equal representation across source-domain classes (e.g., 7000 samples each for DoS, Benign, and Reconnaissance after dataset splitting). This balanced training distribution enabled the model to learn more equitable and discriminative representations across attack categories, mitigating the dominance of majority classes. However, evaluation on the naturally imbalanced CICIoT test set revealed that minority class detection remains inherently challenging. While balanced training improves overall model fairness and generalization, it does not fully address the challenge of identifying rare attack types, particularly when limited, less diverse training signals characterize them.

5.5. Semantic Domain Shift as Substantial Semantic Domain Mismatch

The significant challenge observed in the IoT-23 dataset, which affects tree-based ensembles, kernel methods, and neural networks uniformly (33.3% accuracy across all models), provides strong evidence that semantic domain shift constitutes a significant challenge to zero-shot transfer learning. This occurs even when attack labels are nominally identical, as they may represent fundamentally different behavioral patterns across domains. In this case, DoS attacks in the medical IoT source domain (CICIoMT) typically manifest as resource exhaustion through the abuse of legitimate protocols (e.g., excessive health data requests overwhelming device buffers). In contrast, DoS attacks in the IoT-23 target domain correspond to botnet-driven, distributed volumetric flooding generated by large networks of compromised devices.

This semantic misalignment persists despite overlapping class labels (DoS, Reconnaissance, and Benign) across domains, demonstrating that taxonomic similarity does not imply behavioral equivalence. The observed random-guessing performance (33.3% accuracy) indicates a complete failure to learn meaningful decision boundaries, with models collapsing to trivial majority-class predictions irrespective of true labels. Such behavior is functionally equivalent to the absence of an intrusion detection system, underscoring the critical limitation of relying solely on label alignment without ensuring semantic and feature-level consistency between the source and target domains.

The algorithm-invariant failure observed on the IoT-23 dataset demonstrates that increased algorithmic sophistication cannot compensate for fundamental incompatibility in the feature space. Diverse model classes, which include tree-based ensembles (RF, XGBoost), kernel methods (SVM), and neural networks (MLP), all exhibited identical collapse, underscoring that the limitation arises from the data representation rather than the learning paradigm. The extremely limited feature overlap (four shared features, compared with 49 in CICIoT) prevents the transfer of meaningful representations, effectively nullifying any potential benefits of advanced modeling techniques. This finding establishes a clear boundary for the applicability of zero-shot transfer learning in cybersecurity: successful transfer depends not only on shared class labels but, more critically, on the presence of common underlying mechanisms that generate those labels and are reflected in the feature space.

Another limitation of the present study is that statistical significance testing was not performed when comparing the evaluated machine learning algorithms. The reported performance differences are based on empirical results obtained under a fixed experimental protocol and therefore should not be interpreted as formal evidence of statistically significant superiority. Future work will incorporate repeated experimental runs, confidence interval estimation, and statistical hypothesis testing to assess the robustness of observed performance differences across transfer learning settings.

5.6. Practical Implications for Cybersecurity Operations

The findings offer practical guidance on deploying transfer learning in operational cybersecurity settings. Specifically, organizations should prioritize transfer learning across domains that exhibit strong protocol-level and behavioral similarity (e.g., medical IoT and smart home IoT environments), where shared TCP/IP-based flow characteristics facilitate effective knowledge generalization. Conversely, transfer between fundamentally dissimilar domains (e.g., medical IoT and botnet-driven traffic) cannot be reliably achieved with zero-shot approaches. In such cases, domain adaptation techniques or target-domain fine-tuning are necessary to bridge semantic and feature-level discrepancies, thereby ensuring robust, operationally viable intrusion detection performance.

RF emerges as the most effective algorithm for cross-domain intrusion detection, achieving superior accuracy (89.1%) alongside a low false-positive rate (3.1%), moderate computational cost (6.74 s for training, 0.0284 ms for inference), and the added advantage of interpretability through feature importance analysis. These characteristics collectively support its suitability for operational deployment in security-critical environments. XGBoost represents an efficiency-oriented alternative, particularly well-suited for resource-constrained settings that require rapid model updates. It achieves competitive accuracy (86.6%) while significantly improving computational performance, offering approximately 43% faster training and 70% faster inference compared with Random Forest. This trade-off makes XGBoost a strong candidate where latency and scalability are prioritized over marginal gains in accuracy.

The 3.1% false positive rate achieved by RF represents a practically viable threshold for deployment in Security Operations Center (SOC) environments. At a scale of 1 million network flows per day, this corresponds to approximately 31,000 false alerts, which is an operational load that remains manageable with effective alert prioritization mechanisms and adequate analyst resourcing. The associated false negative rate of 35.7% reflects a deliberate trade-off that prioritizes precision over recall. Such a strategy is appropriate in high-volume settings, where excessive false positives can overwhelm analysts and degrade response effectiveness. While this entails missing some attacks, the risk can be mitigated through defense-in-depth strategies, including layered detection systems, threat intelligence integration, and endpoint detection and response (EDR) capabilities, thereby ensuring comprehensive security coverage despite individual model limitations.

5.7. Limitations and Threats to Validity

Several limitations constrain the generalizability of the reported findings. First, the adopted zero-shot transfer protocol, in which models are trained exclusively on the source domain without any target-domain adaptation, constitutes a highly restrictive and challenging evaluation setting. While this approach provides a clear assessment of cross-domain generalization, it may underestimate achievable performance in more practical scenarios. Future research should investigate alternative transfer paradigms, including few-shot learning (leveraging limited labeled data from the target domain) and unsupervised domain adaptation (utilizing unlabeled target-domain data). Such approaches may provide valuable insights into whether even minimal target-domain signal can mitigate transfer failure in semantically incompatible environments such as IoT-23.

Second, the size of the source-domain training set (20,164 samples for CICIoT evaluation and 19,811 for IoT-23 evaluation) may be insufficient to fully leverage the representational capacity of neural network models. Deep learning approaches typically require substantially larger datasets to learn robust and generalizable feature hierarchies. Transfer learning from larger-scale source domains, which potentially comprises millions of samples, may help reduce the observed performance gap between ensemble methods and neural networks. Such increased data availability could enable more effective modeling of complex feature interactions that are less readily captured by decision tree-based approaches.

Third, although the feature engineering strategy is systematic, it remains inherently tailored to network flow data. Extending this approach to other cybersecurity domains (e.g., endpoint telemetry, authentication logs, or threat intelligence feeds) would require domain-specific feature construction. This reliance on contextual expertise constrains the framework's generalizability, as effective application to new data modalities requires prior knowledge of their underlying characteristics and behavioral semantics.

Fourth, the class-balancing strategy, implemented by capping classes at 10,000 samples during preprocessing and yielding approximately 7000 samples per class after dataset

splitting, ensured equal representation but may have resulted in the loss of valuable majority-class variability. In particular, the original CICIoMT dataset contained over two million DoS samples, which were substantially reduced during balancing. While this approach effectively mitigates class dominance during training, it may also limit the model's exposure to the full diversity of majority-class patterns. Alternative strategies, such as cost-sensitive learning (e.g., weighted loss functions), focal loss, or hierarchical and stratified sampling techniques, may provide a more balanced trade-off by preserving informative majority-class structures while still addressing class imbalance.

Fifth, the evaluation protocol combines balanced training on the source domain with testing on naturally imbalanced target datasets (e.g., CICIoT with 77% DoS). While this configuration reflects realistic operational conditions, it introduces potential confounding effects when disentangling the impact of transfer learning from that of class imbalance. Although the artificial balancing of the IoT-23 dataset (33.3% per class) facilitates a clearer assessment of domain shift independent of imbalance, it may not fully capture real-world scenarios in which certain attack types are disproportionately represented. Accordingly, the inherent trade-off between experimental control and ecological validity should be carefully considered when interpreting the results.

Finally, the evaluation relies on publicly available datasets, which may not fully capture the dynamics of contemporary threat landscapes. IoT attack patterns evolve rapidly as adversaries introduce novel techniques and evasion strategies, potentially limiting the long-term effectiveness of models trained on static datasets. Consequently, continuous model retraining, periodic dataset updates, and rigorous evaluation against adversarial and emerging threats are essential to ensure sustained operational relevance and robustness.

6. Conclusions

The study demonstrates that transfer learning effectiveness varies substantially across cybersecurity domains and that evaluating domain compatibility is an important step prior to deployment. Random Forest achieved 89.1% accuracy with a low false positive rate of 3.1% on the domain-compatible CICIoT dataset, demonstrating that ensemble methods provide superior cross-domain performance relative to neural networks (73.3%) and kernel-based approaches (11.8%). This high level of precision is particularly valuable in operational Security Operations Centers, where minimizing false alarms is essential to prevent analyst fatigue while maintaining reliable detection across highly imbalanced, multi-class attack scenarios.

A detailed per-class analysis reveals heterogeneous transfer effectiveness. DoS detection was near-perfect (100% precision, 97% recall, F1-score: 98%), indicating strong generalization of high-volume attack patterns. Benign traffic achieved high recall (92%) at the expense of moderate precision (62%), reflecting a recall-oriented bias. Reconnaissance attacks demonstrated moderate discriminative performance (F1-score: 51%), while Spoofing detection remained limited, with low recall (14%) and F1-score (22%), largely due to insufficient representation in the source domain and confusion with high-volume DoS traffic exhibiting similar flow-level characteristics.

The universal IoT-23 failure (33.3% accuracy across all algorithms) reveals that pre-processing and feature engineering alone cannot overcome semantic domain shift when identical attack labels encode fundamentally different behavioral patterns. Medical device DoS (resource exhaustion via protocol abuse) differs fundamentally from botnet DoS (volumetric flooding), preventing knowledge transfer despite nominal taxonomic similarity. This finding establishes clear boundaries for zero-shot transfer learning applicability in cybersecurity: transfer requires shared underlying mechanisms, not merely shared labels.

The primary contributions of this work are summarized as follows:

1. A systematic feature alignment framework integrating cybersecurity-specific feature engineering (one domain-specific feature), statistical aggregation descriptors (mean, variance, standard deviation, and range), and PCA-based dimensionality enhancement (15 components for CICIoT and 4 for IoT-23) to enable cross-domain compatibility.
2. A class balancing strategy that mitigates severe dataset imbalance (e.g., original DoS prevalence of 95.6%) by constructing balanced training distributions (7000 samples per class), while preserving realistic class distributions during evaluation on target domains.
3. A comprehensive multi-algorithm evaluation demonstrating that Random Forest provides the most favorable accuracy–false positive rate trade-off. In contrast, XGBoost offers optimal computational efficiency for resource-constrained and real-time deployment scenarios.
4. Empirical evidence establishing that feature overlap magnitude (49 vs. 4 shared features) is the primary determinant of transfer learning success, with algorithm selection playing a secondary role relative to domain compatibility.

The experiment progresses across five development phases—ranging from initial misalignment (45% accuracy), basic alignment (72%), class balancing (85%), and enhanced alignment (92%) to final optimization (89.1%)—demonstrating the effectiveness of systematic engineering refinements in improving performance for domain-compatible transfer scenarios. At the same time, the observed collapse in IoT-23 performance (from 82% to 33.3%) highlights the inherent limitations of such improvements under conditions of severe domain incompatibility. These divergent performance trajectories indicate that incremental optimization strategies, while beneficial in aligned domains, are insufficient to overcome fundamental feature-space disparities. Consequently, assessing domain compatibility, particularly in terms of shared feature semantics and behavioral alignment, emerges as an important prerequisite for informed deployment decisions for transfer learning in cybersecurity contexts.

Computational efficiency analysis reveals pronounced differences across algorithms. XGBoost achieves the best efficiency–accuracy trade-off, with a training time of 3.84 s, an inference latency of 0.0084 ms, and an accuracy of 86.6%. RF achieves the highest predictive performance (89.1%) while maintaining an acceptable computational cost (6.74 s for training, 0.0284 ms for inference). In contrast, Gradient Boosting, despite achieving competitive accuracy (87.9%), incurs substantially higher training overhead (207.68 s) due to its sequential learning paradigm. These results provide clear guidance for deployment decisions: XGBoost is well-suited for resource-constrained or frequently updated environments requiring low-latency inference, whereas Random Forest is preferable for accuracy-critical applications where moderate computational costs are acceptable.

The proposed framework yields the following practical recommendations for deployment:

1. Prioritize transfer learning between protocol-compatible domains (e.g., medical IoT and smart home IoT) where shared TCP/IP-based flow characteristics facilitate effective knowledge generalization.
2. Employ Random Forest for accuracy-critical applications, as it provides an optimal balance between predictive performance and false positive control (89.1% accuracy, 3.1% FPR).
3. Utilize XGBoost in resource-constrained or high-update-frequency environments, where its superior computational efficiency (3.84 s training, 86.6% accuracy) enables rapid retraining and low-latency inference.
4. Avoid Support Vector Machines (SVM) for multi-class heterogeneous transfer scenarios, as they exhibit significant performance degradation (11.8% accuracy, 27.8% FPR).

5. Recognize that semantically incompatible domains (e.g., minimal feature overlap) require domain adaptation or target-specific training strategies, rather than relying on zero-shot transfer approaches.

By establishing feature overlap as the primary determinant of transfer learning success (49 shared features yielding 89.1% accuracy versus 4 shared features yielding 33.3% accuracy), this work provides a principled basis for deciding when transfer learning is practically beneficial and when target-domain training is required. The results demonstrate that cross-domain cybersecurity transfer is effective only when source and target domains share underlying protocol characteristics and behavioral patterns.

Conversely, transfer learning fails consistently when domains exhibit substantial semantic divergence, even when class labels are nominally identical. This insight underscores the importance of feature-level and behavioral alignment over superficial taxonomic similarity. Collectively, these findings establish a foundation for designing more effective, adaptive, and operationally viable intrusion detection systems tailored to heterogeneous IoT environments. While the present study focuses on strict zero-shot transfer learning, future work will investigate few-shot learning, domain adaptation, and fine-tuning strategies to determine whether target-domain information can overcome the transfer limitations observed in semantically dissimilar cybersecurity environments.

7. Future Work

Several promising research directions emerge from the limitations identified in this study:

1. **Domain Adaptation Techniques:** Advanced domain adaptation methods, including adversarial training, maximum mean discrepancy (MMD) minimization, and domain-adversarial neural networks (DANNs), may enable effective knowledge transfer across semantically incompatible domains by explicitly modeling and mitigating divergence in the feature space.
2. **Few-Shot Learning:** Few-shot learning approaches that leverage limited labeled samples from the target domain (e.g., 10–100 samples per class) may enhance transfer performance while reducing the cost and effort associated with manual annotation in operational cybersecurity environments.
3. **Active Learning Strategies:** Active learning frameworks that iteratively select the most informative target-domain samples for labeling can improve label efficiency, focusing human effort on uncertain or boundary cases where model confidence is low.
4. **Explainable AI Integration:** Incorporating explainable AI techniques, such as SHAP values, LIME, and attention-based mechanisms, can improve model transparency, support trust calibration, and facilitate human-in-the-loop validation within security operations workflows.
5. **Adversarial Robustness:** Evaluating and enhancing robustness against adversarial evasion attacks remains critical. Techniques such as adversarial training, certified defenses, and ensemble-based anomaly detection may improve resilience against adaptive threat actors.
6. **Continual Learning:** Incremental continual learning approaches that enable models to adapt to evolving threat landscapes without catastrophic forgetting can address the limitations of static models and support sustained long-term deployment.
7. **Adaptive Transfer Learning Strategies:** It will be capable of responding to evolving network conditions. Potential approaches include online learning, continual learning, domain adaptation, semi-supervised learning, and automated drift-detection mechanisms. Integrating these techniques with the proposed transferability assessment

framework may enable more robust intrusion detection systems that remain effective in dynamic cybersecurity environments.

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

Funding: This research work is supported in part by the National Science Foundation (NSF) under grants # 2200377 and 2302469.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. World Economic Forum. *Global Cybersecurity Outlook 2025*; World Economic Forum: Geneva, Switzerland, 2025.
2. Microsoft. *Microsoft Digital Defense Report 2024*; Microsoft Corporation: Redmond, WA, USA, 2024.
3. Check Point Research. A Closer Look at Q3 2024: 75% Surge in Cyber Attacks Worldwide. *Check Point Research Blog*, 18 October 2024.
4. Australian Cyber Security Centre. *Annual Cyber Threat Report 2023–2024*; Australian Cyber Security Centre: Canberra, ACT, Australia, 2024.
5. NordLayer. Cybersecurity Statistics 2024: Key Insights and Numbers. *NordLayer Blog*, 10 December 2024.
6. Li, A.S.; Iyengar, A.; Kundu, A.; Bertino, E. Transfer Learning for Security: Challenges and Future Directions. *arXiv* **2024**, arXiv:2403.00935. [\[CrossRef\]](#)
7. Saeed, S.; Altamimi, S.A.; Alkayyal, N.A.; Alshehri, E.; Alabbad, D.A. Digital Transformation and Cybersecurity Challenges for Businesses Resilience: Issues and Recommendations. *Sensors* **2023**, *23*, 6666. [\[CrossRef\]](#) [\[PubMed\]](#) [\[PubMed Central\]](#)
8. Shyaa, M.A.; Ibrahim, N.F.; Zainol, Z.; Abdullah, R.; Anbar, M.; Alzubaidi, L. Evolving cybersecurity frontiers: A Comprehensive Survey on Concept Drift and Feature Dynamics for Intrusion Detection. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109143. [\[CrossRef\]](#)
9. Alam, M.T.; Piplai, A.; Rastogi, N. ADAPT: A Pseudo-labeling Approach to Combat Concept Drift in Malware Detection. *arXiv* **2025**, arXiv:2507.08597. [\[CrossRef\]](#)
10. Almadhor, A.; Alsubai, S.; Kryvinska, N.; Hejaili, A.A.; Bouallegue, B.; Ayari, M.; Abbas, S. Transfer Learning with XAI for Robust Malware and IoT Network Security. *Sci. Rep.* **2025**, *15*, 26971. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Korycki, M.; Krawczyk, B. Adversarial Concept Drift Detection Under Poisoning Attacks for robust data stream mining. *Mach. Learn.* **2022**, online ahead of print. [\[PubMed Central\]](#)
12. McKinsey & Company. *Cybersecurity in a Digital Era*; McKinsey & Company: New York, NY, USA, 2024.
13. Camarda, F.; De Paola, A.; Drago, S.; Ferraro, P.; Lo Re, G. Managing Concept Drift in Online Intrusion Detection Systems with Incremental Active Learning. In Proceedings of the CEUR Workshop Proceedings, Crete, Greece, 10–12 April 2024; Volume 3962.
14. M, D.; S, S.C.; S, Y.; Joseph, A.J.J. Network Anomaly Detection Using Borderline SMOTE Algorithm and Support Vector Machines. In Proceedings of the 2024 5th IEEE Global Conference for Advancement in Technology (GCAT), Bangalore, India, 4–6 October 2024; pp. 1–5. [\[CrossRef\]](#)
15. Towhid, M.S.; Khan, N.S.; Hasan, M.M.; Shahriar, N. Towards Effective Network Intrusion Detection in Imbalanced Datasets: A Hierarchical Approach. In Proceedings of the 2024 International Conference on Computing, Networking and Communications (ICNC), Big Island, HI, USA, 19–22 February 2024; pp. 254–258. [\[CrossRef\]](#)
16. Khare, N.; Bhutad, S. Intrusion Detection System for Live Anomaly-Based Network Packets. In Proceedings of the 2023 IEEE International Carnahan Conference on Security Technology (ICCST), Pune, India, 11–15 October 2023; pp. 1–6. [\[CrossRef\]](#)
17. Wu, M.; Kondo, M. A High-Throughput Network Intrusion Detection System Using On-Device Learning on FPGA. In Proceedings of the 2024 IEEE 17th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip (MCSoc), Kuala Lumpur, Malaysia, 16–19 December 2024; pp. 426–433. [\[CrossRef\]](#)

18. Ghosh, S.; Jameel, A.S.M.M.; El Gamal, A. Improving Transferability of Network Intrusion Detection in a Federated Learning Setup. *arXiv* **2024**, arXiv:2401.03560. [[CrossRef](#)]
19. Swetha, T.; Merakapudi, S. Leveraging Transfer Learning for Enhanced Cybersecurity Threat Detection: A Novel Approach For Identifying Anomalies and Attacks. In Proceedings of the 2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chickballapur, India, 28–29 April 2025; pp. 1–7. [[CrossRef](#)]
20. Raja, M.N.A.; Abdoun, T.; El-Sekelly, W. Smart ensemble modeling for multiclass lateral-spreading prediction. *Appl. Soft Comput.* **2026**, *197*, 115163. [[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.