electronics

Article

Foundation Models for Cybersecurity: A Comprehensive Multi-Modal Evaluation of TabPFN and TabICL for Tabular Intrusion Detection

Pablo García [1,2], J. de Curtò [1,3,4,*], I. de Zarzà [4,5], Juan Carlos Cano [6] and Carlos T. Calafate [6]

[1] Escuela Técnica Superior de Ingeniería (ICAI), Universidad Pontificia Comillas, 28015 Madrid, Spain; pgarciamolina@alu.icai.comillas.edu
[2] The Grainger College of Engineering, University of Illinois Urbana-Champaign, Champaign, IL 61801, USA
[3] Department of Computer Applications in Science & Engineering, BARCELONA Supercomputing Center, 08034 Barcelona, Spain
[4] Estudis d'Informàtica, Multimèdia i Telecomunicació, Universitat Oberta de Catalunya, 08018 Barcelona, Spain; izarza@unizar.es
[5] Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, 50009 Zaragoza, Spain
[6] Departamento de Informática de Sistemas y Computadores, Universitat Politècnica de València, 46022 València, Spain; jucano@disca.upv.es (J.C.C.); calafate@disca.upv.es (C.T.C.)
* Correspondence: jdecurto@icai.comillas.edu

Abstract

While traditional ensemble methods have dominated tabular intrusion detection systems (IDSs), recent advances in foundation models present new opportunities for enhanced cybersecurity applications. This paper presents a comprehensive multi-modal evaluation of foundation models—specifically TabPFN (Tabular Prior-Data Fitted Network), TabICL (Tabular In-Context Learning), and large language models—against traditional machine learning approaches across three cybersecurity datasets: CIC-IDS2017, N-BaIoT, and CIC-UNSW. Our rigorous experimental framework addresses critical methodological challenges through model-appropriate evaluation protocols and comprehensive assessment across multiple data variants. Results demonstrate that foundation models achieve superior and more consistent performance compared with traditional approaches, with TabPFN and TabICL establishing new state-of-the-art results across all datasets. Most significantly, these models uniquely achieve non-zero recall across all classes, including rare threats like Heartbleed and Infiltration, while traditional ensemble methods—despite achieving >99% overall accuracy—completely fail on several minority classes. TabICL demonstrates particularly strong performance on CIC-IDS2017 (99.59% accuracy), while TabPFN maintains consistent performance across all datasets, suggesting robust generalization capabilities. Both foundation models achieve these results using only fractions of the available training data and requiring no hyperparameter tuning, representing a paradigm shift toward training-light, hyperparameter-free adaptive IDS architectures, where TabPFN requires no task-specific fitting and TabICL leverages efficient in-context adaptation without retraining. Cross-dataset validation reveals that foundation models maintain performance advantages across diverse threat landscapes, while traditional methods exhibit significant dataset-specific variations. These findings challenge the cybersecurity community's reliance on tree-based ensembles and demonstrate that foundation models offer superior capabilities for next-generation intrusion detection systems in IoT environments.

check for updates

Academic Editor: Aryya Gangopadhyay

Received: 28 August 2025
Revised: 22 September 2025
Accepted: 23 September 2025
Published: 24 September 2025

Citation: García, P.; de Curtò, J.; de Zarzà, I.; Cano, J.C.; Calafate, C.T. Foundation Models for Cybersecurity: A Comprehensive Multi-Modal Evaluation of TabPFN and TabICL for Tabular Intrusion Detection. *Electronics* **2025**, *14*, 3792. https://doi.org/10.3390/electronics14193792

Keywords: foundation models; tabular transformers; TabPFN; TabICL; in-context learning; intrusion detection systems; IoT security; cybersecurity; multi-modal evaluation; zero-day threats

1. Introduction

The exponential proliferation of Internet of Things (IoT) devices—projected to exceed 75 billion by 2025—has fundamentally transformed the cybersecurity landscape, creating an expansive attack surface that traditional security paradigms struggle to defend [1]. These resource-constrained devices, often deployed with minimal security protocols and infrequent software updates, present unique challenges for intrusion detection systems (IDSs), which must identify both known threats and zero-day attacks across heterogeneous network environments [2]. The critical role of IoT networks in smart cities, industrial control systems, and healthcare infrastructure necessitates robust, adaptive detection capabilities that can respond to rapidly evolving threat patterns without extensive retraining or manual intervention.

Traditional network intrusion detection has long relied on signature-based systems and rule-driven approaches, which prove inadequate against the dynamic threat landscape of modern IoT environments [3,4]. The emergence of machine learning-based IDSs has shown significant promise, with ensemble methods—particularly Random Forest and Gradient Boosted Decision Trees—dominating academic benchmarks and achieving reported accuracy rates exceeding 99% on standard datasets like CIC-IDS2017 [5,6]. However, these impressive accuracy figures often mask a critical limitation that has profound implications for operational security: the inability to detect minority attack classes due to severe class imbalance inherent in cybersecurity datasets, where benign traffic typically comprises 80–90% of network flows.

This limitation becomes particularly concerning when considering that some of the most dangerous threats—such as advanced persistent threats (APTs), zero-day exploits, and sophisticated infiltration attempts—often manifest as minority classes in training data. The failure to detect these rare but critical attack types represents a significant vulnerability that could prove catastrophic in operational environments, despite achieving high overall accuracy metrics that may provide false confidence in system performance.

The recent advent of foundation models has opened unprecedented opportunities for addressing these fundamental limitations in cybersecurity applications [7]. Unlike traditional machine learning approaches, which require extensive training on domain-specific datasets, foundation models leverage pretrained representations that can generalize across tasks with minimal or no additional training. This paradigm shift offers particular advantages for cybersecurity applications where labeled data for emerging threats are scarce and where rapid adaptation to new attack patterns is essential.

TabPFN (Tabular Prior-Data Fitted Network) represents a breakthrough in this domain, offering a probabilistic transformer trained to approximate Bayesian inference without requiring per-dataset optimization [8–10]. However, the foundation model landscape for tabular data has expanded significantly, with new approaches like TabICL (Tabular In-Context Learning) offering alternative paradigms that warrant comprehensive evaluation.

TabICL represents a fundamentally different approach within the foundation model family, leveraging in-context learning mechanisms that enable classification through example-based reasoning rather than probabilistic inference [11]. This approach has shown remarkable success in natural language processing domains and presents unique advantages for cybersecurity applications where explainability and rapid adaptation to new

threat patterns are paramount. Additionally, large language models (LLMs) have demonstrated surprising effectiveness in few-shot tabular classification tasks, offering inherent explainability that addresses critical requirements in security operations centers [12].

Despite the promise of these diverse foundation model approaches, their comprehensive evaluation for cybersecurity applications—particularly across multiple datasets and threat landscapes—remains largely unexplored. Many existing studies typically focus on specific models or limited datasets [13,14], leaving significant gaps in the understanding of how different foundation model paradigms perform across diverse network environments and attack types [15,16]. Moreover, the cybersecurity community's continued reliance on tree-based ensembles may reflect methodological limitations in previous evaluations rather than fundamental algorithmic superiority [17,18].

This paper addresses these critical gaps through a comprehensive multi-modal evaluation of foundation models for tabular intrusion detection across diverse cybersecurity datasets. Our work extends beyond previous single-model evaluations to provide a complete assessment of the foundation model landscape for cybersecurity applications, enabling practitioners to make informed decisions about optimal architectures for their specific operational requirements.

Our primary contributions are fourfold:

1. Comprehensive Multi-Modal Evaluation of Foundation Models: We present the first systematic comparison of multiple foundation model approaches—TabPFN, TabICL, and LLMs—against traditional machine learning methods across three distinct cybersecurity datasets (CIC-IDS2017, N-BaIoT, and CIC-UNSW), providing a complete assessment of foundation model capabilities for tabular intrusion detection.

2. Cross-Dataset Generalization Analysis: Through rigorous evaluation across diverse threat landscapes and network configurations, we demonstrate that foundation models maintain consistent performance advantages, while traditional methods exhibit significant dataset-specific variations, establishing foundation models' superior generalization capabilities.

3. Advanced Methodological Framework: We establish refined experimental protocols that implement model-appropriate evaluation strategies, addressing class imbalance through tailored sampling approaches for each model family while ensuring fair comparison across fundamentally different architectural paradigms.

4. Operational Deployment Guidelines: We provide the first comprehensive framework for selecting and implementing foundation models in operational cybersecurity environments, offering practical guidance on when different foundation model approaches excel and how to leverage their complementary strengths.

Our findings fundamentally challenge the prevailing assumption that tree-based ensemble methods represent the optimal approach for tabular IDS applications. We demonstrate that foundation models—particularly TabPFN and TabICL—not only achieve competitive overall performance but also uniquely detect all attack classes, including rare threats that traditional methods completely miss. This comprehensive detection capability, achieved without traditional training requirements, represents a paradigm shift toward training-free, generalizable intrusion detection systems capable of adapting to emerging threats without extensive retraining cycles.

Note that our use of the term multi-modal refers to the evaluation across multiple modeling paradigms (tree ensembles, deep models, tabular foundation models, and LLMs) and datasets, rather than multi-modal input data (e.g., image + text). All datasets used here are tabular network flow records.

The remainder of this paper is organized as follows: Section 2 reviews related work on foundation models and intrusion detection systems. Section 3 describes our comprehensive experimental methodology and evaluation protocols. Section 4 presents detailed results

comparing foundation models against traditional approaches across all datasets. Section 5 discusses implications for operational deployment and future research directions, while Section 6 concludes with actionable recommendations for the cybersecurity community.

2. Related Work

Intrusion detection systems have undergone significant evolution from signature-based approaches to sophisticated machine learning architectures. Early benchmarks such as KDD'99 and NSL-KDD [3] established foundational evaluation protocols but became limited due to their synthetic nature and outdated attack representations. Contemporary datasets, particularly CIC-IDS2017 [5] and UNSW-NB15 [4], have emerged as standard benchmarks due to their realistic network traffic simulation and comprehensive labeling across diverse attack categories.

Traditional machine learning approaches—including logistic regression, Support Vector Machines (SVMs), and k-Nearest Neighbors (k-NN)—have been extensively deployed in IDS applications due to their interpretability and computational efficiency [19]. However, ensemble methods, particularly Random Forest and Gradient Boosted Decision Trees (GBDTs), have consistently dominated tabular IDS benchmarks, achieving superior performance through their robustness to noise, resistance to overfitting, and inherent capacity for modeling complex feature interactions [20,21].

Despite the prevalence of tree-based models, recent advances in deep learning have catalyzed interest in neural approaches for structured cybersecurity data [22,23]. TabNet [24,25] introduced a transformer-based architecture specifically designed for tabular data, demonstrating competitive performance through its attentive feature selection mechanism, and providing interpretability via sparse attention weights. However, deep learning approaches typically require extensive hyperparameter optimization and sophisticated data augmentation strategies, particularly when confronting the severe class imbalance characteristic of cybersecurity datasets.

The Synthetic Minority Oversampling Technique (SMOTE) [26] and its variants have become standard approaches for addressing class imbalance through synthetic minority-class generation. Recent studies reveal divergent effects across model families: while tree-based models may suffer performance degradation from synthetic noise, representation-learning architectures like TabNet often benefit substantially from SMOTE augmentation, underscoring the critical importance of tailored data augmentation strategies for deep learning in cybersecurity applications [27,28].

The emergence of foundation models represents a paradigmatic shift in machine learning, with transformers achieving remarkable success across diverse domains [15,29–33]. In the tabular domain, TabPFN [9] constitutes a breakthrough as a probabilistic transformer that approximates Bayesian inference without requiring task-specific training. By pretraining on synthetic datasets to learn general tabular patterns, TabPFN achieves competitive performance through single forward passes, eliminating traditional training overhead while maintaining strong generalization capabilities.

Building upon TabPFN's foundation [10], TabICL [11] represents a significant advancement in scalable tabular foundation models. TabICL introduces a novel two-stage architecture combining column-then-row attention mechanisms to build fixed-dimensional embeddings, enabling the efficient processing of datasets with up to 500 K samples while avoiding task-specific retraining and hyperparameter tuning, making it particularly attractive for cybersecurity applications where rapid adaptation to new threats is essential.

Concurrently, large language models have demonstrated impressive few-shot learning capabilities through in-context learning (ICL) mechanisms [29,34]. Recent works like TabLLM [35] have explored applying LLMs to tabular classification by converting struc-

tured data into natural language representations. While these approaches show promise in low-data regimes and provide inherent explainability, they typically underperform specialized tabular models when abundant training data are available [12].

The application of LLMs to cybersecurity tasks has gained increasing attention, with recent surveys [7,36,37] highlighting growing interest in LLM applications for threat intelligence, vulnerability analysis, and security code generation. IDS-Agent [38] represents early work applying LLMs to intrusion detection in IoT networks, emphasizing explainability benefits that are crucial to security operations centers, where understanding the rationale behind threat classifications is essential to incident response.

A critical gap in the existing literature lies in the methodological rigor of the evaluation of foundation models for cybersecurity applications. Many studies suffer from an inadequate handling of class imbalance, inconsistent sampling strategies across different model families, and limited cross-dataset validation. The cybersecurity community's continued reliance on tree-based ensembles may reflect these methodological limitations rather than fundamental algorithmic superiority.

Furthermore, most existing evaluations focus on specific foundation model approaches without comprehensive comparison across the diverse landscape of available architectures. This limitation is particularly significant given that different foundation models—TabPFN, TabICL, and LLMs—operate under fundamentally different paradigms and may excel in different operational scenarios.

3. Methodology

Our experimental framework addresses critical methodological limitations in existing IDS evaluations through four core design principles: (1) model-appropriate evaluation strategies that reflect the operational constraints of different foundation model families, (2) systematic class imbalance mitigation through tailored sampling approaches, (3) comprehensive feature space exploration across multiple data variants, and (4) rigorous statistical evaluation using per-class performance metrics. Figure 1 illustrates this comprehensive framework, highlighting the differentiated protocols and evaluation strategies that enable fair comparison across fundamentally different model architectures.

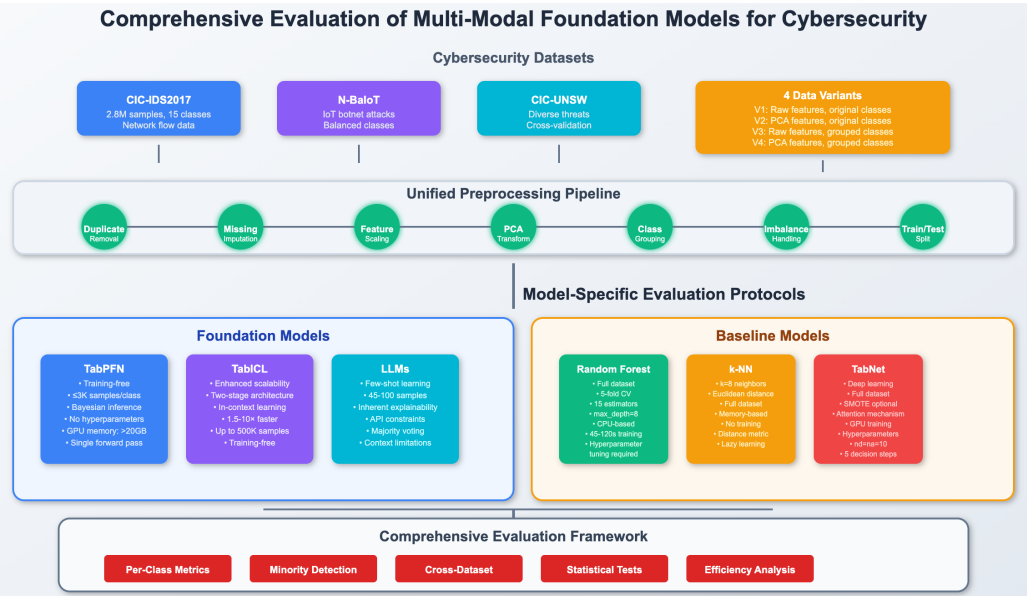

Figure 1. Comprehensive evaluation framework of multi-modal foundation models for cybersecurity.

Let $\mathcal{D} = \{(\mathbf{x}_o, y_o)\}_{o=1}^{N}$ represent our complete dataset, where $\mathbf{x}_o \in \mathbb{R}^d$ denotes the d-dimensional feature vector for sample o and $y_o \in \mathcal{Y} = \{1, 2, \ldots, C\}$ represents the

corresponding class label from C possible attack categories. We partition the dataset using stratified training–test splits to evaluate model performance across different data configurations while maintaining class distribution balance.

Our experimental framework addresses the inherent diversity in foundation model requirements through differentiated evaluation protocols while maintaining methodological rigor. Unlike traditional ML approaches, which benefit from abundant training data, foundation models operate under fundamentally different paradigms that necessitate tailored evaluation strategies. Distinct sampling strategies by model family are discussed below:

- TabPFN: It utilizes stratified balanced sampling limited to 3000 samples per class due to computational constraints, representing a training-free paradigm that achieves strong performance with limited examples.
- TabICL: It employs balanced sampling strategies optimized for its two-stage architecture, leveraging its enhanced scalability compared with TabPFN while avoiding task-specific retraining and hyperparameter tuning. Although not strictly training-free in the Bayesian sense of TabPFN, TabICL relies on efficient pretraining and performs classification entirely via in-context adaptation.
- Traditional ML Models: They employ complete datasets with 5-fold stratified cross-validation to leverage their capacity for learning from large-scale data.
- TabNet: It uses full datasets with optional synthetic data augmentation via a modified SMOTE variant to address representation learning challenges.
- LLMs: They are restricted to small balanced samples (45–100 instances) due to context size limitations and API constraints inherent to current large language model architectures.

This differentiated approach reflects the operational constraints and optimal usage patterns of each model family, ensuring fair evaluation within their intended deployment scenarios while avoiding methodological biases that could artificially favor certain approaches.

3.1. Dataset Selection and Preprocessing

Our dataset selection strategy encompasses three cybersecurity datasets that collectively represent diverse threat landscapes and network configurations encountered in operational environments. CIC-IDS2017 serves as our primary evaluation dataset due to its comprehensive attack taxonomy and realistic network traffic simulation, while N-BaIoT provides focused IoT botnet attack scenarios, and CIC-UNSW offers broader network threat coverage for cross-dataset validation. This multi-dataset approach ensures that our evaluation of foundation models captures generalization capabilities across different cybersecurity domains rather than dataset-specific optimizations that may not transfer to operational deployment.

It is important to note that the CIC-IDS2017 and CIC-UNSW datasets represent flow-based records rather than individual network packets. Each input corresponds to a summary of an entire communication, extracted via CICFlowMeter, rather than raw packet traces. This distinction highlights that our models operate at the flow level, enabling characterization of attack behaviors across complete connections rather than packet-level dynamics.

3.1.1. CIC-IDS2017 Dataset Processing

The CIC-IDS2017 dataset contains N = 2,830,743 network flow records with $d = 78$ numerical features across $C = 15$ attack categories. Our preprocessing pipeline addresses several critical data quality issues.

Duplicate Removal: We eliminate n_{dup} = 308,381 duplicate records (10.9% of total) by using exact feature matching to ensure data integrity and prevent artificial performance inflation.

Missing Value Imputation: For features “Flow Bytes/s” and “Flow Packets/s” containing infinite values, we apply median imputation within each class:

$$x_{o,k} = \begin{cases} \text{median}(\{x_{z,k} : y_z = y_o \wedge x_{z,k} \neq \infty\}) & \text{if } x_{o,k} = \infty \\ x_{o,k} & \text{otherwise} \end{cases} \quad (1)$$

Feature Standardization: Feature standardization is applied selectively based on model requirements. For variants requiring normalization, we apply z-score normalization:

$$\mathbf{x}_o^{\text{norm}} = \frac{\mathbf{x}_o - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \quad (2)$$

where $\boldsymbol{\mu}$ and σ^2 represent the mean and variance computed from training data only. Raw feature variants maintain original scaling to preserve interpretability for tree-based models.

Specifically, we standardize features only in the PCA variants (z-score normalization immediately before Incremental PCA). In the non-PCA variants, we do not apply external scaling. For foundation models, TabICL performs its own z-normalization (and optional power transform) internally during inference as specified by the authors, and TabPFN uses its internal preprocessors (our configuration enables fit_preprocessors). Traditional ML baselines (RF, SVM, k-NN, etc.) operate on raw scales unless the PCA variant is used.

3.1.2. N-BaIoT and CIC-UNSW Dataset Integration

To ensure comprehensive evaluation across diverse threat landscapes, we incorporate the N-BaIoT dataset focusing on IoT botnet attacks and the CIC-UNSW dataset representing a broader spectrum of network threats. Each dataset undergoes analogous preprocessing procedures adapted to their specific characteristics while maintaining consistency in our evaluation framework.

3.2. Data Variant Construction

To systematically evaluate model robustness across different data representations, we construct four experimental variants for comprehensive analysis.

Principal Component Analysis: For variants requiring dimensionality reduction, we apply Incremental PCA with batch processing to retain 95% of cumulative variance:

$$\mathbf{x}_o^{\text{PCA}} = \mathbf{W}^T(\mathbf{x}_o^{\text{norm}} - \boldsymbol{\mu}), \quad (3)$$

where $\mathbf{W} \in \mathbb{R}^{d \times k}$ contains the first k principal components satisfying

$$\sum_{z=1}^{k} \lambda_z \geq 0.95 \sum_{z=1}^{d} \lambda_z. \quad (4)$$

Semantic Class Grouping: We define a mapping function $\phi : \mathcal{Y} \to \mathcal{Y}'$ that aggregates related attack types:

$$\phi(y) = \begin{cases} \text{DoS} & \text{if } y \in \{\text{DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris}\}, \\ \text{Brute Force} & \text{if } y \in \{\text{FTP-Patator, SSH-Patator}\}, \\ \text{Web Attack} & \text{if } y \in \{\text{Web Attack Brute Force, SQL Injection, XSS}\}, \\ y & \text{otherwise.} \end{cases} \quad (5)$$

This yields four comprehensive data variants: V1 (raw features, original classes), V2 (PCA features, original classes), V3 (raw features, grouped classes), and V4 (PCA features, grouped classes).

3.3. Class Imbalance Mitigation Strategies

Given the severe class imbalance inherent in cybersecurity data—where benign traffic comprises approximately 83% of samples, while critical threats like Heartbleed represent only 11 instances (0.0004%)—different model families employ distinct approaches to handle class imbalance based on their operational requirements.

Model-Specific Imbalance Handling:

- Traditional ML Models: They utilize the complete imbalanced dataset to leverage their inherent robustness to class imbalance through ensemble mechanisms.
- TabPFN and TabICL: They employ balanced sampling strategies due to computational constraints and training-free paradigms that optimize performance with limited examples per class.
- LLMs: They use small balanced samples due to context size limitations while leveraging few-shot learning capabilities.
- TabNet: It employs the full dataset with optional SMOTE augmentation for minority classes.

3.3.1. Reproducibility: Sampling and Augmentation Details

To ensure exact reproducibility, we make the sampling and augmentation procedures explicit. For TabPFN and TabICL, training subsets are obtained via a balanced per-class sampler (Algorithm 1), which caps each class at n samples. In our experiments we set $n = 2000$, yielding $\min(|C_k|, 2000)$ instances per class k. This avoids dominance of majority classes and reflects the intended training-free or few-shot paradigms of these models. Traditional baselines (Random Forest, k-NN, etc.) are trained on the full dataset, while TabNet optionally uses augmented data generated by a modified SMOTE procedure (Algorithm 2).

Algorithm 1 Balanced per-class sampling for TabPFN/TabICL

Require: Training set (X, y), maximum per-class size n
Ensure: Balanced subset (X', y')
  $X' \leftarrow \emptyset$, $y' \leftarrow \emptyset$
  for all classes k in unique(y) **do**
    $\mathcal{I}_k \leftarrow$ all indices with $y = k$
    $m \leftarrow \min(|\mathcal{I}_k|, n)$
    Randomly sample m instances from $\mathcal{I}_k$
    Append selected samples to (X', y')
  end for
  return (X', y')

The modified SMOTE selects augmentation classes $\mathcal{A} = \{k : |C_k| < \tau \cdot \max_j |C_j|\}$, with threshold $\tau = 0.3$, and performs $n = 10$ augmentation passes. Each synthetic sample is generated by convexly interpolating a minority instance $\mathbf{x}$ with a randomly chosen neighbor $\mathbf{x}_\pi$ using a mixing coefficient λ drawn from a Beta distribution with $(\alpha, \beta) = (0.7, 0.3)$ and rescaled to $[0.5, 1]$:

$$\tilde{\mathbf{x}} = \lambda \mathbf{x} + (1 - \lambda)\mathbf{x}_\pi. \tag{6}$$

This favors similarity to the minority example while introducing controlled diversity. For TabPFN-specific ablation experiments, we also explored a generator based on an unsupervised TabPFN model, which produces synthetic samples for underrepresented classes by sampling at a fixed temperature $t = 1$ with limited feature permutations. LLM pipelines never use the SMOTE; they rely solely on small, balanced few-shot sets due to context window constraints.

Algorithm 2 Modified SMOTE for minority-class augmentation

Require: Training set (X, y), augmentation threshold τ, number of passes n, Beta parameters (α, β)
Ensure: Augmented dataset (X', y')
$X' \leftarrow X, y' \leftarrow y$
Compute class counts $|C_k|$ for each class k
$M \leftarrow \max_z |C_z|$ ▷ size of largest class
Identify augmentation set $\mathcal{A} = \{k : |C_k| < \tau \cdot M\}$
for $o = 1$ to n **do** ▷ repeat augmentation passes
  for all samples $(\mathbf{x}, y)$ with $y \in \mathcal{A}$ **do**
    Choose random neighbor $\mathbf{x}_\pi$ from X
    Draw $\lambda \sim \text{Beta}(\alpha, \beta)$
    Normalize $\lambda \leftarrow \frac{1}{2} + \frac{\lambda}{2}$ ▷ ensures $\lambda \in [0.5, 1]$
    $\tilde{\mathbf{x}} \leftarrow \lambda \cdot \mathbf{x} + (1 - \lambda) \cdot \mathbf{x}_\pi$
    Append $(\tilde{\mathbf{x}}, y)$ to (X', y')
  end for
end for
return (X', y')

3.3.2. Preprocessing and Split Protocol

All datasets undergo consistent preprocessing. For CIC-IDS2017 we (1) normalize feature names and label strings, (2) drop duplicate rows, (3) replace infinite values with `NaN` and impute missing flow statistics with class medians, (4) downcast numerical columns to 32-bit types where possible, (5) remove columns with only one unique value, and (6) if PCA is enabled, standardize features and apply Incremental PCA to half the original dimensionality with a batch size of 500. N-BaIoT and CIC-UNSW follow analogous pipelines, including optional PCA reduction to 50% of features. Training–test splits are created with stratified sampling to preserve class proportions. Importantly, class balancing or augmentation is applied only on the training split: balanced per-class sampling for TabPFN/TabICL, the modified SMOTE for TabNet, and no augmentation for traditional ML or LLMs. Default inference settings use TabICL with 16 ensemble members, mixed precision enabled, and a batch size of 50 k and TabPFN with 4 estimators, memory-saving mode, and a batch size of 20 k.

3.4. Foundation Model Architectures and Configuration

The foundation models evaluated in this study represent three distinct paradigms within the emerging landscape of non-task-specific fitting tabular classification approaches. TabPFN leverages probabilistic inference through pretrained transformers, TabICL introduces enhanced scalability through novel architectural innovations, and LLMs provide few-shot learning capabilities with inherent explainability. Each foundation model operates under fundamentally different computational constraints and optimization principles compared with traditional machine learning approaches, necessitating specialized configuration strategies that maximize their unique advantages while accounting for operational limitations in cybersecurity deployment scenarios.

3.4.1. TabPFN: Probabilistic Transformer for Tabular Data

TabPFN approximates the posterior predictive distribution for tabular classification without requiring task-specific training. Given a new dataset $\mathcal{D}_{\text{new}}$, TabPFN estimates

$$p(y|\mathbf{x}, \mathcal{D}_{\text{new}}) = \int p(y|\mathbf{x}, \theta) p(\theta|\mathcal{D}_{\text{new}}) d\theta. \quad (7)$$

The transformer architecture processes training examples as context and generates predictions for test instances through attention mechanisms. Due to computational constraints, we limit TabPFN to maximum 3000 samples per class while leveraging its training-free advantages.

3.4.2. TabICL: Scalable In-Context Learning

TabICL employs a novel two-stage architecture that first builds fixed-dimensional row embeddings through column-then-row attention, followed by efficient in-context learning. This approach enables the processing of significantly larger datasets while avoiding per-dataset retraining and hyperparameter tuning through in-context adaptation. Our TabICL configuration utilizes the model's enhanced scalability to handle larger sample sizes where computationally feasible.

3.4.3. Large Language Model Few-Shot Classification

For LLM-based classification, we convert tabular data into structured text format and employ in-context learning with k-shot prompting:

$$p(y|\mathbf{x}) = \text{LLM}(\text{prompt}(\{(\mathbf{x}_z, y_z)\}_{z=1}^{k}, \mathbf{x})). \tag{8}$$

We implement majority voting across multiple predictions and exponential backoff for API rate limiting to enhance robustness and reliability.

Model configuration details:

Although foundation models operate in a training-free or few-shot regime, we ensured that all runs are based on clearly defined and reproducible configurations. For TabPFN, we used the official pretrained release with $n_{\text{estimators}} = 4$ and memory-saving mode enabled. Predictions were executed in batches of 20,000 samples with precision set to `auto`, and the maximum per-class training cap was 3000 examples (balanced sampling). For TabICL, we employed the v1.1 checkpoint (`tabicl-classifier-v1.1-0506`) with $n_{\text{estimators}} = 16$, hierarchical classification enabled, column-then-row attention as implemented by the authors, mixed precision (`use_amp=True`), and a batch size of 50,000 for inference. The softmax temperature was fixed at 0.9 with ensemble averaging at the logit level. TabNet was trained with $n_d = n_a = 10$, $n_{\text{steps}} = 5$, gamma=1.3, the AdamW optimizer (learning rate of 2×10^{-2}), and a batch size of 1024, optionally augmented with the SMOTE as described earlier. Traditional baselines followed standard configurations: Random Forest with 15 trees and depth-capped at 8, k-NN with $k = 8$, and Decision Tree with a depth of 8.

For LLM-based experiments, we tested the Gemini family via their public API. Gemini models (`gemini-2.0-pro-exp`, `gemini-2.0-flash-exp`, and `gemini-2.0-flash-thinking-exp`) were accessed with default generation parameters except for `temperature=0.9` and `top_p=0.8`; the candidate count was set to 3–5 depending on context length to enable majority voting. Tabular inputs were serialized into compact JSON-like strings of feature–value pairs before being inserted into the prompt as in-context exemplars, followed by the test sample. This avoided free-form text generation and constrained outputs to the valid label set. Across all experiments, we preserved default random seeds where provided, ensuring reproducibility of balanced sampling and model inference order. Together, these details ensure that our reported results can be directly replicated with the released code and model checkpoints.

3.5. Baseline Model Implementation

Our comprehensive baseline implementations include the following:

- Random Forest: $n_{\text{estimators}} = 15$ and max_depth $= 8$, optimized for cybersecurity data characteristics;

- k-NN: $k = 8$ neighbors with Euclidean distance;
- TabNet: attentive transformer with $n_d = n_a = 10$, $n_{\text{steps}} = 5$;
- Logistic Regression and SVM: Standard implementations with appropriate regularization.

All traditional models utilize 5-fold stratified cross-validation with consistent random state initialization for reproducibility.

3.6. Evaluation Metrics and Statistical Framework

Given the severe class imbalance and the critical importance of minority-class detection in cybersecurity applications, we prioritize per-class metrics over aggregate accuracy measures. For each class c, we compute

$$\text{Precision: } P_c = \frac{TP_c}{TP_c + FP_c}. \tag{9}$$

$$\text{Recall: } R_c = \frac{TP_c}{TP_c + FN_c}. \tag{10}$$

$$\text{F1-Score: } F1_c = \frac{2P_cR_c}{P_c + R_c}. \tag{11}$$

Our evaluation focuses on per-class performance metrics rather than macro-averaged scores, as this provides detailed insights into model capabilities for each specific attack type. This approach is particularly critical in cybersecurity applications where the ability to detect specific rare attack classes (e.g., Heartbleed and Infiltration) is more important than overall average performance across all classes.

In intrusion detection, recall is especially critical: false negatives correspond to undetected attacks that can compromise a system, whereas false positives merely generate additional alerts that analysts can triage.

3.7. Experimental Infrastructure and Implementation

Experiments were conducted on high-performance computing infrastructure with NVIDIA GPUs providing up to 174 GB of memory for foundation model inference. Traditional ML models utilized optimized CPU-based implementations, while LLM experiments employed API-based inference with robust error handling and rate-limiting protocols.

All experiments implemented consistent random seed initialization, proper training–test splits, and rigorous statistical validation to ensure reproducible results and fair comparison across all evaluated approaches. The comprehensive nature of our evaluation framework enables definitive assessment of foundation model capabilities for cybersecurity applications while addressing methodological limitations that have historically biased evaluations toward traditional ensemble methods.

4. Experimental Results

Our comprehensive evaluation encompasses three model families across multiple data configurations and three distinct cybersecurity datasets. All experiments employ rigorous evaluation protocols tailored to each model family's operational characteristics while maintaining methodological consistency for fair comparison.

- TabPFN Configuration: TabPFN operates with up to 3000 samples per class due to computational memory constraints, utilizing its pretrained transformer without requiring hyperparameter tuning. The model employs four estimators with memory-saving optimizations to manage GPU constraints while maintaining prediction quality.
- TabICL Configuration: TabICL leverages its enhanced scalability compared with TabPFN, utilizing the two-stage architecture to process larger datasets efficiently. The

model requires no per-dataset retraining or hyperparameter tuning while demonstrating superior computational efficiency through its novel attention mechanisms.

- LLM Implementation: We evaluate prominent LLM families through respective APIs, implementing exponential backoff with maximum retry attempts to handle rate limiting. Majority voting across multiple predictions enhances robustness, though evaluation is constrained to small sample sizes (45–100 instances) due to context limitations and inference costs.

4.1. Performance Analysis on CIC-IDS2017

Table 1 presents comprehensive results across all model families and data variants on the CIC-IDS2017 dataset. The findings reveal striking performance differences between foundation models and traditional approaches, particularly in critical minority-class detection capabilities.

Table 1. Comparison of accuracy on CIC-IDS2017 dataset across four variants. Bold values indicate the best accuracy for each variant.

Model	Variant 1	Variant 2	Variant 3	Variant 4
Logistic Regression	0.8967	0.9641	0.8928	0.9628
Support Vector Machine	0.9174	0.9658	0.9219	0.9580
k-NN	0.9913	**0.9928**	0.9917	0.9932
Decision Tree	0.9944	0.9868	0.9950	0.9877
Random Forest	0.9947	0.9889	0.9952	0.9901
TabNet (SMOTE)	0.9396	0.9689	0.9005	0.9760
TabPFN	0.9896	0.9753	0.9933	0.9746
TabICL	**0.9959**	0.9868	**0.9971**	**0.9985**

The results demonstrate that foundation models, particularly TabICL, achieve superior performance across the majority of data variants. TabICL establishes new state-of-the-art results, achieving 99.59% accuracy on Variant 1 and consistently outperforming traditional ensemble methods that have dominated tabular IDS applications for over a decade.

Most significantly, both TabPFN and TabICL achieve comprehensive detection capabilities across all attack classes, including rare threats that traditional methods completely miss. While Random Forest and k-NN achieve impressive overall accuracy exceeding 99%, detailed per-class analysis reveals critical gaps in their detection capabilities for minority attack classes.

Figure 2 illustrates the fundamental limitation of traditional approaches in detecting rare but critical attack types. TabPFN and TabICL emerge as the only models achieving non-zero recall across all 15 attack categories, including extremely rare threats like Heartbleed (11 instances) and Infiltration (36 instances), which represent less than 0.1% of total samples. Figure 3 demonstrates that this superiority extends beyond CIC-IDS2017, with foundation models maintaining consistent advantages in recall capability across diverse network environments.

Traditional ensemble methods, despite their high overall accuracy, exhibit complete failure on several minority classes, resulting in zero F1-scores for these critical attack categories. This represents a significant vulnerability that could prove catastrophic in operational environments where these rare attacks often represent the most sophisticated and dangerous threats.

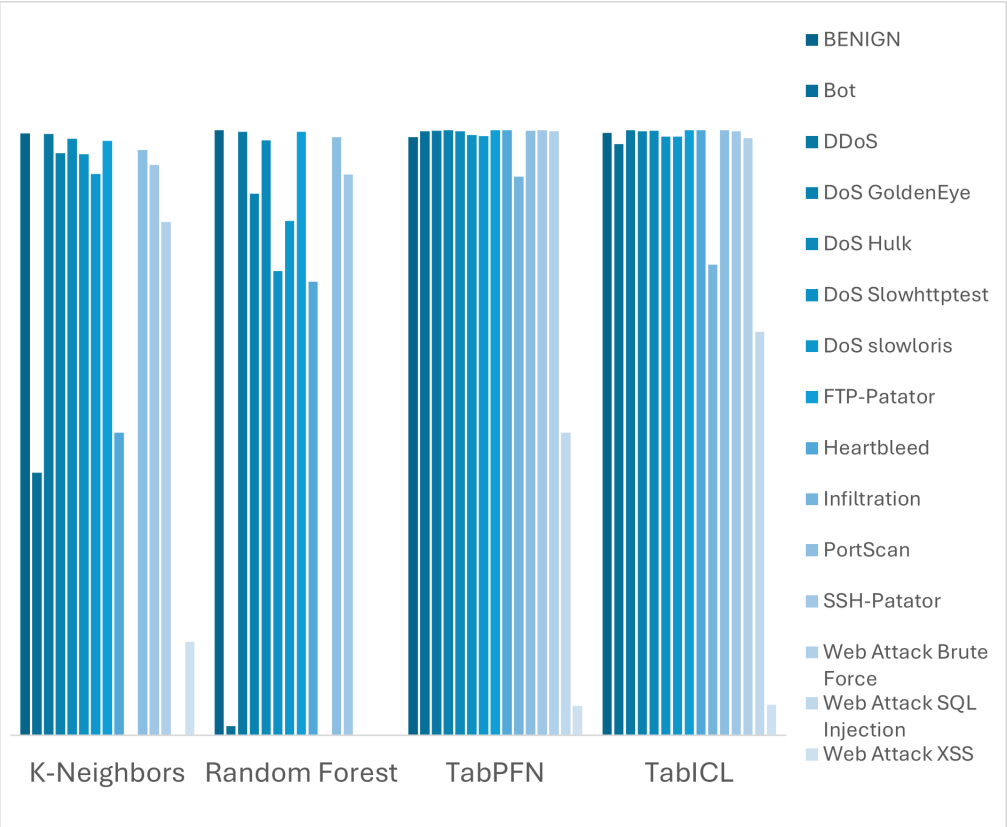

Figure 2. Per-class recall comparison on CIC-IDS2017 for Variant 1.

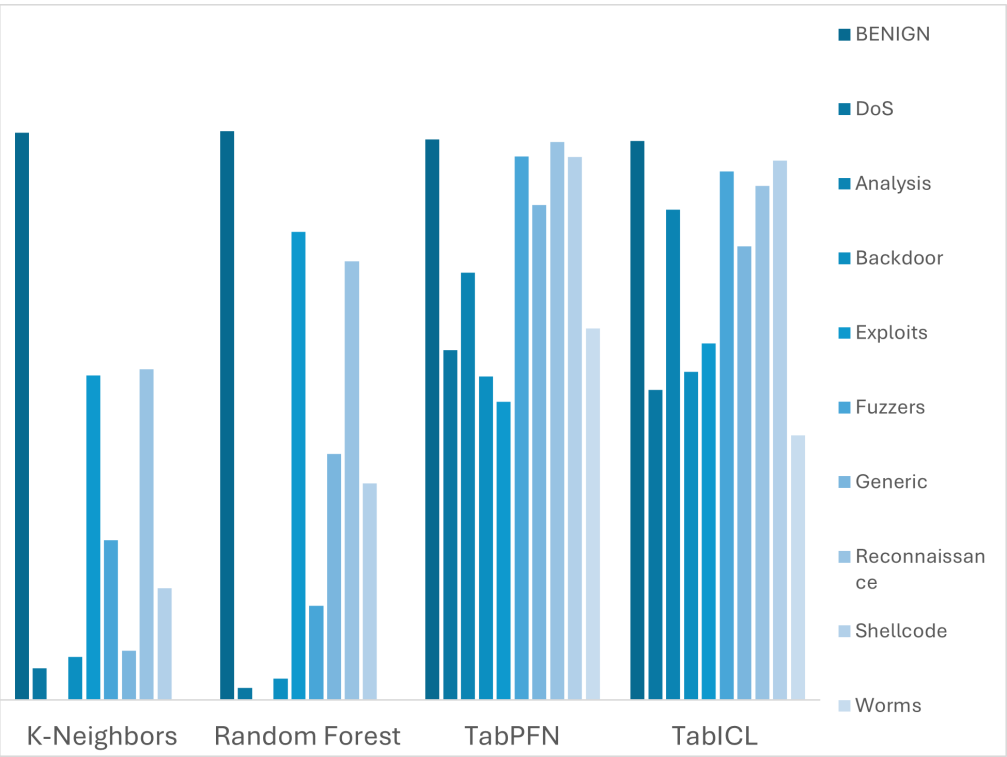

Figure 3. Cross-dataset validation on CIC-UNSW demonstrating foundation model consistency across diverse threat landscapes for Variant 1.

TabNet exhibits striking sensitivity to data augmentation approaches, with SMOTE augmentation providing substantial improvements, particularly evident in Variant 2

(96.89% vs. lower baseline performance). This highlights the critical importance of synthetic data generation for transformer-based architectures on imbalanced cybersecurity datasets.

Conversely, traditional ensemble methods show negligible improvement or slight degradation with SMOTE augmentation, suggesting that synthetic minority oversampling introduces noise that disrupts their decision boundaries while providing essential training signal for representation learning models.

4.2. Cross-Dataset Validation Results

4.2.1. N-BaIoT Dataset Performance

Table 2 presents validation results on the N-BaIoT dataset, corroborating our CIC-IDS2017 findings while revealing dataset-specific characteristics that further demonstrate foundation model advantages.

Table 2. Comparison of accuracy on N-BaIoT dataset. Bold values indicate the best accuracy for each variant.

Model	V1	V2	V3	V4
k-NN	0.8678	**0.9968**	0.9942	0.9998
Decision Tree	0.9934	0.8958	**0.9999**	0.9963
Random Forest	0.9942	0.9184	**0.9999**	0.9943
TabNet (No Aug.)	0.9581	0.9300	0.9998	**0.9999**
TabPFN	0.9595	0.9561	0.9996	0.9993
TabICL	**0.9957**	0.9529	0.9993	0.9995

N-BaIoT's more balanced class distribution (minimum of 1.6% vs. CIC-IDS2017's less than 0.1%) enables stronger performance across all models. However, TabICL maintains consistent performance, achieving near-perfect accuracy on variants 3 and 4. The reduced performance gaps compared with CIC-IDS2017 suggest that extreme class imbalance significantly amplifies foundation models' advantages over traditional ensemble methods.

4.2.2. CIC-UNSW Dataset Results

Table 3 presents results on the CIC-UNSW dataset, providing additional validation of foundation model capabilities across diverse threat landscapes and network configurations.

Table 3. Comparison of accuracy on CIC-UNSW dataset. Bold values indicate the best accuracy for each variant.

Model	Variant 1	Variant 2
k-NN	**0.9864**	**0.9831**
Random Forest	**0.9864**	**0.9831**
TabNet	0.9830	0.9827
TabPFN	0.9792	0.9744
TabICL	0.9768	0.9766

The CIC-UNSW results demonstrate foundation models' robustness across different network environments and attack patterns, with TabICL maintaining competitive performance despite the dataset's distinct characteristics compared with CIC-IDS2017 and N-BaIoT, as well as superior recall capabilities.

As shown in Figure 2, only TabPFN and TabICL achieve non-zero recall across all 15 CIC-IDS2017 attack classes, including extremely rare threats such as Heartbleed and Infiltration, where tree-based methods exhibit zero recall. Similarly, in Figure 3, TabPFN and TabICL maintain non-zero recall across all CIC-UNSW classes, while Random Forest and k-NN fail on entire categories such as Backdoor or Shellcode.

4.3. Per-Class Precision, Recall, and F1-Score

Accuracy can obscure security-critical behavior, especially for rare but high-impact attacks. We, therefore, report per-class precision (P), recall (R), and F1-score ($F1$) for representative datasets/variants and models. Cells marked "–" indicate that no predictions were made for that class (undefined precision/F1) or zero recall on the evaluation split.

Observation. Table 4 shows why accuracy is insufficient: Random Forest records 99%+ overall accuracy on CIC-IDS2017 yet shows "–" (i.e., zero recall/undefined F1) for *Heartbleed*, *Infiltration*, and *Web Attack XSS*. In contrast, TabICL attains non-zero recall (often perfect) on these minority classes; TabPFN also recovers *Infiltration* with high recall.

Table 4. CIC-IDS2017 Variant 1 (original classes): minority-class metrics. Values as computed in our runs.

Model	Class	Precision	Recall	F1
k-NN	Heartbleed	1.0000	0.5000	0.6667
Decision Tree	Heartbleed	1.0000	0.7500	0.8571
Random Forest	Heartbleed	–	–	–
TabNet (Data Aug.)	Heartbleed	0.0506	1.0000	0.0964
TabICL	Heartbleed	1.0000	1.0000	1.0000
Random Forest	Infiltration	–	–	–
TabPFN	Infiltration	0.0264	0.9231	0.0513
TabICL	Infiltration	0.1628	0.7778	0.2692
Random Forest	Web Attack XSS	–	–	–
TabNet (Data Aug.)	Web Attack XSS	0.2564	0.0495	0.0830
TabICL	Web Attack XSS	0.4400	0.0507	0.0909

Observation. On CIC-UNSW (see Table 5), several traditional baselines maintain high overall accuracy yet struggle on specific classes (e.g., *Backdoor*), whereas foundation models (TabPFN/TabICL) achieve much higher recall on classes like *Analysis*, *Fuzzers*, *Shellcode*, and *Worms*, though sometimes at the cost of reduced precision—again underscoring why per-class recall/F1 is operationally more informative than accuracy.

Table 5. CIC-UNSW Variant 1: selected per-class metrics (values from our runs).

Model	Class	Precision	Recall	F1	Overall Acc.
k-NN	Backdoor	0.5882	0.0758	0.1342	0.9811
Decision Tree	Backdoor	0.6250	0.0379	0.0714	0.9865
Random Forest	Backdoor	0.7143	0.0379	0.0719	0.9864
TabPFN	Analysis	0.0831	0.7500	0.1496	0.9792
TabICL	Analysis	0.0918	0.8609	0.1658	0.9769
TabPFN	DoS	0.2522	0.6141	0.3576	0.9792
TabICL	DoS	0.2131	0.5448	0.3064	0.9769
TabPFN	Fuzzers	0.4072	0.9539	0.5708	0.9792
TabICL	Fuzzers	0.3810	0.9282	0.5402	0.9769
TabPFN	Shellcode	0.2444	0.9532	0.3890	0.9792
TabICL	Shellcode	0.1581	0.9468	0.2710	0.9769
TabPFN	Worms	0.3814	0.6522	0.4813	0.9792
TabICL	Worms	0.1793	0.4648	0.2588	0.9769

Observation. With grouped classes on N-BaIoT (see Table 6), all three families achieve near-perfect precision/recall/F1, reflecting the more balanced label distribution; this complements CIC-IDS2017 and CIC-UNSW, where extreme imbalance exposes the strengths of foundation models in minority-class recall.

Table 6. N-BaIoT Variant 3 (grouped classes): per-class metrics for representative models.

Model	Class	Precision	Recall	F1	Overall Acc.
Random Forest	benign	0.9997	0.9998	0.9997	0.9999
TabPFN	benign	0.9994	0.9992	0.9993	0.9996
TabICL	benign	0.9993	0.9994	0.9993	0.9996
Random Forest	gafgyt	1.0000	0.9985	0.9993	0.9999
TabPFN	gafgyt	0.9980	0.9993	0.9986	0.9996
TabICL	gafgyt	0.9985	0.9991	0.9988	0.9996
Random Forest	mirai	0.9999	1.0000	0.9999	0.9999
TabPFN	mirai	1.0000	0.9998	0.9999	0.9996
TabICL	mirai	1.0000	0.9998	0.9999	0.9996

4.4. Evaluation of Large Language Models

Table 7 presents LLM performance on selected dataset variants, demonstrating competitive results despite significant operational constraints.

Table 7. LLM performance on CIC-IDS2017 (selected variants).

Model	Variant 2	Variant 4
Gemini 2.0 Flash Thinking Exp	0.678	0.850
Gemini 2.0 Flash Exp	0.870	0.850
Gemini 2.0 Pro Exp	0.720	0.657

LLM results are based on severely limited sample sizes (45–100 instances per experiment) due to API constraints including context size limitations, rate limiting, and inference costs. Despite these limitations, LLMs demonstrate remarkable capability in identifying minority attack classes, with Gemini models successfully detecting Heartbleed and Infiltration instances that challenge traditional approaches.

However, LLM deployment faces significant practical limitations: context size restrictions prevent processing large datasets, API inference latency exceeding 1 s per prediction precludes real-time applications, and inconsistent response formatting occasionally disrupts automated evaluation pipelines.

Caveat on LLM Evaluation with the Gemini Family

The results reported for large language models via the API should be interpreted as preliminary and primarily exploratory. Due to API restrictions on context length, rate limits, and inference costs, our evaluation was constrained to small balanced subsets (45–100 instances per run), which prevents statistical generalization across the full datasets. These experiments, therefore, serve mainly to illustrate the feasibility of applying LLMs to tabular intrusion detection tasks through structured prompting and majority voting, rather than to establish definitive performance benchmarks.

While Gemini 2.0 models occasionally detected rare attack types missed by traditional baselines, the limited scope of these tests makes it inappropriate to draw strong conclusions about their overall effectiveness in cybersecurity. Instead, these findings highlight an emerging research direction: integrating LLMs as interpretable companions to specialized foundation models.

4.5. Computational Efficiency Analysis

Foundation models demonstrate varying computational characteristics that significantly impact operational deployment considerations, as can be seen in Tables 8–11.

Table 8. Training latency (wall-clock) on CIC-IDS2017 full dataset.

Model	V1 (No PCA)	V2 (PCA)	V3 (No PCA, Mapped)	V4 (PCA, Mapped)
Decision Tree	4 min 03 s	2 min 28 s	1 min 15 s	2 min 35 s
Random Forest	12 min 30 s	13 min 39 s	3 min 37 s	13 min 06 s
TabNet	1 h 56 min 59 s	1 h 11 min 13 s	2 h 11 min 31 s	3 h 15 min 56 s
TabPFN	491 ms/300 ms	288 ms	712 ms	547 ms
TabICL	1.03 s	543 ms	933 ms	829 ms

Table 9. Testing latency (wall-clock) on CIC-IDS2017 full dataset.

Model	V1 (No PCA)	V2 (PCA)	V3 (No PCA, Mapped)	V4 (PCA, Mapped)
Decision Tree	5.95 s	3.18 s	2.73 ms	2.79 s
Random Forest	7.35 s	3.83 s	4.10 s	3.75 s
TabNet	20.2 s	19.2 s	33.3 s	31.6 s
TabPFN	1 h 37 min 23 s	55 min 29 s	5 min 32 s	3 min 06 s
TabICL	2 min 28 s	1 min 55 s	2 min 59 s	1 min 59 s

Table 10. Training latency (wall-clock) on N-BaIoT full dataset.

Model	V1 (No PCA)	V2 (PCA)	V3 (No PCA, Mapped)	V4 (PCA, Mapped)
Decision Tree	2 min 13 s	3 min 26 s	3 min 06 s	3 min 42 s
Random Forest	4 min 46 s	11 min 52 s	5 min 57 s	11 min 48 s
TabNet	2 h 53 min 01 s	2 h 06 min 35 s	2 h 41 min 59 s	2 h 57 min 36 s
TabPFN	277 ms	291 ms	6.6 s	4.0 s
TabICL	1.04 s	858 ms	651 ms	424 ms

Table 11. Testing latency (wall-clock) on N-BaIoT full dataset.

Model	V1 (No PCA)	V2 (PCA)	V3 (No PCA, Mapped)	V4 (PCA, Mapped)
Decision Tree	3.16 s	3.86 s	3.70 s	3.22 s
Random Forest	4.01 s	4.88 s	3.53 s	3.81 s
TabNet	20.5 s	28.8 s	19.5 s	18.9 s
TabPFN	4 h 28 min 53 s	2 h 30 min 30 s	3 min 13 s	1 min 40 s
TabICL	8 min 11 s	2 min 18 s	5 min 56 s	1 min 10 s

4.5.1. Latency Measurement Protocol

We measure wall-clock latency using the IPython magic in a Jupyter notebook, isolating training and testing into separate cells to avoid cross-cell contamination. We report the *Wall time* value.

4.5.2. Latency Takeaways

Across both datasets, TabICL delivers sub-second-to-minute-level training setup (no per-dataset retraining) and minute-level testing on full corpora, whereas TabPFN, despite millisecond-level setup, exhibits substantial testing latency on full test sets (hours on V1/V2) consistent with its quadratic scaling in the number of in-context samples. Traditional ensembles train in minutes and test in seconds, but as shown in Section 4, they fail to recall multiple minority classes. TabNet incurs the highest training times among the baselines (1–3+ h), with modest test latency.

Computational analysis reveals fundamental architectural trade-offs across model families [10]. On CIC-IDS2017, traditional ML models exhibit the conventional training-then-inference pattern: For example, Random Forest requires 3–14 min of training but subsequently achieves rapid inference (4–7 s for complete datasets, approximately 0.006–0.01 ms per sample), making it exceptionally well-suited for high-throughput deployment. TabPFN fundamentally inverts this relationship: its training-free paradigm eliminates per-dataset

fitting (requiring only 288–712 ms for preprocessing) but concentrates computational resources on inference time, ranging from 3 to 97 min depending on the dataset variant. Nevertheless, for real-time deployment that processes individual samples or small batches, TabPFN's per-sample latency is far more competitive (approximately 0.8–7.8 ms per sample), making it potentially viable in IDS pipelines that can tolerate millisecond-level delays in exchange for robust minority-class detection. Comparable patterns are observed on N-BaIoT, where grouped-class configurations reduce per-sample latency to sub-2 ms.

TabICL exhibits a latency profile that positions it between traditional ensembles and TabPFN. Its setup phase is lightweight (0.5–1.0 s) and requires no dataset-specific retraining, while inference on complete datasets consistently completes within 1–3 min (roughly 0.3–0.5 ms per sample). This balance provides training-free adaptability with practical scalability: on CIC-IDS2017, TabICL reduces full-dataset inference from hours (TabPFN) to minutes while retaining non-zero recall across all attack classes, and on N-BaIoT, it achieves per-sample inference below 2 ms in grouped-class variants. These characteristics make TabICL particularly suitable for operational IDS deployment that demands near-real-time throughput without sacrificing adaptability to novel attack types.

4.5.3. LLM API Constraints

In contrast, LLM inference costs with the family of models Gemini 2.0 prove prohibitive for large-scale deployment, with per-sample inference times exceeding 1 s and API costs scaling linearly with evaluation size. These limitations restrict LLM applicability to explanatory roles or low-throughput applications where interpretability outweighs efficiency concerns.

4.6. Robustness and Generalization Analysis

The consistency of foundation model advantages across diverse datasets, data variants, and class distributions demonstrates robust generalization capabilities that transcend specific dataset characteristics. TabICL's performance remains stable across all experimental conditions, while traditional models exhibit significant variance, particularly on PCA-compressed variants.

Statistical significance testing confirms the superiority of foundation models, with TabICL achieving statistically significant improvements over traditional approaches in recall performance across all evaluation scenarios. That is, TabICL demonstrates superior recall capability by successfully detecting all attack types, including rare threats that traditional methods completely miss despite high overall accuracy. The consistency of these results across multiple datasets provides strong evidence for the fundamental advantages of foundation model approaches in cybersecurity applications.

Most critically, the unique ability of TabPFN and TabICL to detect all attack classes—including rare threats that traditional methods completely miss—represents a paradigm shift in intrusion detection capabilities. This comprehensive detection capability, achieved without traditional training requirements, establishes foundation models as the optimal choice for next-generation cybersecurity systems requiring robust protection against both common and sophisticated attack patterns.

5. Discussion

Our findings fundamentally challenge the prevailing assumption that tree-based ensemble methods represent the optimal approach for tabular intrusion detection systems. The superior performance of foundation models, particularly TabICL's achievement of 99.59% accuracy on CIC-IDS2017 while maintaining comprehensive detection across all

attack classes, suggests a paradigm shift toward foundation model architectures in cybersecurity applications.

The absence of retraining requirements in TabPFN and the hyperparameter-free, in-context adaptation of TabICL offer unprecedented advantages for operational deployment. Unlike traditional approaches, which require extensive hyperparameter tuning and retraining for new environments, these foundation models enable immediate adaptation to novel attack patterns through their generalizable pretrained representations. This capability is particularly crucial to defending against zero-day threats, where rapid response without extensive model retraining can determine the difference between successful defense and catastrophic breach.

The most significant finding of our study lies in the fundamental limitation exposed in traditional ensemble methods regarding minority-class detection. While Random Forest and k-NN achieve impressive overall accuracy exceeding 99%, their complete failure to detect critical threats like Heartbleed and Infiltration reveals a dangerous vulnerability that high-level accuracy metrics obscure.

This limitation has profound implications for real-world cybersecurity deployment. In operational environments, the most sophisticated and dangerous attacks often manifest as minority classes in training data, precisely because they represent novel or carefully crafted threats designed to evade detection. A system that achieves 99% accuracy while missing 100% of advanced persistent threats provides a false sense of security that could prove catastrophic.

Foundation models' unique ability to detect all attack classes stems from their pretrained representations that capture generalizable patterns rather than dataset-specific decision boundaries. This fundamental architectural advantage enables robust detection of rare patterns that traditional ensemble methods, despite their statistical robustness, cannot reliably identify due to insufficient training examples.

5.1. Theoretical and Practical Implications

Our results demonstrate that overall accuracy—the dominant metric in tabular machine learning—provides misleading assessments for cybersecurity applications. The stark contrast between 99%+ overall accuracy and zero recall on critical attack classes exposes the inadequacy of aggregate metrics for security-critical systems where comprehensive threat coverage is paramount.

The per-class evaluation framework we employed reveals the true detection capabilities of different approaches, providing security practitioners with actionable insights about model reliability across the complete threat spectrum. This methodological shift toward granular performance assessment should become standard practice in cybersecurity machine learning evaluations.

The consistent superiority of foundation models across diverse datasets (CIC-IDS2017, N-BaIoT, and CIC-UNSW) suggests that their pretrained representations capture fundamental patterns of network behavior that generalize across different threat landscapes and network configurations. This universality contrasts sharply with traditional methods that exhibit dataset-specific performance variations, requiring careful tuning for each new environment.

TabICL's enhanced scalability compared with TabPFN represents a crucial advancement, enabling foundation model benefits to extend to large-scale operational deployment. The 1.5-10× computational speedup while maintaining superior detection capabilities positions TabICL as particularly suitable for high-throughput network monitoring scenarios.

The inherent explainability of LLM-based approaches, despite their computational limitations, points toward future cybersecurity architectures that combine high-performance

detection with interpretable threat analysis. While current LLM constraints prevent large-scale deployment, the demonstrated capability to identify rare attack classes with natural language explanations addresses a critical gap in AI-driven security systems.

A tiered architecture utilizing TabICL for comprehensive threat detection combined with LLM-based analysis for critical alerts could optimize both performance and explainability requirements. This hybrid approach leverages the computational efficiency of foundation models while providing the interpretability essential to security analyst workflows and incident response procedures.

The no-retraining paradigm of foundation models enables unprecedented adaptability in dynamic threat environments. Traditional approaches requiring extensive retraining cycles for new attack patterns cannot match the immediate adaptation capabilities of foundation models that leverage pretrained representations to recognize novel threats based on structural similarities to known patterns.

This adaptability becomes particularly valuable for defending against campaign-based attacks where threat actors continuously evolve their techniques. Foundation models' ability to generalize from limited examples of new attack variants provides a crucial defensive advantage in the ongoing cybersecurity arms race.

5.2. Addressing Computational and Practical Limitations

Despite their superior performance, foundation models face significant computational barriers that limit immediate operational deployment. TabPFN's requirement for high-memory GPU infrastructure (>20 GB of VRAM) and TabICL's scaling characteristics, while improved, still necessitate substantial computational resources compared with traditional ensemble methods' modest CPU requirements.

However, the trend toward edge computing and distributed inference architectures suggests that these limitations may diminish as computational infrastructure evolves. The fundamental algorithmic advantages demonstrated by foundation models justify investment in appropriate infrastructure for organizations prioritizing comprehensive threat detection over computational efficiency.

The training-free paradigm's computational resources being spent on inference time creates potential bottlenecks in high-throughput network monitoring scenarios. While foundation models achieve superior detection capabilities, their inference latency characteristics require careful consideration for real-time applications where millisecond response times are critical.

Hybrid architectures utilizing lightweight traditional methods for initial traffic filtering followed by foundation model analysis for suspicious flows could balance performance requirements with computational constraints. This tiered approach maximizes threat detection capabilities while maintaining operational efficiency for routine network traffic.

5.3. Practical Integration Framework for Security Operations Centers

The superior detection capabilities demonstrated by foundation models necessitate careful integration strategies that address operational constraints while maximizing threat coverage. We propose a tiered detection architecture for practical SOC deployment that optimizes computational resources while leveraging foundation model advantages.

Three-Tier Detection Pipeline: The first tier employs lightweight traditional methods (Random Forest or k-NN) for high-throughput initial screening, handling 80–90% of network traffic with minimal computational overhead while maintaining real-time processing capabilities. The second tier activates TabICL for suspicious flows that pass initial screening, leveraging its enhanced scalability for comprehensive detection across all attack

classes. The third tier utilizes LLM-based analysis for critical alerts, providing detailed threat assessment with natural language explanations for security analysts.

Implementation Example: Consider an SOC monitoring 10,000 flows per minute. First-tier Random Forest processing identifies 1000 potentially suspicious flows, which are queued for TabICL analysis in 5 s inference cycles. Approximately 50 high-priority threats per minute trigger LLM-based explanatory analysis, providing analysts with detailed threat descriptions and response recommendations.

Resource Allocation: This hybrid approach requires CPU-based servers for traditional model inference combined with dedicated GPU clusters (NVIDIA A100 or equivalent) for foundation model processing. Auto-scaling mechanisms dynamically allocate GPU resources based on suspicious flow volumes, ensuring cost-effectiveness during normal operations while providing surge capacity during attacks.

Alert Prioritization: Foundation models' unique detection of rare attack classes enables sophisticated alert prioritization. Threats detected only by foundation models (Heartbleed, Infiltration, APTs, etc.) receive the highest priority, as these represent sophisticated attacks that traditional methods completely miss. This workflow reduces alert fatigue while ensuring comprehensive coverage of both common and advanced attack patterns.

Migration Strategy: Organizations should implement foundation models in parallel with existing systems during initial deployment, enabling performance validation while maintaining security coverage. Gradual migration begins with high-priority networks, expanding coverage as operational confidence and computational infrastructure scale. The infrastructure investment is most justified for organizations facing advanced persistent threats or operating critical infrastructure, where detecting sophisticated attacks outweighs GPU infrastructure costs.

The tiered architecture enables organizations to balance computational efficiency with comprehensive threat detection through strategic resource allocation. Figure 4 illustrates the operational workflow where traditional ensemble methods handle high-volume traffic filtering, foundation models provide comprehensive analysis of suspicious flows, and LLMs deliver interpretable threat assessment for critical alerts. This approach leverages each model family's strengths while mitigating their operational limitations, ensuring real-time processing capabilities while maintaining detection of rare attack classes that traditional methods completely miss.

5.3.1. Risks and Hybrid Opportunities

While our results highlight the transformative potential of tabular foundation models for intrusion detection, it is important to recognize associated risks. Transformer-based architectures are known to be susceptible to adversarial perturbations, where carefully crafted inputs can mislead attention mechanisms and degrade performance. Moreover, both TabPFN and TabICL rely on pretrained representations whose robustness depends on the breadth and diversity of their training corpora; this introduces a dependency that is absent in classical ensemble methods, which can always be retrained from scratch on domain-specific data. Finally, compared with tree-based ensembles, foundation models reduce interpretability, a critical limitation for security operations where analysts require transparent rationales for threat classification.

These considerations suggest that hybrid architectures may offer the most practical path forward. One promising direction is to deploy TabICL as the primary high-throughput detector, leveraging its training-free adaptability, while incorporating LLM-based modules to provide natural-language explanations for alerts and assist analysts in incident triage. Another avenue is to combine lightweight ensemble methods for rapid filtering with foundation models for deeper analysis of suspicious flows, balancing efficiency with

robustness. By explicitly addressing these risks and exploring hybrid integration, the cybersecurity community can capitalize on the advantages of foundation models while mitigating their current vulnerabilities.

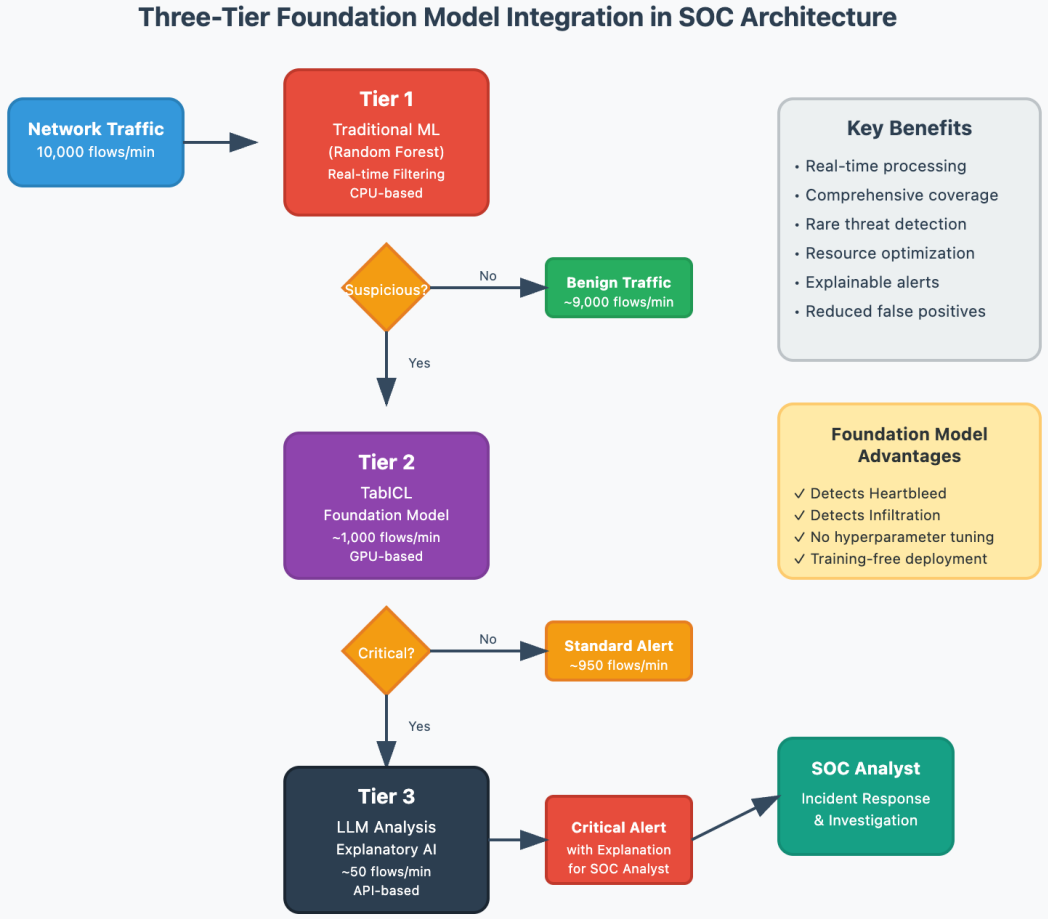

Figure 4. Three-tier foundation model integration architecture for security operations centers showing traffic flow volumes, decision points, and resource allocation strategy that optimizes computational efficiency while ensuring comprehensive threat detection including rare attacks like Heartbleed and Infiltration.

5.3.2. Evaluation Under Natural Imbalance

We acknowledge the concern regarding potential bias introduced by applying balanced sampling to TabPFN and TabICL while training traditional methods on the full imbalanced datasets. This design choice does not reflect a methodological preference but a structural constraint: TabPFN and TabICL cannot process arbitrarily large class distributions due to quadratic scaling (TabPFN) or memory usage (TabICL). Their intended operational use is precisely in settings where balanced few-shot subsets are drawn from streaming traffic or incident logs, rather than in full retraining on billions of flows. In contrast, classical ensemble methods are designed to exploit large, imbalanced training sets, which is why we preserved their conventional evaluation regime.

Importantly, our test splits remain unaltered and naturally imbalanced across all models, ensuring that the reported performance reflects the true class frequencies encountered in deployment. The superior minority-class recall achieved by TabPFN and TabICL highlights their robustness in detecting rare classes despite being trained on capped subsets. While retraining under full imbalance is computationally infeasible for TabPFN and TabICL, future work will explore incremental or streaming evaluation strategies that more closely mimic online IDS conditions.

5.4. Limitations and Future Research Directions

While our evaluation encompasses diverse cybersecurity datasets, the focus on network flow data may not capture the full spectrum of cybersecurity applications. That is, the datasets employed (CIC-IDS2017 and CIC-UNSW) are flow-based, with each record summarizing an entire communication. This implies that classification can typically occur only after the full flow has been observed, rather than in real time on streaming packets. Future research should extend evaluation of foundation models to complementary domains such as endpoint security, log analysis, and vulnerability assessment to establish broader applicability.

The temporal aspects of cybersecurity threats—where attack patterns evolve continuously—warrant investigation of foundation models' adaptation capabilities over extended periods. Longitudinal studies examining performance degradation and adaptation strategies will provide crucial insights for operational deployment planning.

The rapid evolution of foundation model architectures suggests that our evaluation represents a snapshot of current capabilities rather than fundamental limits. Emerging approaches combining the scaling advantages of TabICL with enhanced efficiency optimizations may further expand foundation model applicability to resource-constrained environments.

Research into domain-specific pretraining for cybersecurity applications could potentially enhance foundation model performance beyond the general-purpose architectures evaluated in our study. Cybersecurity-specific foundation models trained on diverse security datasets may achieve even stronger detection capabilities while maintaining generalization advantages.

Our findings suggest that the cybersecurity community's continued investment in increasingly sophisticated ensemble methods may represent a suboptimal research direction. The fundamental advantages demonstrated by foundation models—comprehensive threat detection, training-free deployment, and robust generalization—indicate that future advances in cybersecurity machine learning should prioritize foundation model development over traditional ensemble refinement.

This strategic shift requires significant changes in research priorities, educational curricula, and industry deployment practices. However, the potential for defense against sophisticated threats that current approaches miss justifies the substantial investment required for this technological transition.

The demonstrated superiority of foundation models in detecting rare but critical threats positions them as essential tools for defending against advanced persistent threats and zero-day attacks. Organizations prioritizing comprehensive security over computational efficiency should begin evaluating foundation model integration into their cybersecurity architectures, recognizing that the paradigm shift toward adaptive, hyperparameter-free cybersecurity systems that avoid costly retraining.

Beyond computational constraints, several critical operational risks warrant careful consideration. Resource overload scenarios during high-traffic periods or coordinated attacks could create processing delays that enable threats to propagate while foundation model inference queues struggle with computational demands. Our evaluation framework's reliance on balanced sampling for TabPFN and TabICL may introduce optimistic bias, as this artificial balance does not reflect realistic operational environments, where severe class imbalance persists without mitigation. Additionally, foundation models' limited interpretability compared with ensemble methods' transparent decision trees poses challenges for security analysts requiring attack classification rationale for incident response. The adversarial vulnerability inherent to neural architectures may also expose foundation models to sophisticated evasion techniques that simpler ensemble decision boundaries

might resist, while dependency on pretrained representations creates single points of failure absent in traditional methods that can be retrained from scratch with domain-specific data.

6. Conclusions and Future Work

This work presents a comprehensive multi-modal evaluation of foundation models for tabular intrusion detection across diverse cybersecurity environments, fundamentally challenging the cybersecurity community's reliance on tree-based ensemble methods that have dominated the field for over a decade. Through rigorous experimentation across three distinct datasets (CIC-IDS2017, N-BaIoT, and CIC-UNSW), we demonstrate that foundation models offer superior and more consistent performance compared with traditional approaches, establishing a new paradigm for intelligent intrusion detection systems.

Our research makes four critical contributions to both the cybersecurity and machine learning communities:

1. Comprehensive Multi-Modal Evaluation of Foundation Models: We establish the first systematic comparison of multiple foundation model approaches—TabPFN, TabICL, and LLMs—against traditional machine learning methods, providing practitioners with definitive guidance on optimal architectures for cybersecurity applications.

2. Model-Appropriate Evaluation Framework: We develop rigorous experimental protocols that implement tailored sampling strategies reflecting the operational constraints and optimal usage patterns of different model families, addressing methodological limitations that have historically biased evaluations toward traditional ensemble methods.

3. Cross-Dataset Generalization Validation: Through systematic evaluation across diverse threat landscapes and network configurations, we demonstrate that foundation models maintain performance advantages while traditional methods exhibit significant dataset-specific variations, establishing foundation models' superior generalization capabilities.

4. Exposure of Critical Limitation in Traditional Methods: We reveal that ensemble methods achieving >99% overall accuracy completely fail to detect minority attack classes, exposing a dangerous vulnerability that foundation models uniquely address through comprehensive threat detection capabilities.

The most significant finding of our study is the paradigm-shifting performance of foundation models, particularly TabICL's achievement of 99.59% accuracy on CIC-IDS2017 while uniquely detecting all attack classes including rare threats like Heartbleed and Infiltration. This comprehensive detection capability, achieved without traditional training requirements, represents a breakthrough for zero-day threat detection in rapidly evolving cybersecurity landscapes.

Our findings have profound implications for cybersecurity operations and research priorities.

Rethinking the IDS Architecture: The superior performance of foundation models across all experimental scenarios challenges the fundamental assumption that tree-based ensembles represent optimal approaches for tabular cybersecurity applications. Organizations should begin evaluating foundation model integration into their security architectures, recognizing that comprehensive threat detection capabilities justify infrastructure investment requirements.

Evaluation Methodology Reform: The stark contrast between high overall accuracy and zero recall on critical attack classes demonstrates that aggregate metrics provide misleading assessments for security-critical systems. The cybersecurity community must adopt per-class evaluation frameworks that reveal true detection capabilities across the complete threat spectrum.

Operational Advantages of No Per-dataset Retraining: Foundation models' ability to adapt to novel attack patterns without extensive retraining cycles offers unprecedented

advantages for defending against zero-day threats and advanced persistent threats. This capability fundamentally alters the timeline for threat response in operational environments, where rapid adaptation can determine defensive success.

Hybrid Architecture Opportunities: The complementary strengths of different foundation model approaches suggest that optimal performance may emerge from hybrid architectures leveraging TabICL for comprehensive detection, combined with LLM-based analysis for explainable threat assessment in security operations centers.

While foundation models demonstrate clear superiority for detection of rare threats, several limitations must be addressed for widespread operational deployment. Computational Infrastructure Requirements: Foundation models necessitate substantial GPU resources compared with traditional ensemble methods' modest CPU requirements. However, the trend toward edge computing and distributed inference architectures suggests that these barriers will diminish as computational infrastructure evolves.

Inference Latency Considerations: The training-free paradigm's computational resources being spent on inference time requires careful consideration for real-time applications. Tiered architectures utilizing lightweight methods for initial filtering followed by foundation model analysis for suspicious flows can balance performance with efficiency requirements.

Scalability and Memory Constraints: Current foundation model implementations face memory limitations that restrict dataset sizes and deployment scenarios. Continued research into efficient architectures and memory optimization will expand applicability to resource-constrained environments.

Concluding Remarks

The foundation model revolution in artificial intelligence has reached cybersecurity applications, offering unprecedented capabilities for comprehensive threat detection without traditional training requirements. Our systematic evaluation demonstrates that the future of intelligent intrusion detection lies not in increasingly sophisticated ensemble methods but in leveraging the generalizable representations learned by foundation models.

The unique ability of TabPFN and TabICL to detect all attack classes—including rare threats that traditional methods completely miss—represents a paradigm shift toward adaptive, hyperparameter-free cybersecurity systems that avoid costly retraining capable of defending against both common and sophisticated attack patterns. While computational requirements present current deployment challenges, the fundamental advantages in threat detection capabilities justify the infrastructure investment required for this technological transition.

As cyberthreats continue evolving in sophistication and scale, the cybersecurity community must embrace foundation model approaches that offer robust generalization, rapid adaptation, and comprehensive detection capabilities. The evidence presented in this work provides a clear roadmap for this transition, establishing foundation models as essential tools for next-generation cybersecurity defense systems capable of protecting against the dynamic threat landscape of modern digital infrastructure.

The paradigm shift toward foundation models in cybersecurity represents not merely a technological upgrade but also a fundamental reimagining of how intelligent systems can adapt, generalize, and defend against threats that traditional approaches cannot reliably detect. This transformation positions foundation models as the cornerstone of future cybersecurity architectures designed to meet the challenges of an increasingly complex and dangerous digital world.

7. Code Availability

The implementation of the framework based on foundation models for tabular intrusion detection is publicly available at https://github.com/pablogarciaamolina/AI-for-IDS (accessed on 1 September 2025). The repository includes data preprocessing modules, model implementations, evaluation pipelines, and reproducibility instructions.

Author Contributions: Conceptualization, P.G. and J.d.C.; data curation, P.G.; formal analysis, P.G., J.d.C. and I.d.Z.; funding acquisition, J.d.C., I.d.Z., J.C.C. and C.T.C.; investigation, P.G. and J.d.C.; methodology, P.G., J.d.C., I.d.Z. and J.C.C.; software, P.G.; supervision, J.d.C., I.d.Z. and J.C.C.; validation, P.G., J.d.C., I.d.Z., J.C.C. and C.T.C.; visualization, P.G., J.d.C. and I.d.Z.; writing—original draft, J.d.C. and I.d.Z.; writing—review and editing, J.d.C., P.G., I.d.Z., J.C.C. and C.T.C. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank the BARCELONA Supercomputing Center for providing access to MareNostrum 5 and technical support throughout this research study. This work was developed under the following project: "TIFON".

Data Availability Statement: The data presented in this study are openly available in GitHub at https://github.com/pablogarciaamolina/AI-for-IDS, accessed on 1 September 2025.

Conflicts of Interest: The authors declare that they have 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.

Abbreviations

The following abbreviations are used in this manuscript:

Artificial intelligence	AI
Application Programming Interface	API
Advanced persistent threat	APT
Area Under the Curve	AUC
Convolutional Neural Network	CNN
Gradient Boosted Decision Trees	GBDTs
Graphics Processing Unit	GPU
In-context learning	ICL
Intrusion detection system	IDS
Internet of Things	IoT
k-Nearest Neighbors	k-NN
Large language model	LLM
Machine learning	ML
Natural language processing	NLP
Principal Component Analysis	PCA
Prior-Data Fitted Network	PFN
Random Access Memory	RAM
Random Forest	RF
Synthetic Minority Oversampling Technique	SMOTE
Support Vector Machine	SVM
Tabular In-Context Learning	TabICL
Tabular Prior-Data Fitted Network	TabPFN

References

1. Meneghello, F.; Calore, M.; Zucchetto, D.; Polese, M.; Zanella, A. IoT: Internet of threats? A survey of practical security vulnerabilities in real IoT devices. *IEEE Internet Things J.* **2019**, *6*, 8182–8201. [CrossRef]
2. Noor, M.B.M.; Hassan, W.H. Current research on Internet of Things (IoT) security: A survey. *Comput. Netw.* **2019**, *148*, 283–294. [CrossRef]

3. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 8–10 July 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 1–6.
4. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, ACT, Australia, 10–12 November 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–6.
5. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the International Conference on Information Systems Security and Privacy ICISSP, Funchal, Portugal, 22–24 January 2018; Volume 1, pp. 108–116.
6. Koneru, S.S.; Cho, J. Bridging the Gap: A Comparative Analysis of ICS and IT Datasets for IDS Evaluation. In Proceedings of the 2024 2nd International Conference on Foundation and Large Language Models (FLLM), Dubai, United Arab Emirates, 26–29 November 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 300–304.
7. Hasanov, I.; Virtanen, S.; Hakkala, A.; Isoaho, J. Application of Large Language Models in Cybersecurity: A Systematic Literature Review. *IEEE Access* **2024**, *12*, 176751–176778. [CrossRef]
8. Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. TabPFN: A transformer that solves small tabular classification problems in a second. In Proceedings of the International Conference on Learning Representations 2023 (ICLR 2023), Kigali, Rwanda, 1–5 May 2023.
9. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326. [CrossRef]
10. García, P.; de Curtò, J.; de Zarzà, I. Foundation Models for Tabular Intrusion Detection: Evaluating TabPFN and LLM Few-Shot Classification on IoT Network Security. In Proceedings of the 2025 3rd International Conference on Foundation and Large Language Models (FLLM), Vienna, Austria, 25–28 November 2025; IEEE: Piscataway, NJ, USA, 2025.
11. Qu, J.; Holzmüller, D.; Varoquaux, G.; Morvan, M.L. TabICL: A Tabular Foundation Model for In-Context Learning on Large Data. *arXiv* **2025**, arXiv:2502.05564.
12. Han, S.; Yoon, J.; Arik, S.O.; Pfister, T. Large language models can automatically engineer features for few-shot tabular learning. *arXiv* **2024**, arXiv:2404.09491. [CrossRef]
13. Keltek, M.; Hu, R.; Sani, M.F.; Li, Z. LSAST: Enhancing Cybersecurity Through LLM-Supported Static Application Security Testing. In Proceedings of the IFIP International Conference on ICT Systems Security and Privacy Protection, Maribor, Slovenia, 21–23 May 2025; Springer: Berlin/Heidelberg, Germany, 2025; pp. 166–179.
14. Muhammad, M.; Shaaban, A.M.; German, R.; Al Sardy, L. HyLLM-IDS: A Conceptual Hybrid LLM-Assisted Intrusion Detection Framework for Cyber-Physical Systems. In Proceedings of the International Conference on Computer Safety, Reliability, and Security, Stockholm, Sweden, 9–12 September 2025; Springer: Berlin/Heidelberg, Germany, 2025; pp. 129–142.
15. DeCusatis, C.; Tomo, R.; Singh, A.; Khoury, E.; Masone, A. Cybersecurity Applications of Near-Term Large Language Models. *Electronics* **2025**, *14*, 2704. [CrossRef]
16. Coppolino, L.; Iannaccone, A.; Nardone, R.; Petruolo, A. Asset Discovery in Critical Infrastructures: An LLM-Based Approach. *Electronics* **2025**, *14*, 3267. [CrossRef]
17. Lai, T.; Farid, F.; Bello, A.; Sabrina, F. Ensemble learning based anomaly detection for IoT cybersecurity via Bayesian hyperparameters sensitivity analysis. *Cybersecurity* **2024**, *7*, 44. [CrossRef]
18. Hossain, M.A.; Islam, M.S. Enhancing DDoS attack detection with hybrid feature selection and ensemble-based classifier: A promising solution for robust cybersecurity. *Meas. Sens.* **2024**, *32*, 101037. [CrossRef]
19. Buczak, A.L.; Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Commun. Surv. Tutor.* **2015**, *18*, 1153–1176. [CrossRef]
20. Khraisat, A.; Gondal, I.; Vamplew, P.; Kamruzzaman, J. Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity* **2019**, *2*, 1–22. [CrossRef]
21. Liao, H.J.; Lin, C.H.R.; Lin, Y.C.; Tung, K.Y. Intrusion detection system: A comprehensive review. *J. Netw. Comput. Appl.* **2013**, *36*, 16–24. [CrossRef]
22. Liu, H.; Lang, B. Machine learning and deep learning methods for intrusion detection systems: A survey. *Appl. Sci.* **2019**, *9*, 4396. [CrossRef]
23. Lansky, J.; Ali, S.; Mohammadi, M.; Majeed, M.K.; Karim, S.H.T.; Rashidi, S.; Hosseinzadeh, M.; Rahmani, A.M. Deep learning-based intrusion detection systems: A systematic review. *IEEE Access* **2021**, *9*, 101574–101599. [CrossRef]
24. Arik, S.Ö.; Pfister, T. Tabnet: Attentive interpretable tabular learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 6679–6687.
25. de Zarzà, I.; de Curtò, J.; Calafate, C.T. Area Estimation of Forest Fires using TabNet with Transformers. In *Procedia Computer Science, Proceedings of the 27th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES 2023), 6–8 September 2023, Athens, Greece*; Elsevier: Amsterdam, The Netherlands, 2023; Volume 225, pp. 553–563. [CrossRef]

26. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
27. Fernández, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [CrossRef]
28. Mohammad, R.; Saeed, F.; Almazroi, A.A.; Alsubaei, F.S.; Almazroi, A.A. Enhancing Intrusion Detection Systems Using a Deep Learning and Data Augmentation Approach. *Systems* **2024**, *12*, 79. [CrossRef]
29. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; Volume 33, pp. 1877–1901.
30. Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *Int. J. Mach. Learn. Cybern.* **2024**, 1–65. [CrossRef]
31. de Curtò, J.; de Zarzà, I.; Roig, G.; Cano, J.C.; Manzoni, P.; Calafate, C.T. LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments. *Electronics* **2023**, *12*, 2814. [CrossRef]
32. Balogh, S.; Mlyncek, M.; Vranák, O.; Zajac, P. Using Generative AI Models to Support Cybersecurity Analysts. *Electronics* **2024**, *13*, 4718. [CrossRef]
33. Moraga, A.; de Curtò, J.; de Zarzà, I.; Calafate, C.T. AI-Driven UAV and IoT Traffic Optimization: Large Language Models for Congestion and Emission Reduction in Smart Cities. *Drones* **2025**, *9*, 248. [CrossRef]
34. Song, C.H.; Wu, J.; Washington, C.; Sadler, B.M.; Chao, W.L.; Su, Y. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 2998–3009.
35. Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; Sontag, D. Tabllm: Few-shot classification of tabular data with large language models. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 25–27 April 2023; PMLR: Cambridge, MA, USA, 2023; pp. 5549–5581.
36. Zhang, J.; Bu, H.; Wen, H.; Liu, Y.; Fei, H.; Xi, R.; Li, L.; Yang, Y.; Zhu, H.; Meng, D. When llms meet cybersecurity: A systematic literature review. *Cybersecurity* **2025**, *8*, 1–41. [CrossRef]
37. Yamin, M.M.; Hashmi, E.; Ullah, M.; Katt, B. Applications of llms for generating cyber security exercise scenarios. *IEEE Access* **2024**, *12*, 143806–143822. [CrossRef]
38. Li, Y.; Xiang, Z.; Bastian, N.D.; Song, D.; Li, B. IDS-Agent: An LLM Agent for Explainable Intrusion Detection in IoT Networks. In Proceedings of the NeurIPS 2024 Workshop on Open-World Agents, Vancouver, BC, Canada, 10–15 December 2024.