Digital Transformation
RESEARCH SOCIETY

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

The First International Conference on Security and Cybersecurity in the AI and Digital Context
CYBERSEC 2026 (https://www.dtrsociety.org/cybersec2026/)

PROF. DR. ALEXANDER LAWALL

Reliable Intrusion Detection of Zero-Day Attacks:

A Two-Stage Architecture Combining Deep Learning and Conformal AI

Keynote
Porto, June 2026

PROF. DR. ALEXANDER LAWALL

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Academic Roles

- Program Director, B.Sc. & M.Sc. Cyber Security and Cyber Security Management
- Professor in Cyber Security (Distance & On-site Learning)

Expertise

- System & Network Security
- Web Application & Cloud Security
- IoT and Industrial IT Security

Professional Affiliations

- Leadership Committee, "Management of Information Security" (Society for Informatics, GI)
- Scientific Advisory Board (DTR Society)
- Conference Chair of the CYBERSEC (DTR Society)
- Professional Lead, "Security & GRC in IT" (Summit Leipzig)
- Member, Association of Cyber Forensics and Threat Investigators (ACFTI)
- Member, Zentrum Digitalisierung Bayern (ZD.B)
- Conference Committees Board Chair (IARIA)
- Steering Committee of the Conference SECURWARE, CYBER & IoT-AI (IARIA)

Research & Publications

- Focus Areas: Cyber Security, Information Security, Industry 4.0/5.0, IoT, Rights Management, AI in Cyber Security
- Publications in national/international Journals and Conferences
- Keynote Speaker, Program Chair, Panel Expert of International Conferences

AGENDA

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Introduction and Motivation	**1**
Part I – The Generalization Crisis	**2**
Part II – Trustworthy AI & Foundation Models	**3**
Part III – Zero-Day Adaptive Defense	**4**
Conclusion and Future Work	**5**

Why IDS fail?

How can AI express uncertainty?

Trustworthy adaptive zero-day defense systems (two-stage architecture)

INTRODUCTION AND MOTIVATION

Network Intrusion Detection

„A **Network Intrusion Detection** System (NIDS) is a software that performs packet sniffing and **network traffic analysis** to **identify suspicious activity** and **record relevant information**.“

Quelle: https://csrc.nist.gov/glossary/term/NIDS & (NISTIR 8183A Vol. 2, NISTIR 8183A Vol. 3)

INTRODUCTION AND MOTIVATION

Zero-Day Attacks

„A **zero-day attack** is a **new type of cyberattack** that uses vulnerabilities that have **not** been **publicly identified**. These attacks are often **undetected** because they **use new methods** that **avoid detection** by existing security tools. This allows attackers to target specific systems **without being noticed**.“

Quellen:
Singh, Umesh Kumar, Chanchala Joshi, and Suyash Kumar Singh. "Zero day attacks defense technique for protecting system against unknown vulnerabilities."
International Journal of Scientific Research in Computer Science and Engineering 5.1 (2017): 13-18.

Bilge, Leyla, and Tudor Dumitraş. "Before we knew it: an empirical study of zero-day attacks in the real world."
Proceedings of the 2012 ACM conference on Computer and communications security. 2012.

Guo, Yang. "A review of machine learning-based zero-day attack detection: Challenges and future directions."
Computer communications 198 (2023): 175-185.

INTRODUCTION AND MOTIVATION

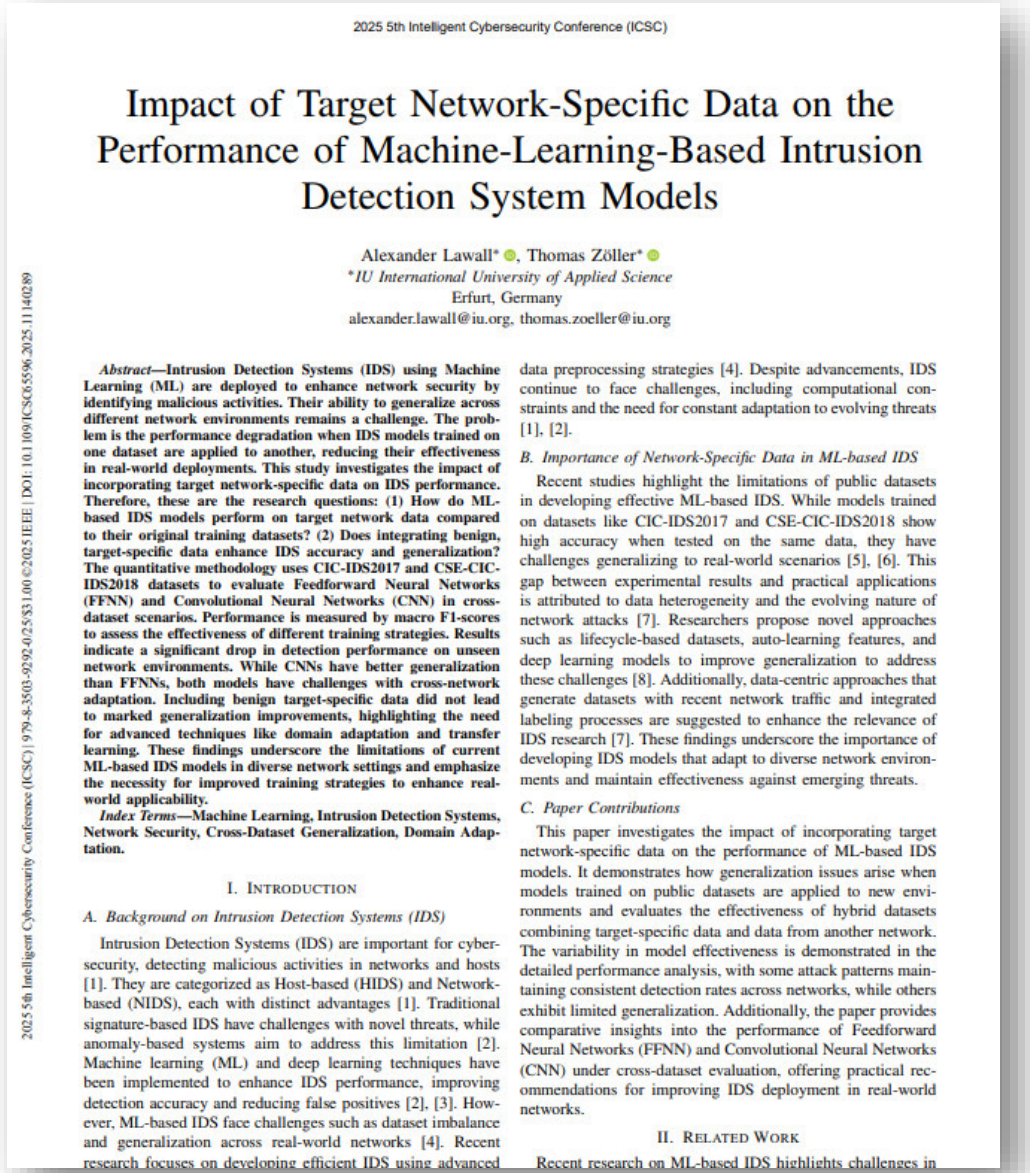
Fundamental Problems of Modern AI-based NIDS

1. **Generalization Failure** → *Why classical ML-based IDS fail in real-world environments*

2. **Overconfidence under Distribution Shift** → *How Foundation Models and Conformal Prediction make uncertainty visible and actionable*

3. **The Zero-Day Adaptation Bottleneck** → *How these concepts lead to a practical, adaptive, trustworthy NIDS architecture*

PART I – THE GENERALIZATION CRISIS



2025 5th Intelligent Cybersecurity Conference (ICSC)

Impact of Target Network-Specific Data on the Performance of Machine-Learning-Based Intrusion Detection System Models

Alexander Lawall* , Thomas Zöller*
*IU International University of Applied Science
Erfurt, Germany
alexander.lawall@iu.org, thomas.zoeller@iu.org

2025 5th Intelligent Cybersecurity Conference (ICSC) | 979-8-3503-9292-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICSC65596.2025.11140289

***Abstract*—Intrusion Detection Systems (IDS) using Machine Learning (ML) are deployed to enhance network security by identifying malicious activities. Their ability to generalize across different network environments remains a challenge. The problem is the performance degradation when IDS models trained on one dataset are applied to another, reducing their effectiveness in real-world deployments. This study investigates the impact of incorporating target network-specific data on IDS performance. Therefore, these are the research questions: (1) How do ML-based IDS models perform on target network data compared to their original training datasets? (2) Does integrating benign, target-specific data enhance IDS accuracy and generalization? The quantitative methodology uses CIC-IDS2017 and CSE-CIC-IDS2018 datasets to evaluate Feedforward Neural Networks (FFNN) and Convolutional Neural Networks (CNN) in cross-dataset scenarios. Performance is measured by macro F1-scores to assess the effectiveness of different training strategies. Results indicate a significant drop in detection performance on unseen network environments. While CNNs have better generalization than FFNNs, both models have challenges with cross-network adaptation. Including benign target-specific data did not lead to marked generalization improvements, highlighting the need for advanced techniques like domain adaptation and transfer learning. These findings underscore the limitations of current ML-based IDS models in diverse network settings and emphasize the necessity for improved training strategies to enhance real-world applicability.**

***Index Terms*—Machine Learning, Intrusion Detection Systems, Network Security, Cross-Dataset Generalization, Domain Adaptation.**

I. Introduction

A. Background on Intrusion Detection Systems (IDS)

Intrusion Detection Systems (IDS) are important for cybersecurity, detecting malicious activities in networks and hosts [1]. They are categorized as Host-based (HIDS) and Network-based (NIDS), each with distinct advantages [1]. Traditional signature-based IDS have challenges with novel threats, while anomaly-based systems aim to address this limitation [2]. Machine learning (ML) and deep learning techniques have been implemented to enhance IDS performance, improving detection accuracy and reducing false positives [2], [3]. However, ML-based IDS face challenges such as dataset imbalance and generalization across real-world networks [4]. Recent research focuses on developing efficient IDS using advanced data preprocessing strategies [4]. Despite advancements, IDS continue to face challenges, including computational constraints and the need for constant adaptation to evolving threats [1], [2].

B. Importance of Network-Specific Data in ML-based IDS

Recent studies highlight the limitations of public datasets in developing effective ML-based IDS. While models trained on datasets like CIC-IDS2017 and CSE-CIC-IDS2018 show high accuracy when tested on the same data, they have challenges generalizing to real-world scenarios [5], [6]. This gap between experimental results and practical applications is attributed to data heterogeneity and the evolving nature of network attacks [7]. Researchers propose novel approaches such as lifecycle-based datasets, auto-learning features, and deep learning models to improve generalization to address these challenges [8]. Additionally, data-centric approaches that generate datasets with recent network traffic and integrated labeling processes are suggested to enhance the relevance of IDS research [7]. These findings underscore the importance of developing IDS models that adapt to diverse network environments and maintain effectiveness against emerging threats.

C. Paper Contributions

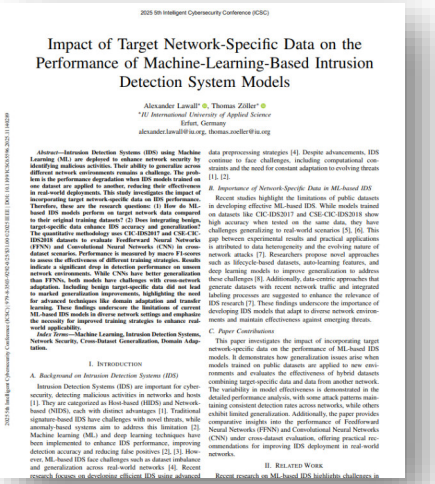
This paper investigates the impact of incorporating target network-specific data on the performance of ML-based IDS models. It demonstrates how generalization issues arise when models trained on public datasets are applied to new environments and evaluates the effectiveness of hybrid datasets combining target-specific data and data from another network. The variability in model effectiveness is demonstrated in the detailed performance analysis, with some attack patterns maintaining consistent detection rates across networks, while others exhibit limited generalization. Additionally, the paper provides comparative insights into the performance of Feedforward Neural Networks (FFNN) and Convolutional Neural Networks (CNN) under cross-dataset evaluation, offering practical recommendations for improving IDS deployment in real-world networks.

II. Related Work

Recent research on ML-based IDS highlights challenges in

Quelle: A. Lawall and T. Zöller, “Impact of Target Network-Specific Data on the Performance of Machine-Learning-Based Intrusion Detection System Models”,
2025 5th Intelligent Cybersecurity Conference (ICSC), Tampa, FL, USA, 2025, pp. 290-297, doi: 10.1109/ICSC65596.2025.11140289.

PART I – THE GENERALIZATION CRISIS



RQ1: *How do ML-based IDS models perform on target network data compared to their original training datasets?*

RQ2: *Does integrating benign, target-specific data enhance IDS accuracy and generalization?*

"High benchmark accuracy does NOT imply real-world security"

PART I – THE GENERALIZATION CRISIS

Datasets

CIC-IDS2017:

- ~2 million network flows
- Simpler network setup
- Used as target network in hybrid approach

CSE-CIC-IDS2018:

- ~16 million network flows (reduced by 95% benign traffic)
- Complex infrastructure (420 machines)
- Attack source in hybrid approach

PART I – THE GENERALIZATION CRISIS

Data Processing & Feature Selection

Preprocessing:

- Reduced benign traffic in CSE-CIC-IDS2018 by 95%
- Selected 7 attack categories present in both datasets
- Excluded attacks with fewer than 100 samples
- Removed incomplete records and handled infinite values

Feature Selection:

- Combined chi-squared tests and mutual information scores
- Applied correlation threshold of $|0.7|$ to reduce multicollinearity
- Selected 39 features from original 81 network flow features

PART I – THE GENERALIZATION CRISIS

Experimental Setup: Model Architectures

FFNN:

- 2 dense layers (64, 32 neurons)
- Dropout (30%)
- L2 regularization
- ReLU activation

CNN:

- 3 convolutional layer
- Decreasing filters (120, 60, 30)
- Increasing kernel sizes (2, 3, 4)
- ReLU activation

PART I – THE GENERALIZATION CRISIS

Experimental Setup: Configurations

Baseline:

- Train and test on same dataset

Cross-Network:

- Train on A, test on B
- Train on B, test on A

Hybrid Approach:

- Benign traffic from target
- Attack traffic from source

PART I – THE GENERALIZATION CRISIS

Results: Baseline Results

Both architectures showed excellent performance when trained and tested on the same dataset:

Model	CIC-IDS2017	CSE-CIC-IDS2018
FFNN	0.98	0.98
CNN	1.00	1.00

- Strong detection capabilities across all attack categories
- CNN showed marginally better consistency across attack types
- High F1-scores for both malicious and benign traffic

PART I – THE GENERALIZATION CRISIS

Results: Cross-Network Generalization

Notable performance degradation when models were tested on different network environments:

Model	Training Set	Test Set	F1-Score
FFNN	CIC-IDS2017	CSE-CIC-IDS2018	0.36
FFNN	CSE-CIC-IDS2018	CIC-IDS2017	0.31
CNN	CIC-IDS2017	CSE-CIC-IDS2018	0.46
CNN	CSE-CIC-IDS2018	CIC-IDS2017	0.52

- CNN shows better generalization (avg. F1-score: 0.495) compared to FFNN (avg. F1-score: 0.345)

PART I – THE GENERALIZATION CRISIS

Results: Hybrid Approach Results

Integrating benign traffic from target network with attack traffic from source network:

Model	F1-Score	Comparison to Cross-Network
FFNN	0.35	Similar (0.31-0.36)
CNN	0.52	Similar (0.46-0.52)

- No significant improvement over direct cross-network testing
- Hybrid approach did not solve the generalization challenge
- CNN maintains better performance than FFNN

PART I – THE GENERALIZATION CRISIS

Results: Attack-specific Analysis

Poor Generalization:

- Botnet Ares (F1≈0)
- DDoS-LOIC-HTTP (F1≈0)
- DoS GoldenEye (F1=0.05-0.50)
- DoS Hulk (F1=0-0.75)

Better Generalization:

- DoS Slowloris (F1=0.78-0.97)
- Infiltration-NMAP (F1=0.33-0.92)
- SSH-BruteForce (F1=0.66-1.00)

PART I – THE GENERALIZATION CRISIS

Results: Attack-specific Analysis: Feature Distribution

DDoS-LOIC-HTTP:

(High variability between datasets)

- Mean KS statistic: 0.380
- Standard deviation: 0.411

DoS Slowloris:

(more consistent)

- Mean KS statistic: 0.233
- Standard deviation: 0.229

PART I – THE GENERALIZATION CRISIS

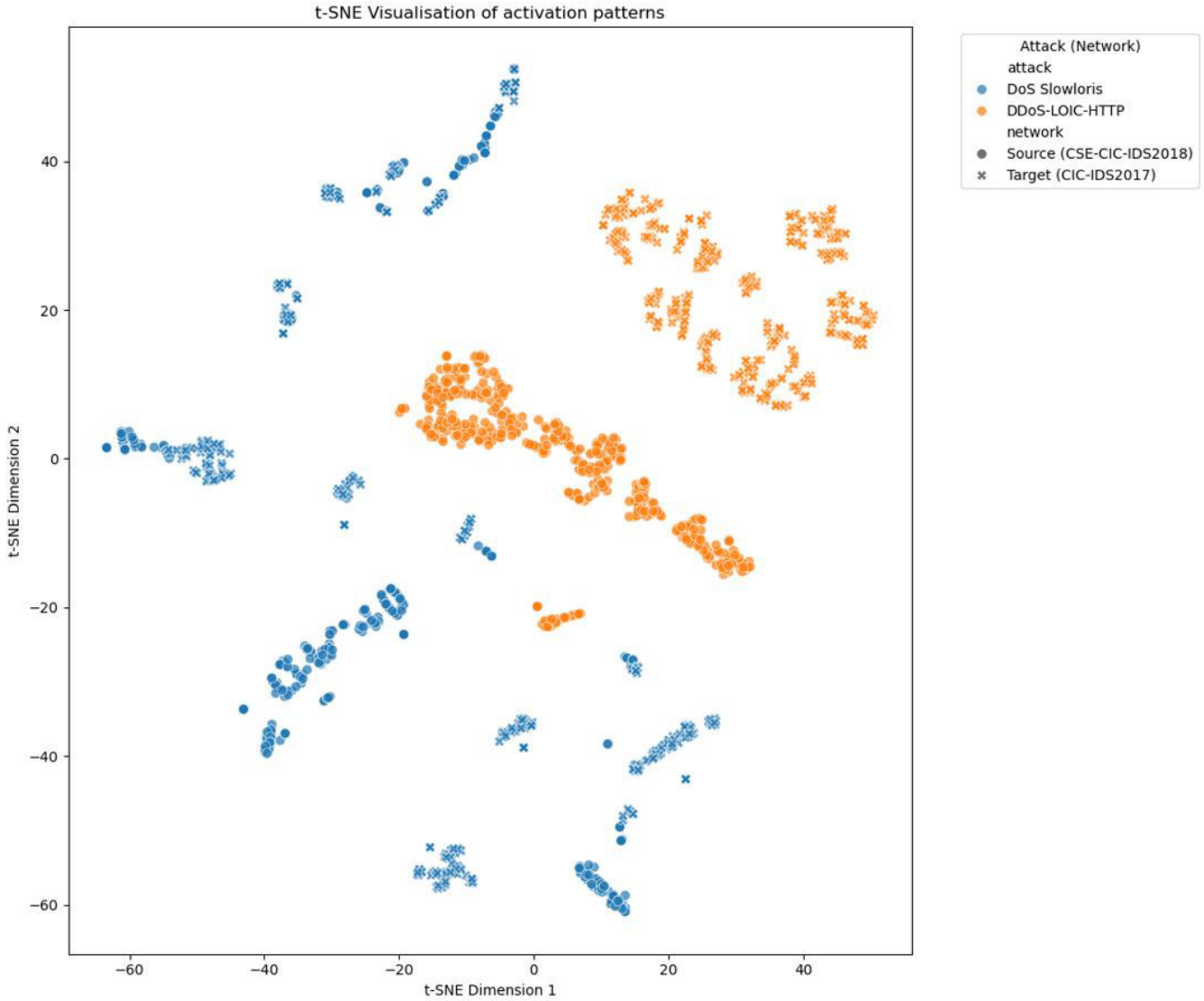
Results: Attack-specific Analysis: Network Activation

Metric	DDoS-LOIC-HTTP	DoS Slowloris
Pearson Correl. Coef.	0.346	0.944
Cosine Similarity	0.377	0.950
Euclidean Distance	12.397	2.544

- DoS Slowloris shows remarkably consistent internal representations across networks
- DDoS-LOIC-HTTP exhibits network-specific rather than attack-specific patterns

PART I – THE GENERALIZATION CRISIS

Results: Attack-specific Analysis



PART I – THE GENERALIZATION CRISIS

Conclusion

- **Substantial performance drop when IDS models are deployed to new environments**
  - FFNN F1-scores drop from 0.98 to 0.31-0.36
  - CNN F1-scores drop from 1.00 to 0.46-0.52

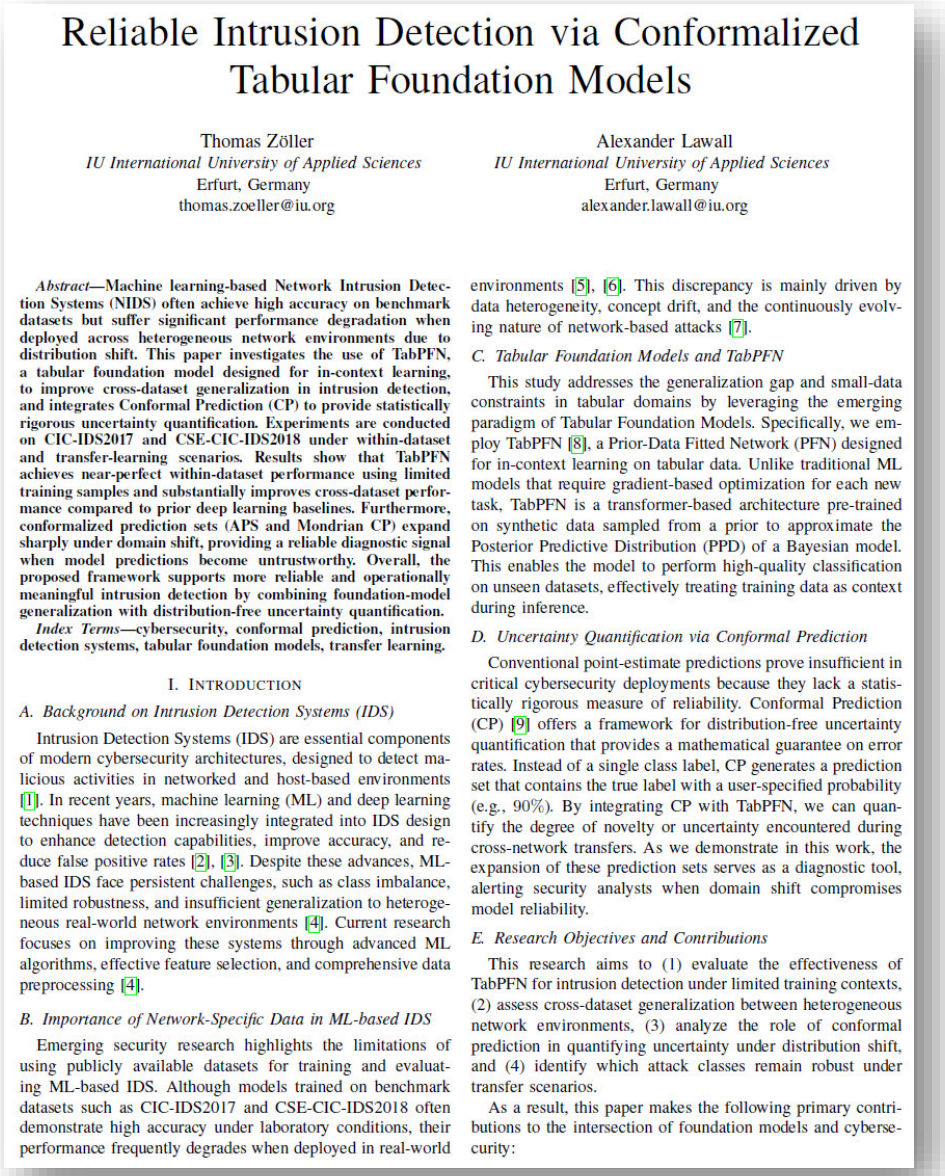
- **CNN models generalize better than FFNN models**

- **Hybrid approach (target-specific benign data) provides limited improvement**

- **Attack detection generalization varies remarkably by attack type**
  - Some attacks maintain consistent detection rates (e.g., DoS Slowloris)
  - Others exhibit limited generalization (e.g., DDoS-LOIC-HTTP)

"At this point, we realized: the problem was not only detection performance. The real problem was that the model had no idea when it was wrong."

PART II – TRUSTWORTHY AI & FOUNDATION MODELS



Reliable Intrusion Detection via Conformalized Tabular Foundation Models

Thomas Zöller
IU International University of Applied Sciences
Erfurt, Germany
thomas.zoeller@iu.org

Alexander Lawall
IU International University of Applied Sciences
Erfurt, Germany
alexander.lawall@iu.org

***Abstract*—Machine learning-based Network Intrusion Detection Systems (NIDS) often achieve high accuracy on benchmark datasets but suffer significant performance degradation when deployed across heterogeneous network environments due to distribution shift. This paper investigates the use of TabPFN, a tabular foundation model designed for in-context learning, to improve cross-dataset generalization in intrusion detection, and integrates Conformal Prediction (CP) to provide statistically rigorous uncertainty quantification. Experiments are conducted on CIC-IDS2017 and CSE-CIC-IDS2018 under within-dataset and transfer-learning scenarios. Results show that TabPFN achieves near-perfect within-dataset performance using limited training samples and substantially improves cross-dataset performance compared to prior deep learning baselines. Furthermore, conformalized prediction sets (APS and Mondrian CP) expand sharply under domain shift, providing a reliable diagnostic signal when model predictions become untrustworthy. Overall, the proposed framework supports more reliable and operationally meaningful intrusion detection by combining foundation-model generalization with distribution-free uncertainty quantification.**

***Index Terms*—cybersecurity, conformal prediction, intrusion detection systems, tabular foundation models, transfer learning.**

I. Introduction

A. Background on Intrusion Detection Systems (IDS)

Intrusion Detection Systems (IDS) are essential components of modern cybersecurity architectures, designed to detect malicious activities in networked and host-based environments [1]. In recent years, machine learning (ML) and deep learning techniques have been increasingly integrated into IDS design to enhance detection capabilities, improve accuracy, and reduce false positive rates [2], [3]. Despite these advances, ML-based IDS face persistent challenges, such as class imbalance, limited robustness, and insufficient generalization to heterogeneous real-world network environments [4]. Current research focuses on improving these systems through advanced ML algorithms, effective feature selection, and comprehensive data preprocessing [4].

B. Importance of Network-Specific Data in ML-based IDS

Emerging security research highlights the limitations of using publicly available datasets for training and evaluating ML-based IDS. Although models trained on benchmark datasets such as CIC-IDS2017 and CSE-CIC-IDS2018 often demonstrate high accuracy under laboratory conditions, their performance frequently degrades when deployed in real-world environments [5], [6]. This discrepancy is mainly driven by data heterogeneity, concept drift, and the continuously evolving nature of network-based attacks [7].

C. Tabular Foundation Models and TabPFN

This study addresses the generalization gap and small-data constraints in tabular domains by leveraging the emerging paradigm of Tabular Foundation Models. Specifically, we employ TabPFN [8], a Prior-Data Fitted Network (PFN) designed for in-context learning on tabular data. Unlike traditional ML models that require gradient-based optimization for each new task, TabPFN is a transformer-based architecture pre-trained on synthetic data sampled from a prior to approximate the Posterior Predictive Distribution (PPD) of a Bayesian model. This enables the model to perform high-quality classification on unseen datasets, effectively treating training data as context during inference.

D. Uncertainty Quantification via Conformal Prediction

Conventional point-estimate predictions prove insufficient in critical cybersecurity deployments because they lack a statistically rigorous measure of reliability. Conformal Prediction (CP) [9] offers a framework for distribution-free uncertainty quantification that provides a mathematical guarantee on error rates. Instead of a single class label, CP generates a prediction set that contains the true label with a user-specified probability (e.g., 90%). By integrating CP with TabPFN, we can quantify the degree of novelty or uncertainty encountered during cross-network transfers. As we demonstrate in this work, the expansion of these prediction sets serves as a diagnostic tool, alerting security analysts when domain shift compromises model reliability.

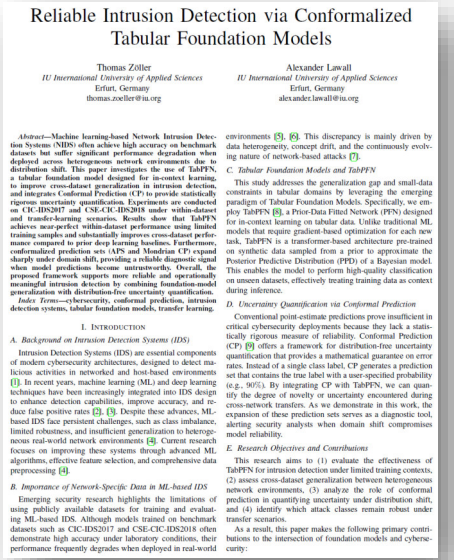
E. Research Objectives and Contributions

This research aims to (1) evaluate the effectiveness of TabPFN for intrusion detection under limited training contexts, (2) assess cross-dataset generalization between heterogeneous network environments, (3) analyze the role of conformal prediction in quantifying uncertainty under distribution shift, and (4) identify which attack classes remain robust under transfer scenarios.

As a result, this paper makes the following primary contributions to the intersection of foundation models and cybersecurity:

Quelle: T. Zöller and A. Lawall, “Reliable Intrusion Detection via Conformalized Tabular Foundation Models”, 2026 IEEE International Conference on Cyber Security and Resilience (IEEE CSR), Lisbon, Portugal, 2026.

PART II – TRUSTWORTHY AI & FOUNDATION MODELS



Reliable Intrusion Detection via Conformalized Tabular Foundation Models

Thomas Zöller
IU International University of Applied Sciences
Erfurt, Germany
thomas.zoeller@iu.org

Alexander Lawall
IU International University of Applied Sciences
Erfurt, Germany
alexander.lawall@iu.org

***Abstract*—Machine learning-based Network Intrusion Detection Systems (NIDS) often achieve high accuracy on benchmark datasets but suffer significant performance degradation when deployed across heterogeneous network environments due to distribution shift. This paper investigates the use of TabPFN, a tabular foundation model designed for in-context learning, to improve cross-dataset generalization in intrusion detection, and integrates Conformal Prediction (CP) to provide statistically rigorous uncertainty quantification. Experiments are conducted on CIC-IDS2017 and CSE-CIC-IDS2018 under within-dataset and transfer-learning scenarios. Results show that TabPFN achieves near-perfect within-dataset performance using limited training samples and substantially improves cross-dataset performance compared to prior deep learning baselines. Furthermore, conformalized prediction sets (APS and Mondrian CP) expand sharply under domain shift, providing a reliable diagnostic signal when model predictions become untrustworthy. Overall, the proposed framework supports more reliable and operationally meaningful intrusion detection by combining foundation-model generalization with distribution-free uncertainty quantification.**

***Index Terms*—cybersecurity, conformal prediction, intrusion detection systems, tabular foundation models, transfer learning.**

RO1: *Evaluate the effectiveness of TabPFN for intrusion detection under limited training contexts*

RO2: *Assess cross-dataset generalization between heterogeneous network environments*

RO3: *Analyze the role of conformal prediction in quantifying uncertainty under distribution shift*

RO4: *Identify which attack classes remain robust under transfer scenarios*

“Cybersecurity AI must not only predict – it must communicate uncertainty.”

PART II – TRUSTWORTHY AI & FOUNDATION MODELS

TabPFN and Conformal Prediction

Main idea:

“Instead of retraining for every new task, the model performs inference through in-context learning.”

Conformal Prediction (Simple):

Traditional AI says: ‘This is an attack.’

Conformal AI says: ‘I am 90% confident that the correct label lies within this prediction set.’

PART II – TRUSTWORTHY AI & FOUNDATION MODELS

Results: TabPFN

TABLE I: Performance Metrics for Same-Network Scenarios

Metric	CIC-IDS2017	CSE-CIC-IDS2018
Accuracy	0.9975	0.9971
Macro F1-Score	1.00	1.00

TABLE II: Cross-Dataset F1-Score Comparison

Study	Architecture	Datasets	F1-Score
This Study	**TabPFN**	**17/18**	**0.62**
Lawall and Zöller '25 [12]	FFNN/CNN	17/18	0.31/0.52
Al-Riyami et al. [29]	ANN/CNN	NSL/gure	0.43/0.42

TABLE III: Per-Class Cross-Network Performance

Attack Type	Precision	Recall	F1-Score
BENIGN	0.58	1.00	0.73
Botnet Ares	1.00	1.00	1.00
DDoS-LOIC-HTTP	0.00	0.00	0.00
DoS GoldenEye	1.00	0.25	0.40
DoS Hulk	0.00	0.00	0.00
DoS Slowloris	0.95	0.99	0.97
Infiltration-NMAP	1.00	0.85	0.92
SSH-BruteForce	1.00	0.96	0.98

PART II – TRUSTWORTHY AI & FOUNDATION MODELS

Results: Conformal Prediction (& TabPFN)

***Experiment 4**: (IDS17) Adaptive Prediction Sets (APS) and Mondrian variants to establish baseline prediction set sizes.*

***Experiment 5**: (IDS18) establish baseline uncertainty metrics on the more complex topology.*

***Experiment 6 [**transfer learning scenario**]**: (IDS17 to IDS18) analyze how domain shift affects prediction set expansion and reliability.*

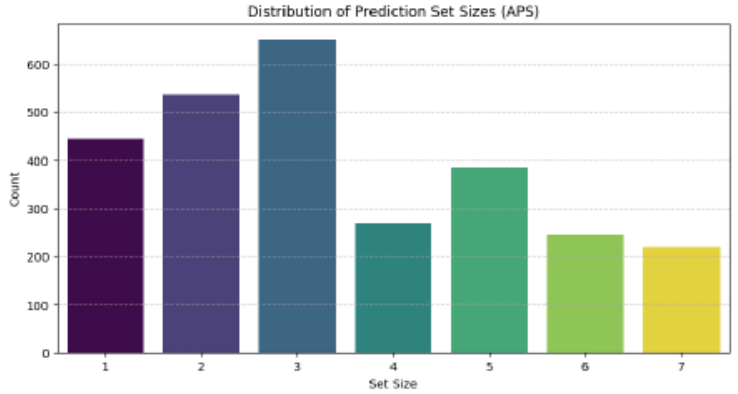
TABLE IV: Median Prediction Set Sizes Across All Experiments ($\alpha = 0.1$)

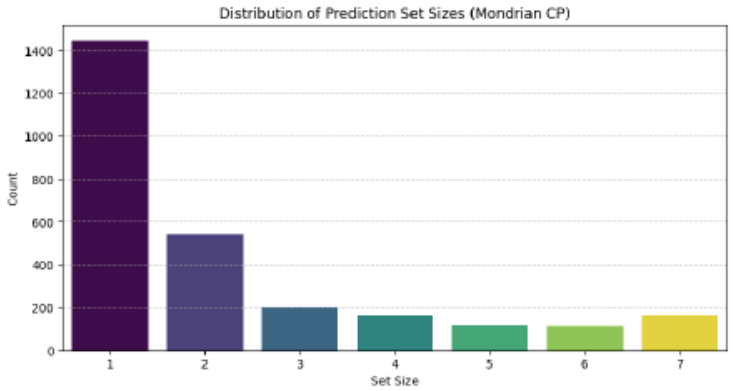
Attack Type	IDS17 (Exp 4)		IDS18 (Exp 5)		Transfer (Exp 6)	
	APS	Mondrian	APS	Mondrian	APS	Mondrian
BENIGN	3	2	3	2	3	4
Botnet Ares	3	1	4	1	7	1
DDoS-LOIC-HTTP	2	2	2	1	7	7
DoS GoldenEye	1	3	1	3	7	7
DoS Hulk	1	3	3	1	4	5
DoS Slowloris	4	3	2	3	7	1
Infiltration-NMAP	5	1	5	1	6	6
SSH-BruteForce	3	1	1	1	7	1

TABLE IV: Median Prediction Set Sizes Across All Experiments ($\alpha = 0.1$)

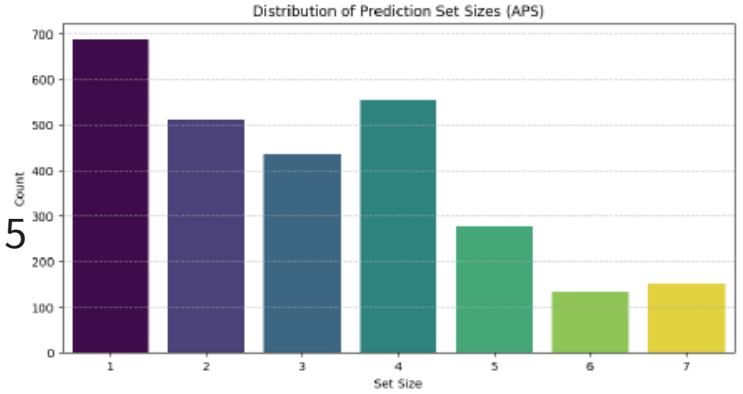
Attack Type	IDS17 (Exp 4)		IDS18 (Exp 5)		Transfer (Exp 6)	
	APS	Mondrian	APS	Mondrian	APS	Mondrian
BENIGN	3	2	3	2	3	4
Botnet Ares	3	1	4	1	7	1
DDoS-LOIC-HTTP	2	2	2	1	7	7
DoS GoldenEye	1	3	1	3	7	7
DoS Hulk	1	3	3	1	4	5
DoS Slowloris	4	3	2	3	7	1
Infiltration-NMAP	5	1	5	1	6	6
SSH-BruteForce	3	1	1	1	7	1

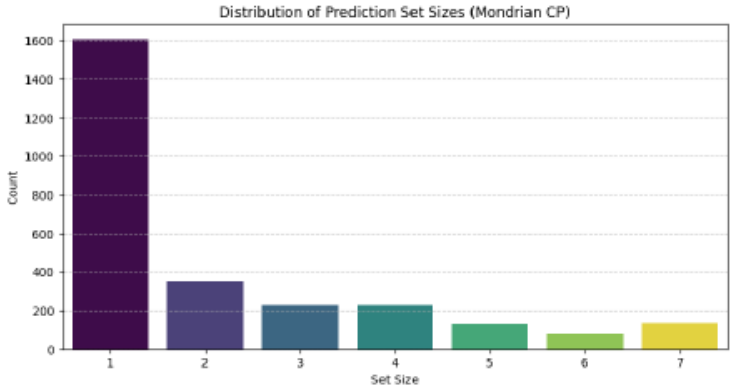
Exp. 4



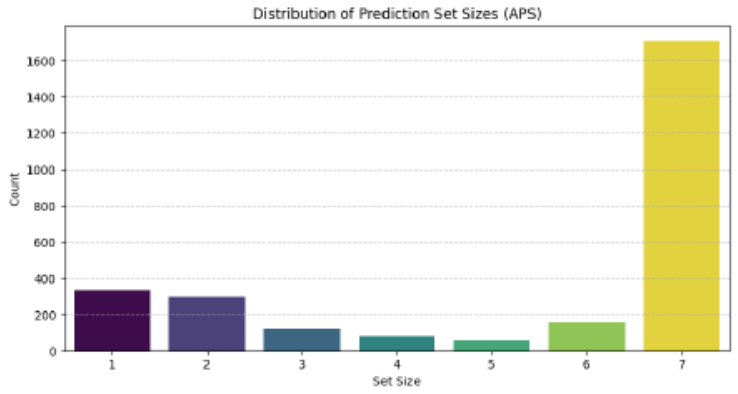


Exp. 5





Exp. 6



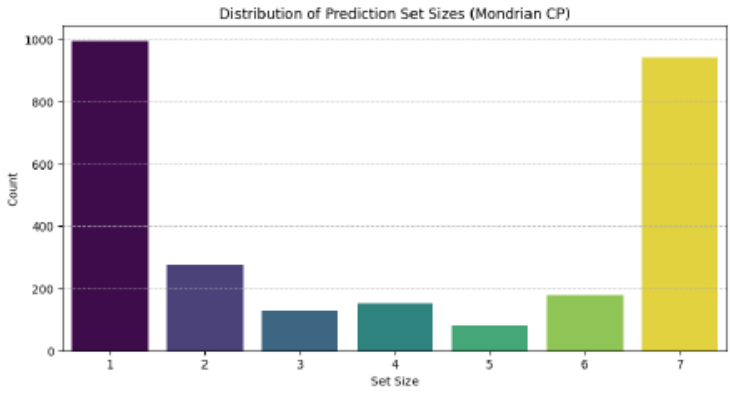
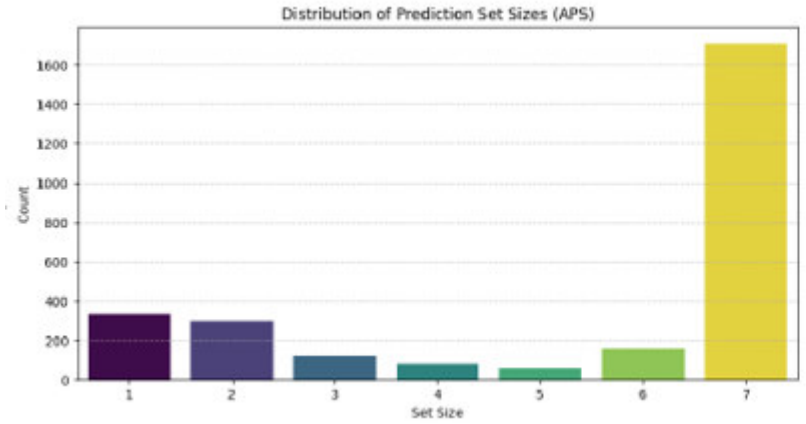


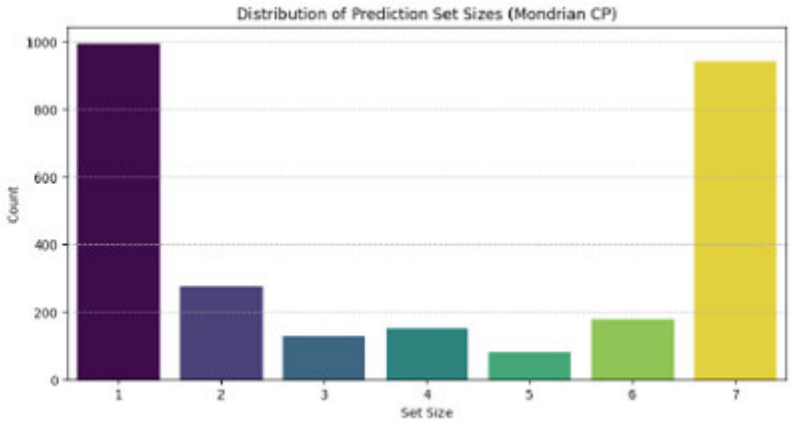
Fig. 1: Histograms of CP set sizes for within-dataset and cross-dataset evaluation scenarios: Each row corresponds to a distinct experiment (Exp. 4–6), with columns comparing the APS (left) and Mondrian (right) conformal prediction strategies.

PART II – TRUSTWORTHY AI & FOUNDATION MODELS

Conclusion

- **Prediction Set Expansion under Domain Shift (Transfer Learning)**





"Uncertainty becomes a security signal:
A mathematically grounded indicator of distributional shift."

PART III – ZERO-DAY ADAPTIVE DEFENSE

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Trustworthy Zero-Day Detection in NIDS via Mondrian Conformal Gating and In-Context Learning

Anonymized Authors
University
City, State, Country
name@university.org

Abstract

Modern Network Intrusion Detection Systems (NIDS) rely on high-throughput tree-based ensembles for real-time traffic analysis. However, these models suffer from uncalibrated overconfidence when encountering Out-of-Distribution (OOD) Zero-Day attacks. We present a hybrid architecture that leverages Mondrian Conformal Gating to detect novel threats using behavioral network semantics. Our system employs a LightGBM "Fast-Path" for known traffic, providing rigorous marginal and category-conditional coverage guarantees. We demonstrate that while known traffic yields tight prediction sets ($|\hat{C}| \approx 1$), novel attacks trigger significant expansion in set size ($|\hat{C}| \gg 1$), serving as a mathematically grounded gating mechanism. Critical anomalies are routed to a Tabular Foundation Model (TabPFN), which, via In-Context Learning (ICL), enables rapid adaptation to novel attack vectors. Our evaluation on behavioral NIDS benchmarks shows that this architecture achieves near-instantaneous mitigation with substantially higher sample efficiency than traditional model retraining, while preserving the reliability and high-throughput performance of the primary classifier.

Keywords

Network Intrusion Detection, Conformal Prediction, Zero-Day Mitigation, Tabular Foundation Models, In-Context Learning, Mondrian Conformal Gating, Machine Learning Security

ACM Reference Format:

Anonymized Authors. 2026. Trustworthy Zero-Day Detection in NIDS via Mondrian Conformal Gating and In-Context Learning. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (ACM CSS '26)*. ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The integrity of modern cybersecurity architectures is fundamentally dependent on the efficacy of Intrusion Detection Systems (IDS) [19]. As malicious activities grow in complexity, the integration of machine learning (ML) and deep learning (DL) has become standard practice to enhance detection capabilities and reduce the burden of manual alert validation [2, 8]. In the context of high-throughput network monitoring, tree-based ensembles, specifically Gradient Boosted Decision Trees (GBDTs) like LightGBM, have emerged as the preferred baseline due to their superior performance on tabular data and their low-latency inference profiles [10, 18].

1.1 The Reliability Gap in NIDS

Despite these advancements, ML-based NIDS consistently face challenges regarding robustness and generalization [17]. Most current systems are trained on static benchmarks, yet they are deployed into highly non-stationary network environments characterized by continuous distribution shifts and evolving attack patterns [7]. As established in recent literature, models that exhibit near-perfect accuracy under controlled experimental conditions frequently experience catastrophic degradation when exposed to unseen network topologies or novel Zero-Day exploits [5, 6].

A critical failure mode of these standard classifiers is their inherent lack of self-awareness. When presented with Out-of-Distribution (OOD) traffic, GBDTs tend to project novel samples onto established partitions of the feature space with high confidence [9]. This results in "silent failures", where the system issues authoritative but incorrect classifications for novel threats. In high-stakes security operations, a point-estimate prediction without a rigorous measure of its own reliability is insufficient [15].

Problem Definition. Let $(x, y) \sim P_{\text{ID}}$ denote in-distribution network flows and labels, and let $x^\star \sim P_{\text{OOD}}$ denote zero-day traffic drawn from an unknown distribution. The objective is to design a classifier f and uncertainty estimator $\mathcal{U}$ such that:

- $f(x)$ is accurate for $x \sim P_{\text{ID}}$
- $\mathcal{U}(x^\star)$ is high for $x^\star \sim P_{\text{OOD}}$
- adaptation to P_{OOD} requires minimal labeled samples k

1.2 The Mitigation Dilemma and Analyst Load

When a novel attack is identified, the standard response is adversarial retraining. However, this introduces the "Analyst Bottleneck": during the early stages of a Zero-Day event, human analysts can typically only provide a handful of labeled samples ($k \leq 50$). Traditional models struggle in this ultra-low sample regime. To force a model trained on millions of benign flows to recognize a tiny minority of k attack flows, administrators often resort to aggressive class-weighting. We identify a systemic risk in this approach which we term *Boundary Collapse*, where the gradient updates required to accommodate the new attack manifold distort the decision boundaries for existing, stable traffic, thereby degrading the system's overall security posture.

ACM CSS '26, Hague, Netherlands

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

PART III – ZERO-DAY ADAPTIVE DEFENSE

RC1: *Provide* **Conformal Fast-Path Reliability** with LightGBM-based primary classifier using Mondrian Conformal Prediction

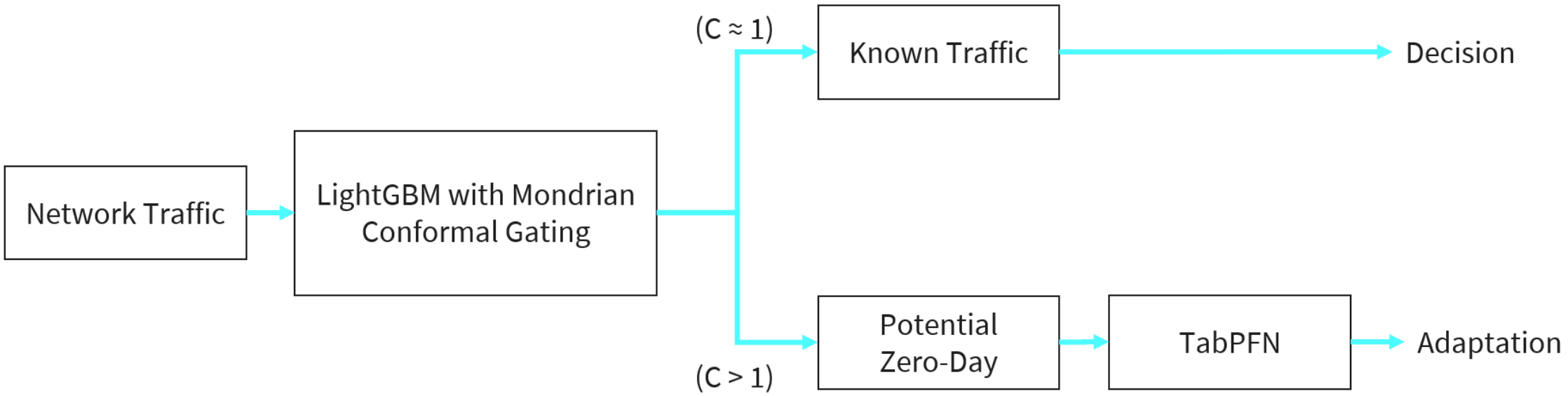
RC2: **Semantic Gating via Set Expansion** of prediction sets ($C \gg 1$), potential Zero-Day traffic

RC3: **Foundation-Model Driven Mitigation:** TabPFN-based high-accuracy rescue path for critical cases (e.g., 0-days)

RC4: **Ultra-Fast Adaptation:** *in-context learning* allows the system to adapt to novel attacks/0-days with small sample size

PART III – ZERO-DAY ADAPTIVE DEFENSE

$$\mathcal{R}(x) = \begin{cases} \text{Fast-Path} & \text{if } |\hat{C}(x,\alpha)| = 1 \\ \text{Rescue Path (TabPFN)} & \text{if } |\hat{C}(x,\alpha)| > 1 \\ \text{Reject} & \text{if } |\hat{C}(x,\alpha)| = 0 \end{cases}$$



PART III – ZERO-DAY ADAPTIVE DEFENSE

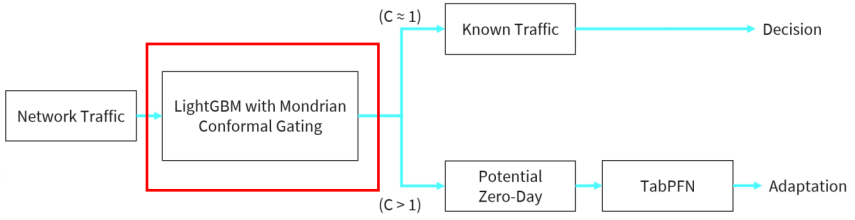
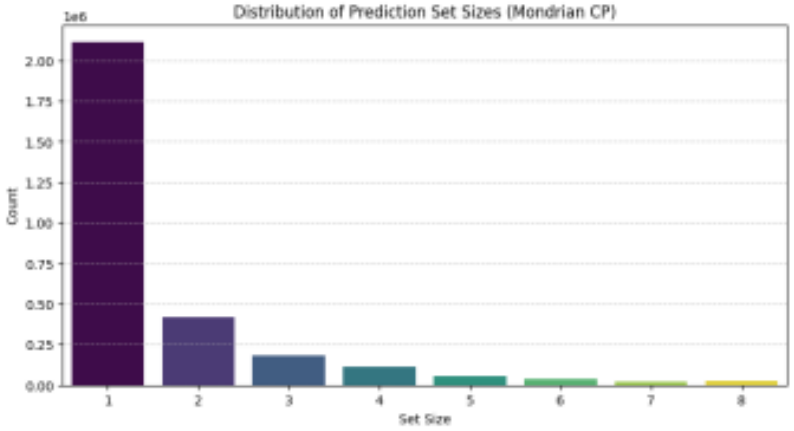
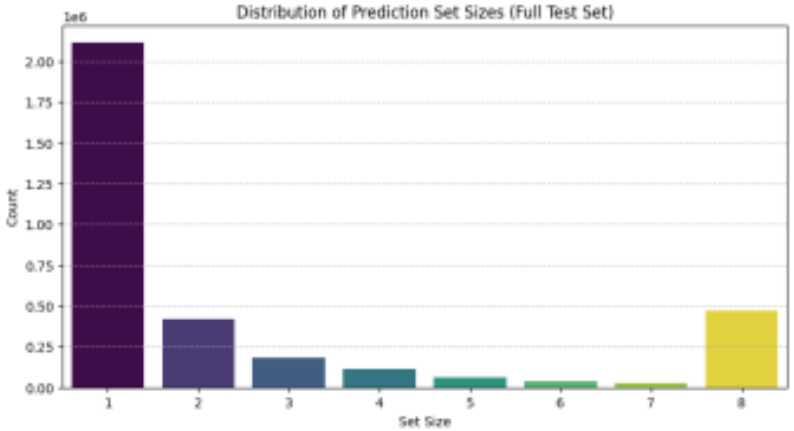


Table 1: Class Distribution in LycoS-Unicas-IDS2018

Class	Flow Count
Benign	10,000,000
DoS Hulk	1,802,966
DDoS HOIC	1,074,379
DDoS LOIC-HTTP	289,328
FTP-Patator	190,300
DoS Slowhttptest	105,550
Bot	96,154
SSH-Patator	92,648
DoS GoldenEye	26,861
DoS Slowloris	10,274
Excluded Classes:	
DDoS LOIC-UDP	2,382
Web Attack - Brute Force	260
Web Attack - XSS	116
Web Attack - Sql Injection	50



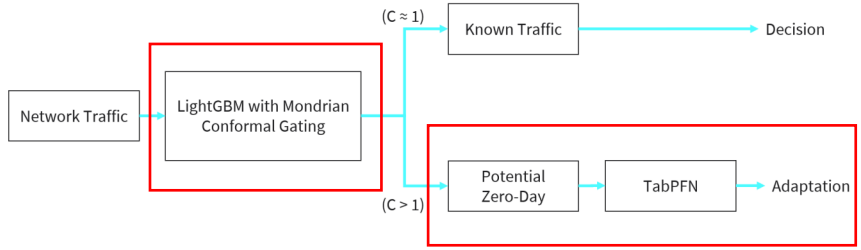
(a) All classes known at training / calibration.



(b) Using `DoS Hulk` as the zero-day attack

Figure 2: Overall distribution of prediction set sizes under Mondrian Conformal Prediction for the Fast-Path classifier.

PART III – ZERO-DAY ADAPTIVE DEFENSE



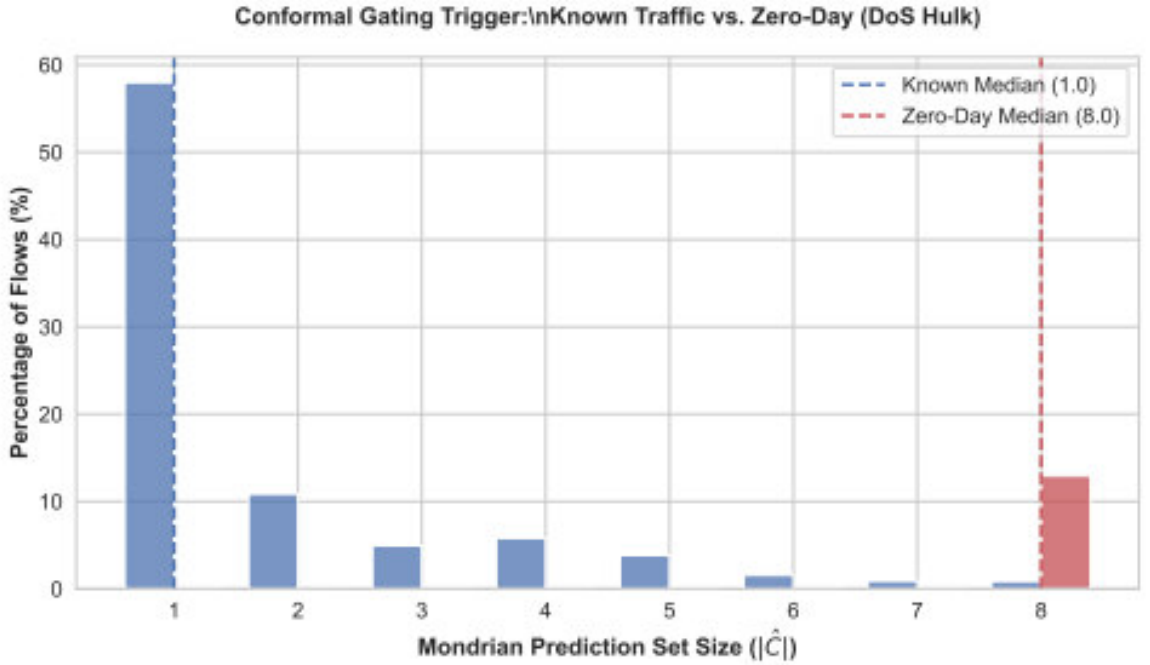
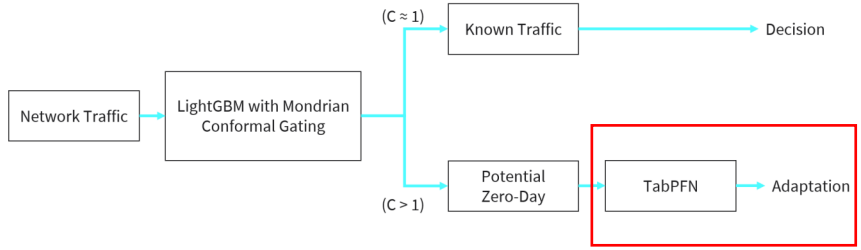


Figure 4: Visualizing the conformal gating mechanism in the case of DoS Hulk.

PART III – ZERO-DAY ADAPTIVE DEFENSE



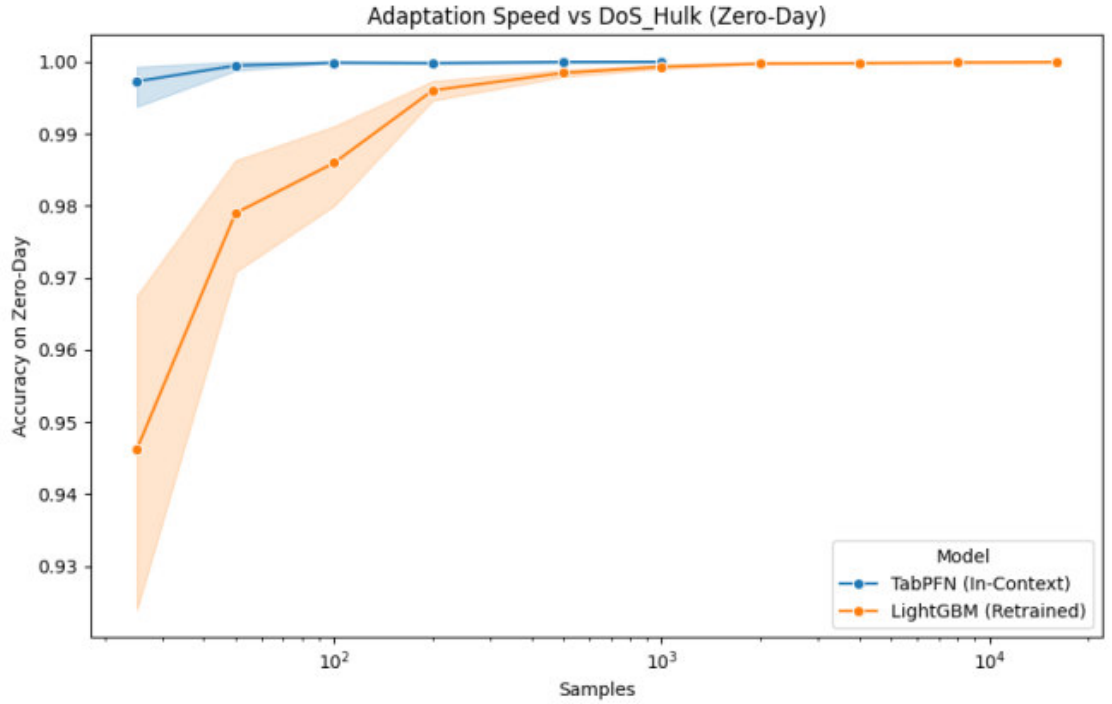


Figure 6: Learning performance of tabPFN vs. LightGBM in the case of DoS Hulk.

PART III – ZERO-DAY ADAPTIVE DEFENSE

Conclusion

- **Prediction set expansion** provides a **practical indicator** of distribution shift (i.a., 0-day)
- **Mondrian Conformal Gating** transforms **uncertainty** into a **routing** mechanism
- **TabPFN** enables rapid **few-shot adaptation** to novel attacks (sample size $k \leq 50$)
- **Adaptation** can be achieved **without** destructive **retraining**

"Trustworthy cyber defense begins when uncertainty becomes action."

AI-supported detection of cyber attacks (espec. zero-day exploits) on IT infrastructures
https://www.iu.de/en/research/projects/kidz/

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Project is based on the sub-problems (SP) related to zero-day exploits:

SP1: Delayed detection due to the time lag between the emergence of threats and the adaptation of AI/ML models.

SP2: Insufficient quantification of uncertainty and lack of transparency in AI/ML models.

Prof. Dr. Alexander Lawall
alexander.lawall@iu.org