POLITECNICO DI TORINO
Repository ISTITUZIONALE

MERGE: Multi-Scenario Embedding for Robust Generalization in Network Anomaly Detection

*Original*
MERGE: Multi-Scenario Embedding for Robust Generalization in Network Anomaly Detection / Colella, C., Serra, C., Sacco, A., Marchetto, G.. - ELETTRONICO. - (2026), pp. 463-468. (IEEE NetSoft 2026 - 5th International Workshop on Edge Network Softwarization Berlin, Germany 29 June - 03 July 2026) [10.1109/netsoft70012.2026.11603479].

*Availability:*
This version is available at: 11583/3013596 since: 2026-08-21T13:26:20Z

*Publisher:*
IEEE

*Published*
DOI:10.1109/netsoft70012.2026.11603479

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

11 September 2026

# MERGE: Multi-Scenario Embedding for Robust Generalization in Network Anomaly Detection

Christian Colella, Cristiano Serra, Alessio Sacco, Guido Marchetto
*Department of Control and Computer Engineering, Politecnico di Torino,*
Turin, Italy.

***Abstract*—Modern computer networks are increasingly complex, heterogeneous, and exposed to a growing spectrum of cyber threats, making robust network anomaly detection (NAD) a critical challenge. While machine learning (ML) has significantly improved detection accuracy, most existing NAD systems remain tightly coupled to the datasets on which they are trained, leading to poor generalization across diverse infrastructures and evolving traffic patterns. To fill this gap, we present MERGE, a NAD solution that shifts from dataset-specific optimization toward architectural unification. We address dataset heterogeneity by training a single unified autoencoder and a shared classifier that operate across multiple data sources, establishing a universal latent space and decision boundary for anomaly detection. Unlike standard approaches that rely on environment-specific calibration, MERGE leverages eXplainable AI (XAI) as an architectural optimization tool to identify a universal feature set. This enables distilling multiple specialized pipelines into a single, compact model instance. Experimental results over three benchmarks (enterprise, IoT, and hybrid traffic) demonstrate that MERGE achieves near-parity with dataset-specific baselines, exhibiting only a 1-2% difference in F1-Score. Crucially, this unification reduces model management overhead and parameter footprint by a factor of $M$ (where $M$ is the number of datasets), providing a scalable and interpretable solution for large-scale network environments.**

***Index Terms*—anomaly detection, ML, AI, representation learning**

## I. Introduction

The modern digital landscape is characterized by the deep integration of enterprise Information Technology (IT) and the Internet of Things (IoT), bringing large numbers of heterogeneous, often unmanaged devices onto IP networks and expanding the operational attack surface [1], [2]. In such environments, defenders must detect a broad range of anomalous behaviors, spanning unauthorized access attempts in corporate infrastructures to malicious communications among compromised IoT devices and internal reconnaissance (e.g., scanning) preceding lateral movement. Traditional misuse- and signature-based defenses, while effective for known threats, are brittle against novel or rapidly evolving attacks and therefore provide limited coverage for zero-day behaviors [3], [4]. As a result, Network Anomaly Detection (NAD) has increasingly moved toward data-driven machine learning (ML) approaches that model normal network behavior and flag deviations across contexts, enabling detection beyond predefined signatures [5]–[7].

A significant barrier to the real-world deployment of these models is the operational complexity of model proliferation. Most NAD systems are developed and validated in isolated data environments, leading to a domain shift in which a model optimized for one type of traffic fails when applied to another [8], [9]. This forces network administrators to maintain, update, and secure a diverse ensemble of specialized detectors–one for each network segment–resulting in a fragmented security posture and high administrative costs.

To fill this gap, we propose MERGE, an ML-based model that shifts the paradigm from dataset-specific tuning toward operational unification. The proposed framework employs a single Autoencoder (AE) to learn a compact latent representation of network traffic, mapping heterogeneous traffic patterns into a unified latent space. Unlike existing methods that rely on environment-specific calibration, MERGE utilizes a shared Feed-Forward Neural Network (FFNN) classifier and a single decision threshold across all data sources. This transformation mitigates environment-specific inconsistencies and yields a standardized representation of network data, enabling a "plug-and-play" security deployment. Building on this foundation, we leverage Explainable Artificial Intelligence (XAI), specifically SHAP (SHapley Additive exPlanations) [10], as a mechanism for architectural optimization rather than post-hoc explanation. In MERGE, SHAP is used to identify and rank the most influential features for anomaly detection across environments. By isolating the features that contribute the most predictive power globally, we enable the distillation of multiple specialized pipelines into a single, compact architecture, ensuring that the model focuses on transferable, high-impact features rather than dataset-specific noise.

To evaluate adaptability, we conduct experiments on three fundamentally different benchmarks: CIC-IDS2017 [11] (enterprise traffic), Bot-IoT [12] (IoT traffic), and UNSW-NB15 [13] (hybrid modern traffic). This heterogeneous evaluation demonstrates that a unified representation of network behavior can be learned even under extreme variability [14], [15]. Empirical results show that MERGE achieves cross-dataset generalization performance closely matching dataset-specific baselines, with F1-scores ranging approximately from $0.91$ to $1.00$ across all evaluated benchmarks. While MERGE exhibits a minor "unification tax" of 1–2% in F1-score compared to specialized models, this delta is offset by a reduction in parameter footprint and management overhead by a factor of $M$ (where $M$ is the number of monitored datasets). These findings confirm that a compact, explainability-guided architecture can effectively capture the most influential features and

operate robustly across diverse network environments while significantly reducing model complexity and computational residency.

The remainder of the paper is organized as follows. Section II reviews the state of the art, and Section III overviews the proposed system. Section IV details the implementation, reports experimental results, and describes the iterative process used to design MERGE. Finally, Section V concludes the paper and outlines directions for future research.

## II. Related Work

While many Network Anomaly Detection (NAD) models achieve high accuracy under standard evaluation settings, most overlook the challenge of cross-dataset generalization. Classical machine learning (ML) classifiers typically perform well when trained and tested on the same dataset, but their effectiveness significantly degrades across diverse data sources [16]–[18]. Such performance drops are mainly caused by differences in feature distributions, variations in attack representations, and heterogeneous traffic characteristics across datasets, leading models to overfit dataset-specific patterns and to diverge in real-world environments where traffic characteristics evolve over time. Layeghy et al. [17] provide a detailed analysis of how these factors affect generalization, while Iwanowski et al. [18] show that training data selection and preprocessing choices strongly influence cross-dataset performance. These studies collectively highlight the need for methods that explicitly address dataset heterogeneity.

Several recent works have attempted to mitigate this problem. Learn-IDS [9] enables cross-dataset training by unifying raw packet processing across heterogeneous traffic sources; however, its evaluation relies on balanced subsets, potentially overestimating performance in realistic, highly imbalanced scenarios. DI-NIDS [19] addresses domain shift through adversarial domain adaptation to learn domain-invariant features, but it assumes the availability of NetFlow representations, which may not always be feasible in operational settings. META [14] introduces a unified multimodal dataset to support cross-dataset learning from flow-level, payload-level, temporal, and contextual features from multiple sources, though its validation is limited to basic ML models, leaving its effectiveness with more advanced approaches unexplored. PTN-IDS [20] explores few-shot learning by representing attacks as prototype vectors, showing promising transfer between CIC-IDS2017 and CIC-IDS2018; however, its evaluation is restricted to a small set of similar attack types, limiting conclusions about broader generalization. Finally, de Oliveira et al. [21] propose a distributed ensemble-based NAD that improves scalability and cross-dataset performance through majority voting among shallow classifiers. Despite relative improvements, absolute performance remains limited, indicating that ensemble strategies alone are insufficient to fully address severe distribution shifts across datasets.

In contrast, MERGE focuses on robust feature selection to systematically identify features that are both discriminative and transferable across datasets. By explicitly targeting feature-level robustness, it addresses limitations of both classical classifiers and embedding-based methods, offering a more reliable foundation for cross-dataset network intrusion detection.

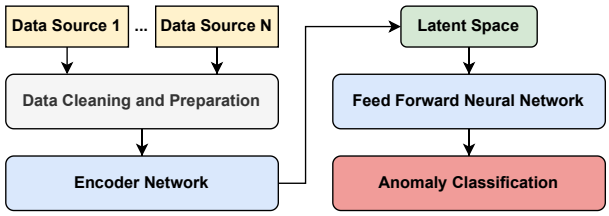


Fig. 1: MERGE's Architecture Overview

## III. System Overview

MERGE aims to address cross-domain NAD by learning a compact, shared representation that generalizes across heterogeneous network environments. Rather than relying on dataset-specific models or handcrafted feature alignment, the system enforces architectural and decision-level unification to promote domain-invariant detection capabilities. An overview of the proposed unified detection pipeline is illustrated in Fig. 1, which summarizes the data flow from heterogeneous network sources through data cleaning and preparation, encoder-based feature extraction, latent space compression, and final anomaly classification.

Each input dataset undergoes a standardized preprocessing pipeline to ensure numerical stability and consistency across environments. Missing values, infinite entries, and duplicated records are removed, and all features are normalized to the $[0, 1]$ range using Min–Max scaling. To prevent information leakage, normalization parameters are computed exclusively on the training partition and reused unchanged during validation and testing.

At the core of the framework is a single, shared Autoencoder (AE) [22] that maps input vectors to an N-dimensional latent space. This bottleneck is tuned to capture salient statistical regularities while discarding dataset-specific noise. Once trained, the encoder is retained as a universal feature compressor.

The resulting latent embeddings are fed into a single feed-forward neural network (FFNN) classifier. Unlike systems that require environment-specific calibration, MERGE utilizes a shared decision threshold across all datasets. This design explicitly enforces a unified decision function, reflecting realistic deployment scenarios where prior knowledge of the target infrastructure is unavailable. This structural parsimony provides a significant advantage in parameter economy; by consolidating $M$ environment-specific models into a single instance, MERGE achieves a theoretical reduction in the parameter footprint by a factor of $M$. For the three environments evaluated, this represents a 66.7% reduction in architectural overhead, enabling deployment in resource-constrained edge gateways.

To further optimize the architecture, the framework incorporates SHAP for feature selection [23]. We identify a compact, static Top-K feature set to prioritize operational efficiency

and consistent input dimensionality. While iterative selection methods exist, our static approach ensures a fixed architectural footprint suitable for hardware-accelerated middleboxes. This explainability-driven refinement allows for the distillation of multiple specialized pipelines into a single, compact model instance that balances detection performance, interpretability, and computational residency in dynamic network environments.

## IV. Evaluations and Results

This section describes the experimental methodology and empirical results of the proposed framework. We first outline the experimental setup and training configuration, then define the evaluation protocol and metrics used in the study. Finally, we present the results, providing a detailed robustness analysis through the various architectural configurations introduced in Section III.

### A. Experimental Setup

All experiments are conducted on the datasets introduced in Section III, using the same preprocessing pipeline, feature normalization strategy, and data partitioning scheme. Train, validation, and test splits are fixed at 55%, 15%, and 30%, respectively. The experimental evaluation considers three main configurations: (i) dataset-specific baselines, where each dataset is modeled independently, (ii) unified cross-domain training, where latent representations extracted from multiple environments are jointly used to train a single classifier, and (iii) XAI-driven refined configurations, where SHAP-based feature selection is applied to reduce the input dimensionality and simplify the architecture.

The neural network architectures and training pipeline were implemented using PyTorch and accelerated via NVIDIA A30 GPUs with Mixed Precision (FP16) to optimize throughput.

*1) Model Architectures:* The AE used for feature extraction follows a symmetrical bottleneck design. The encoder consists of three linear layers with size [input, 128, 64, $N$], where $N$ represents the latent dimension. The primary aim of varying $N$ is to identify the optimal information bottleneck required to isolate domain-invariant traffic characteristics from dataset-specific noise. To ensure stable training, Batch Normalization and LeakyReLU activations ($\alpha = 0.2$) are applied after each hidden layer, while the bottleneck layer uses a Sigmoid activation to constrain the latent space. The decoder mirrors this structure to reconstruct the original input dimensionality. The learning objective for the AE is the minimization of the Mean Squared Error (MSE) loss, which ensures that the latent representation N captures the most salient statistical features. The classification component is an FFNN comprising four layers with a hidden dimension of 512. We employ ReLU activations and Layer Normalization to facilitate robust gradient flow. The final output is a single logit that represents the probability of a sample being malicious. The FFNN is optimized using Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss), which combines a Sigmoid layer and the BCE loss in a single class for increased numerical stability.

*2) Training Configuration:* The model is optimized using the Adam optimizer with an initial learning rate of $1 \times 10^{-3}$ and a batch size of 1024. Training is capped at a maximum of 150 epochs, with early stopping implemented to prevent overfitting. The patience is set to 10 epochs. To further refine convergence, we use the `ReduceLROnPlateau` scheduler, which scales the learning rate by a factor of 0.5 if the validation loss stagnates for 5 consecutive epochs. To evaluate the impact of the bottleneck capacity on feature extraction, we test a discrete range of latent dimensions $N \in \{4, 8, 16, 32, 64\}$. For the XAI-driven refinement, SHAP's `GradientExplainer` is configured with 100 background samples and 100 test samples to generate robust feature importance rankings.

### B. Evaluation Protocol and Metrics

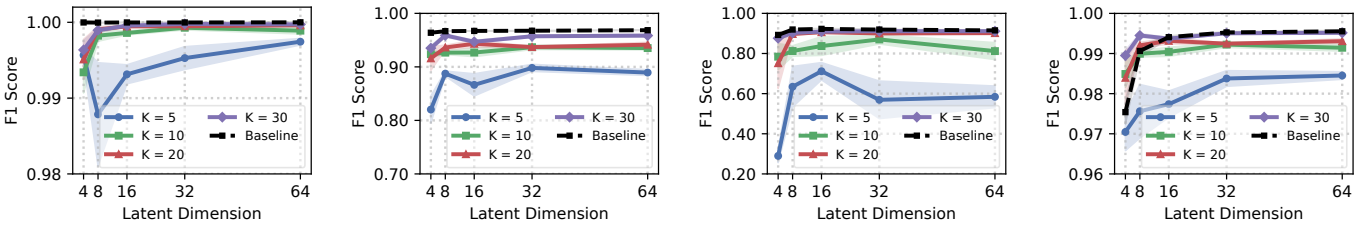
Due to the inherent class imbalance in network security datasets, we adopt the F1-Score as the primary evaluation metric, as it provides a harmonic mean between Precision and Recall. A critical component of our protocol is the Dynamic Threshold Selection. Rather than using a fixed global threshold (e.g., 0.5), we determine the optimal decision threshold for each model configuration by analyzing the Precision-Recall (PR) curve on the validation set. We select the threshold that maximizes the F1-score and subsequently apply this value to the unseen test partition for final evaluation. This approach ensures that no data leakage occurs, as the test set is never used during threshold optimization, and fair comparison, by evaluating each architectural configuration at its peak empirical capability.

### C. Experimental Results

We evaluate the fully unified MERGE model, which utilizes a single AE, a shared classifier, and fixed SHAP-selected features. To ensure statistical significance, all metrics report the mean and $90\%$ confidence intervals across $5$ independent runs. Experiments span three diverse benchmarks with varying feature spaces—CIC-IDS2017 [11] (enterprise), Bot-IoT [12] (IoT), and UNSW-NB15 [13] (hybrid)—which typically preclude training a single model. Finally, we establish an upper-bound baseline using a *unified* training dataset randomly sampled from all three sources.

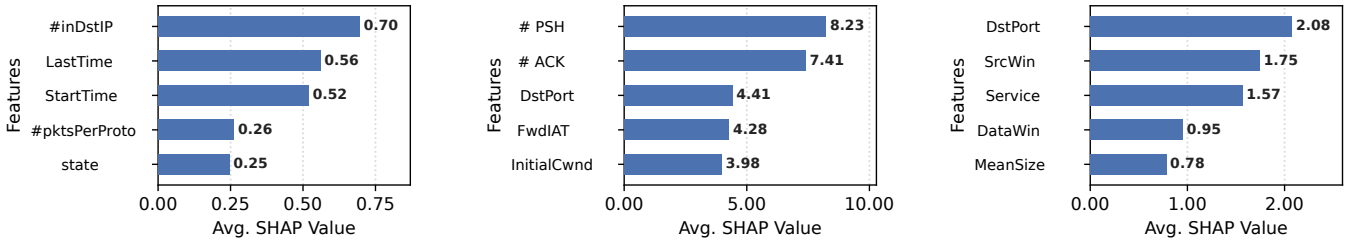
Fig. 2 reports the F1-Score obtained by the final MERGE model as a function of the latent dimensionality $N$, evaluated for different numbers of retained input features $K$ as a result of SHAP. The different settings are tested separately on each dataset and on the unified test set. For reference, a *dataset-specific baseline* as in [24] is also shown, representing a model trained independently per dataset without any feature selection.

Bot-IoT traffic exhibits near-saturated performance across all latent dimensionality, confirming that its discriminative structure is preserved even under full architectural unification. CIC-IDS2017 achieves consistently high F1-scores, with only marginal improvements as the bottleneck size increases, indicating that moderate latent capacity is sufficient to capture the dominant traffic patterns. UNSW-NB15 shows a more




(a) Bot-IoT (b) CIC-IDS (c) UNSW-NB (d) Unified

Fig. 2: MERGE F1-Score vs. latent dimension $N$ for varying top-$K$ SHAP-selected features. Each subplot shows one dataset or the unified test set. Baseline (all features) included for reference.




(a) Bot-IoT (b) CIC-IDS (c) UNSW-NB

Fig. 3: Average SHAP Values across the three considered datasets. This information is used to select the most significant features.

pronounced dependency on the latent dimensionality. Performance improves as $N$ increases from very small to moderate values and stabilizes thereafter. This behavior reflects the higher heterogeneity of the dataset, where overly compact embeddings fail to capture sufficient variability, while larger latent spaces offer diminishing returns. Across all datasets, the gap between MERGE and the dataset-specific baseline remains limited, typically within 1-2% F1-score. This marginal delta is a negligible performance trade-off compared to the operational advantage of protecting three distinct environments with a single model instance. This confirms that the unified model preserves most of the discriminative power of specialized pipelines while enforcing a single shared representation and decision function.

Performance on the unified test set consistently improves with increasing latent dimensionality and rapidly converges to near-optimal values. Notably, strong performance is achieved without requiring large bottleneck sizes, demonstrating that the learned latent space is compact yet expressive.

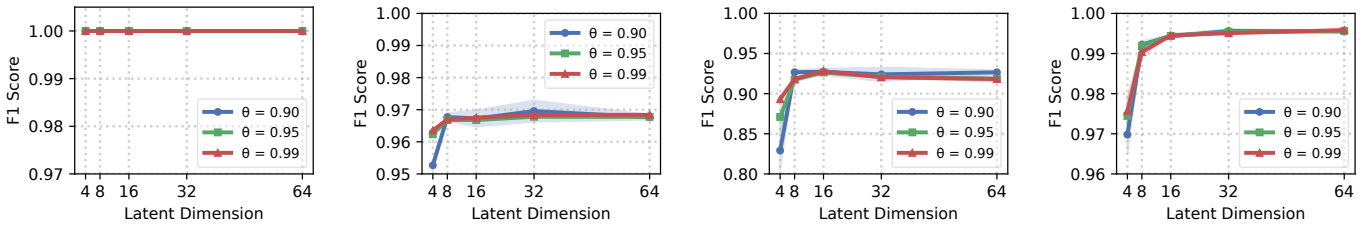
TABLE I: F1-Score Comparison between MERGE and other ML models.

| Dataset | RF | DT | XGB | [21] | Baseline | MERGE |
|---|---|---|---|---|---|---|
| CIC | 0.968 | 0.965 | 0.957 | 0.560 | 0.968 | 0.957 |
| BoT | 0.999 | 0.999 | 0.999 | 1.000 | 1.000 | 1.000 |
| UNSW | 0.911 | 0.929 | 0.887 | 0.770 | 0.919 | 0.911 |
| Unified | - | - | - | - | 0.995 | **0.995** |

Table I compares the best MERGE model ($N = 32$) with selected state-of-the-art approaches. Random Forest (RF), Decision Tree (DT), and XGBoost (XGB) are common models for NAD and are included as strong classical baselines, while the ensemble-based solution by de Oliveira et al. [21] represents a recent cross-dataset NAD approach. MERGE achieves competitive performance across all datasets and consistently outperforms ensemble-based cross-dataset solutions. While classical models trained and tested on the same dataset may achieve slightly higher scores, MERGE is the only method in the comparison that operates with a single instance, without dataset-dependent training or calibration, thus advancing model generalization. This result highlights MERGE's core objective: enabling cross-dataset generalization through architectural unification.

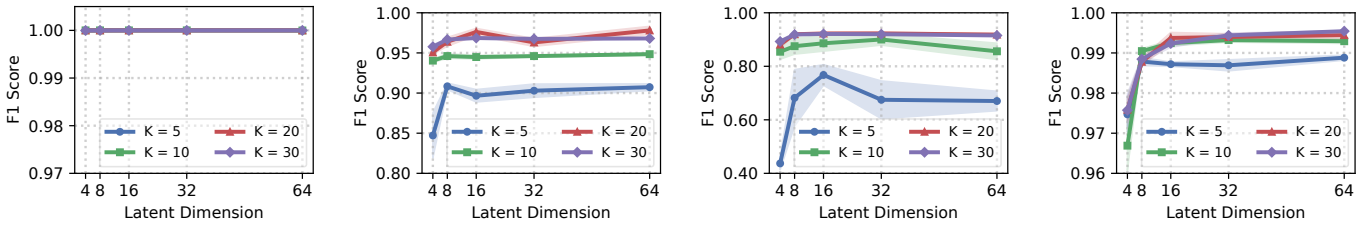
### *D. SHAP Analysis Results*

Fig. 3 presents the top five features for each dataset based on their average contribution to the model's prediction, measured via SHAP values. The values are computed using the baseline model with $N = 32$ and averaged over five independent runs. These graphs provide insights into the most significant features that guide our design. For example, in Bot-IoT, the most influential features are dominated by connection-level statistics, including destination IP counts, flow timing metrics, and protocol-level packet counts. CIC-IDS2017 relies primarily on TCP flag counts and packet inter-arrival times. In contrast, UNSW-NB15 emphasizes transport-layer attributes such as destination port, TCP window size, and service type. These SHAP-based rankings inform the subsequent feature selection process: by identifying a compact set of high-impact features for each dataset, we establish the basis for the unified input representation adopted in MERGE.




(a) Bot-IoT (b) CIC-IDS (c) UNSW-NB (d) Unified

Fig. 4: F1-Score as a function of latent dimension $N$ and cumulative SHAP threshold $\theta$ for each dataset.




(a) Bot-IoT (b) CIC-IDS (c) UNSW-NB (d) Unified

Fig. 5: F1-Score as a function of latent dimension $N$ and top-$K$ SHAP-selected features for each dataset.

### *E. Explainability-Driven Feature Selection Process*

Rather than treating explainability solely as a post-hoc diagnostic, SHAP guides systematic feature selection. We explicitly employ a static Top-$K$ strategy rather than iterative methods (e.g., recursive elimination) to maintain a fixed computational footprint and consistent input dimensionality, which are critical for deployment in real-time network hardware.

*1) Cumulative SHAP thresholding:* Fig. 4 shows the effect of cumulative SHAP-based feature selection. Features are ranked by mean absolute SHAP values, and top-ranked features are retained until their cumulative importance reaches a threshold $\theta$. While this dynamic selection yields dataset-specific feature counts—precluding a single unified model—it serves as a crucial diagnostic step. It demonstrates SHAP's efficacy in preserving baseline performance with a reduced feature subset, providing the empirical justification for the rigid Top-$K$ selection required for final architectural unification.

For Bot-IoT, performance remains saturated across all cumulative thresholds and bottleneck sizes. CIC-IDS2017 is highly robust to cumulative feature reduction: although very compact latent spaces induce a slight performance drop, accuracy quickly stabilizes as latent dimensionality increases, with minimal sensitivity to the threshold, indicating substantial feature redundancy. In contrast, UNSW-NB15 is more sensitive to aggressive pruning under small latent representations, but the performance gap across thresholds narrows as latent dimensionality increases, suggesting that sufficient representational capacity can compensate for reduced input dimensionality. The unified configuration consistently benefits from larger latent spaces and rapidly converges across thresholds. Beyond a moderate bottleneck size, threshold differences become negligible, confirming that SHAP-based feature selection preserves cross-dataset generalization and supports stable, transferable representations.

*2) Top-$K$ Feature Selection:* Fig. 5 evaluates a fixed top-$K$ feature selection strategy guided by SHAP importance scores. Unlike cumulative thresholding, this approach retains a fixed number of the most influential features, enabling further architectural simplification while directly controlling the input dimensionality. The extensive performance curves presented across these experiments serve as a comprehensive robustness ablation study; the visual consistency of the trends across vastly different benchmarks demonstrates that the model maintains a stable decision boundary under aggressive feature reduction.

For Bot-IoT, detection performance remains essentially unaffected even for very small values of $K$. Increasing either the number of retained features or the bottleneck size does not yield measurable performance improvements. CIC-IDS2017 exhibits a clear dependence on the number of selected features: aggressive reduction leads to noticeable performance degradation, particularly with compact latent representations, but performance rapidly stabilizes as $K$ increases. Beyond a moderate feature budget, additional features provide diminishing returns, indicating that most discriminative information is captured by a limited subset of attributes. UNSW-NB15 is more sensitive to feature count, with aggressive pruning causing substantial degradation and increased variance, especially at low latent dimensionalities. As more informative features are retained, performance steadily improves and becomes less dependent on bottleneck size, indicating improved robustness in modeling heterogeneous traffic. The unified configuration consistently benefits from increasing both $K$ and latent dimensionality. While very small feature sets degrade performance under compact bottlenecks, the model quickly converges as either the

feature budget or representational capacity increases. Notably, competitive performance is achieved with moderate values of $K$, confirming that a compact, explainability-guided feature set is sufficient to support effective cross-dataset generalization.

TABLE II: Comparison of mean F1-scores across experimental configurations ($N = 32$)

| **Dataset** | **Baseline** | **Cumulative** ($\theta = 0.90$) | **Top-K** ($K = 30$) | **MERGE** ($K = 30$) |
|---|---|---|---|---|
| CIC | 0.968 | 0.970 | 0.968 | 0.957 |
| BoT | 1.000 | 1.000 | 1.000 | 1.000 |
| UNSW | 0.919 | 0.924 | 0.920 | 0.911 |
| Unified | 0.995 | 0.996 | 0.994 | 0.995 |

*3) Comparative Summary:* Table II compares all methods at the latent dimensionality identified as optimal in prior experiments ($N = 32$). The results show that a single autoencoder architecture achieves detection performance comparable to dataset-specific baselines, with only a marginal F1-score gap of 1–2% relative to specialized models. By leveraging SHAP-guided feature selection with $K = 30$, the unified model effectively suppresses domain-specific noise, enabling robust generalization across heterogeneous traffic. This design substantially reduces computational complexity and model management overhead, demonstrating that architectural simplification can be achieved without sacrificing cross-domain detection accuracy.

## V. Conclusions

This paper presents MERGE, a unified learning model for cross-dataset network anomaly detection that integrates unsupervised representation learning, explainability-driven feature selection, and a shared classifier to achieve robust generalization across heterogeneous network environments. Experimental results show that, by retaining only high-impact features, a single autoencoder–classifier pipeline can match dataset-specific baselines, reducing the parameter footprint by a factor of $M$ (where $M$ is the number of environments) while improving interpretability. Overall, MERGE demonstrates that effective cross-dataset generalization can be achieved through principled representation learning and feature selection, offering a practical and scalable approach to intrusion detection. Future work will explore more expressive feature extractors, alternative classifiers, and extensions to additional feature spaces, including packet-level and protocol-specific representations.

## References


[1] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A.-R. Sadeghi, "Dïot: A federated self-learning anomaly detection system for iot," in *2019 IEEE 39th International conference on distributed computing systems (ICDCS)*. IEEE, 2019, pp. 756–767.

[2] C. Hazman, A. Guezzaz, S. Benkirane, and M. Azrour, "lids-sioel: intrusion detection framework for iot-based smart environments security using ensemble learning," *Cluster Computing*, vol. 26, no. 6, pp. 4069–4083, 2023.

[3] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network intrusion detection for iot security based on learning techniques," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2671–2701, 2019.

[4] K. He, D. D. Kim, and M. R. Asghar, "Adversarial machine learning for network intrusion detection systems: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 538–566, 2023.

[5] I. H. Sarker, "Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective," *SN Computer Science*, vol. 2, no. 3, p. 154, 2021.

[6] D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, and K. Rieck, "Dos and don'ts of machine learning in computer security," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 3971–3988.

[7] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," *Computers & security*, vol. 86, pp. 147–167, 2019.

[8] M. Sarhan, S. Layeghy, N. Moustafa, and M. Portmann, "Netflow datasets for machine learning-based network intrusion detection systems," in *International Conference on Big Data Technologies and Applications*. Springer, 2020, pp. 117–135.

[9] M. Wang, N. Yang, Y. Guo, and N. Weng, "Learn-ids: Bridging gaps between datasets and learning-based network intrusion detection," *Electronics*, vol. 13, no. 6, p. 1072, 2024.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[11] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani *et al.*, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." *ICISSp*, vol. 1, no. 2018, pp. 108–116, 2018.

[12] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

[13] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015, pp. 1–6.

[14] S. Wali, Y. A. Farrukh, I. Khan, and N. D. Bastian, "Meta: Towards a unified, multimodal dataset for network intrusion detection systems," *IEEE data descriptions*, 2024.

[15] H. Hindy, D. Brosset, E. Bayne, A. Seeam, C. Tachtatzis, R. Atkinson, and X. Bellekens, "A taxonomy and survey of intrusion detection system design techniques, network threats and datasets," *arXiv preprint arXiv:1806.03517*, 2018.

[16] M. Cantone, C. Marrocco, and A. Bria, "Machine learning in network intrusion detection: A cross-dataset generalization study," *IEEE Access*, 2024.

[17] S. Layeghy and M. Portmann, "On generalisability of machine learning-based network intrusion detection systems," *arXiv preprint arXiv:2205.04112*, 2022.

[18] M. Iwanowski, D. Olszewski, W. Graniszewski, J. Krupski, and F. Pelc, "The choice of training data and the generalizability of machine learning models for network intrusion detection systems," *Applied Sciences*, vol. 15, no. 15, p. 8466, 2025.

[19] S. Layeghy, M. Baktashmotlagh, and M. Portmann, "Di-nids: Domain invariant network intrusion detection system," *Knowledge-Based Systems*, vol. 273, p. 110626, 2023.

[20] N. Niknami, V. Mahzoon, and J. Wu, "Ptn-ids: Prototypical network solution for the few-shot detection in intrusion detection systems," in *2024 IEEE 49th Conference on Local Computer Networks (LCN)*. IEEE, 2024, pp. 1–9.

[21] V. M. De Oliveira, H. M. De Oliveira, G. M. Santos, J. Geremias, and E. K. Viegas, "A big data framework for scalable and cross-dataset capable machine learning in network intrusion detection systems," *IEEE Access*, 2025.

[22] J. Zhai, S. Zhang, J. Chen, and Q. He, "Autoencoder and its various variants," in *2018 IEEE international conference on systems, man, and cybernetics (SMC)*. IEEE, 2018, pp. 415–419.

[23] C. Zilli, A. Sacco, F. Esposito, and G. Marchetto, "Clearnet: Enhancing transparency in opaque network models using explainable ai (xai) for efficient traffic engineering," *IEEE Transactions on Network and Service Management*, vol. 22, no. 4, pp. 3617–3631, 2025.

[24] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, 2014, pp. 4–11.