SAGE: Sample-Aware Guarding Engine for Robust Intrusion Detection Against Adversarial Attacks

Jing Chen*, Onat Gungor*, Zhengli Shang, and Tajana Rosing

***Abstract*—The rapid proliferation of the Internet of Things (IoT) continues to expose critical security vulnerabilities, necessitating the development of efficient and robust intrusion detection systems (IDS). Machine learning-based intrusion detection systems (ML-IDS) have significantly improved threat detection capabilities; however, they remain highly susceptible to adversarial attacks. While numerous defense mechanisms have been proposed to enhance ML-IDS resilience, a systematic approach for selecting the most effective defense against a specific adversarial attack remains absent. To address this challenge, we previously proposed DYNAMITE, a dynamic defense selection approach that identifies the most suitable defense against adversarial attacks through an ML-driven selection mechanism. Building on this foundation, we propose SAGE (Sample-Aware Guarding Engine), a substantially improved defense algorithm that integrates active learning with targeted data reduction. It employs an active learning mechanism to selectively identify the most informative input samples and their corresponding optimal defense labels, which are then used to train a second-level learner responsible for selecting the most effective defense. This targeted sampling improves computational efficiency, exposes the model to diverse adversarial strategies during training, and enhances robustness, stability, and generalizability. As a result, SAGE demonstrates strong predictive performance across multiple intrusion detection datasets, achieving an average F1-score improvement of 201% over the state-of-the-art defenses. Notably, SAGE narrows the performance gap to the Oracle to just 3.8%, while reducing computational overhead by up to 29×.**

***Index Terms*—IoT Security, Intrusion Detection, Machine Learning, Adversarial Attacks, Defense Selection**

I. Introduction

The Internet of Things (IoT) systems connect numerous devices that communicate and share data, enabling smart applications in sectors like healthcare, manufacturing, and transportation [1]. IoT systems are particularly susceptible to cyber threats due to their inter-connectivity, resource constraints, and diverse configurations [2]. Consequently, ensuring robust security measures is essential to safeguard these systems against potential attacks. Intrusion Detection Systems (IDS) play a crucial role in identifying and responding to malicious activities within IoT networks by monitoring network traffic and system behavior [1]. The integration of machine learning (ML) into IDS has significantly improved their effectiveness in detecting and mitigating cyber threats. ML-IDS possess the capability to analyze vast amounts of data, identify latent patterns, and detect cyberattacks that conventional methods may overlook [3]. Thus, ML-IDS serve as a robust approach for enhancing IoT security by addressing evolving threats. However, the rise of adversarial attacks poses a significant challenge to the effectiveness of ML-IDS [4]. These attacks allow malicious activities to go undetected and harm the security of IoT systems, leading to compromised operations, data breaches, and significant financial losses [5].

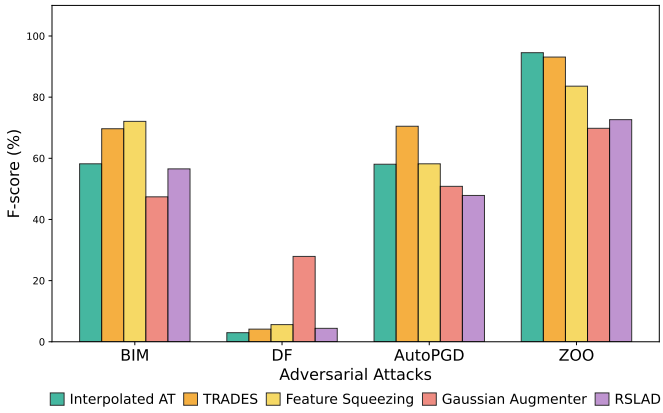

Fig. 1. State-of-the-art Defense Performance Against Adversarial Attacks

Developing effective defenses against adversarial attacks is crucial for maintaining the reliability and robustness of ML-IDS [6]. Several strategies, both general and specific to ML-IDS, have been proposed, including adversarial training [7], [8], modifications to the training process [9], [10], input transformation techniques [11], and methods for adversarial attack detection [12], [13]. However, the effectiveness of defense mechanisms varies depending on the specific type of attack they are intended to mitigate [14]. Given that adversarial attacks can differ in their techniques and objectives, tailored defense strategies are necessary to effectively address each distinct scenario. Fig. 1 demonstrates that no single defense model (represented by different colors) is universally effective against all adversarial attacks, with the optimal defense varying depending on the specific nature of the attack (as shown on the x-axis). This variability highlights the limitation of relying on a singular defense mechanism for comprehensive protection. It further emphasizes the importance of a dynamic defense selection mechanism that adaptively assigns the most appropriate defense for each attack scenario. Such an approach is crucial for achieving robust security, as it ensures the real-time deployment of the most effective defense in response to the evolving nature of adversarial attacks.

Building upon our preliminary work, DYNAMITE [15],

* Jing Chen and Onat Gungor contributed equally to this work.
Department of Computer Science and Engineering, University of California, San Diego, CA 92093 USA (email: {jic128, ogungor, z4shang, tajana}@ucsd.edu).

this paper addresses two key opportunities for advancing ML-IDS defense frameworks: enhancing computational efficiency and strengthening robustness against previously unseen adversarial attacks. Maximizing efficiency is critical for practical ML-IDS defense deployment in IoT environments, where training on large datasets and running complex defense models pose significant challenges. To this end, our extended framework employs a targeted subsample selection strategy that curates a minimal yet highly informative training set, substantially reducing computational overhead. Furthermore, enhancing robustness against novel threats is essential in an evolving adversarial landscape, where static defenses can be quickly circumvented. To address this, we leverage the Entropic Open-set Active Learning (EOAL) algorithm [16], which is particularly suited to our objectives. Together, these advances yield a defense framework that is both scalable and resilient, systematically enhancing efficiency and robustness.

We introduce the Sample-Aware Guarding Engine (SAGE), an adaptive and efficient ML-IDS defense framework. Unlike static defense mechanisms, SAGE dynamically selects the optimal defense strategy against adversarial attacks. The SAGE pipeline, depicted in Fig. 2, begins with data preprocessing and the training of a baseline ML-IDS model alongside several SOTA defense models. To simulate diverse threat landscapes, we generate adversarial samples using various attack strategies and intensities. Each defense model is then evaluated against these samples, and performance metrics are used to label each sample with its most effective defense. To optimize the training of our defense selection mechanism, SAGE employs Entropic Open-set Active Learning (EOAL) [16] to identify and prioritize the most informative and diverse samples. This approach significantly reduces the training data volume while preserving high performance. A second-level learner is then trained on this reduced dataset to predict the most effective defense for previously unseen adversarial attacks. Our experiments on multiple realistic intrusion detection datasets show that SAGE surpasses state-of-the-art defenses, yielding an average F1-score improvement of 201%. Moreover, SAGE remains robust against previously unseen adversarial attacks, demonstrating its ability to adapt to new threats. Notably, by leveraging EOAL's focus on sparsely represented regions, SAGE achieves comparable or superior robustness using only 1% of labeled data, with gains of up to 3.4 points over full supervision. SAGE also substantially improves computational efficiency, achieving up to a 29× speedup compared to the Oracle, while maintaining only a 3.8% gap in F1-score. These findings position SAGE as a robust and efficient defense framework that achieves high accuracy while substantially reducing computational overhead.

II. Background and Related Work

A. ML-based Intrusion Detection Systems (ML-IDS)

The growing dependence on computer networks and the expansion of the Internet of Things (IoT) have introduced significant security challenges, driven by the increasing complexity and diversity of these interconnected systems [17]. Intrusion Detection Systems (IDS) are designed to monitor network activity and detect malicious behavior. IDS methods are broadly categorized into two types: signature-based and anomaly-based [4]. Signature-based IDS rely on predefined signatures of known attacks to identify malicious activities, offering high accuracy for detecting known threats but struggling with zero-day attacks. In contrast, anomaly-based IDS detect deviations from normal network behavior, enabling them to identify previously unseen attacks, though they may suffer from higher false-positive rates. This limitation has motivated the integration of ML techniques into IDS, enhancing their ability to identify complex and evolving attack patterns with greater accuracy [18]. A range of ML models, such as Decision Trees (DT), Random Forests (RF), and Deep Neural Networks (DNN), have been utilized in ML-IDS to improve detection capabilities [19]. By leveraging large datasets, these models can learn intricate patterns of network behavior, surpassing the performance of traditional methods. However, despite their effectiveness, ML-IDS solutions remain vulnerable to adversarial attacks, where malicious actors manipulate input data to evade detection or degrade system performance [4].

B. Adversarial Attacks

Adversarial attacks manipulate ML models by introducing small, intentional, and often imperceptible perturbations to input data [20]. These attacks are especially critical in ML-IDS since they can exploit model vulnerabilities to evade detection [4]. While white-box attacks have full access to the model's details for gradient-based perturbations, black-box attacks generate adversarial examples without internal knowledge, using queries or transfer methods [21]. Overall, we select nine state-of-the-art, widely used white-box and black-box adversarial attacks: Fast Gradient Sign Method (FGSM) [22], Basic Iterative Method (BIM) [23], Projected Gradient Descent (PGD) [7], Auto Projected Gradient Descent (AutoPGD) [24], DeepFool (DF) [25], Zeroth Order Optimization (ZOO) [26], Scale-Invariant Nesterov Iterative Fast Gradient Sign Method (SINI-FGSM) [27], Variance-tuned Nesterov Iterative Fast Gradient Sign Method (VNI-FGSM) [28], and Cost-aware Feasible Attack (CaFA) [29].

C. Adversarial Defenses

Defense mechanisms aim to protect ML models from adversarial attacks by enhancing their robustness or reducing the effectiveness of adversarial perturbations [30]. We evaluate ten representative defenses across four principal categories: (i) *Adversarial Training (AT)*: methods that incorporate adversarial examples during training to improve model robustness, including Projected Gradient Descent AT (PGD-AT) [7], Interpolated AT (IAT) [31], TRadeoff-inspired Adversarial DEfense via Surrogate-loss Minimization (TRADES) [8], and Free AT (FAT) [32]. (ii) *Training-Process Modification*: approaches that enhance robustness via training-time augmentation or architectural adjustments without explicit inner maximization, exemplified by Gaussian Augmenter (GA) [33] and Robust Generative Adaptation Network (RGAN) [34]. (iii) *Robust Network Design*: architectural or knowledge-distillation-based

methods that smooth decision boundaries and improve stability, including Defensive Distillation (DD) [9] and Robust Soft Label Adversarial Distillation (RSLAD) [10]. (iv) *Input Preprocessing*: inference-time transformations applied to inputs to mitigate adversarial perturbations, represented by Feature Squeezing (FS) [11] and Gaussian Noise (GN) [35]. These defenses form the static portfolio from which the selector assigns the most appropriate defense to each sample.

D. Adversarial Defense Mechanisms in ML-IDS

Several efforts have been directed toward developing adversarial defense mechanisms specifically tailored to enhance the robustness of ML-IDS against adversarial attacks. Han et al. [36] address traffic-space attacks targeting ML-based NIDS, proposing a defense scheme that reduces evasion rates across multiple attack scenarios. Debicha et al. [12] introduce AdvBot, a framework for generating adversarial botnet traffic to test and strengthen IDS defenses. Debicha et al. [13] present a transfer learning-based framework that employs multiple adversarial detectors to improve detection rates. Alslman et al. [37] enhance ML-IDS resilience in 5G networks with a CycleGAN-based adversarial recovery mechanism. Similarly, Holla et al. [38] propose a dual-layered defense for ML-IDS in cloud environments, which combines adversarial training with SHAP-based feature selection. Existing studies on ML-based IDS often rely on isolated or manually selected defense mechanisms, which limits their generalizability. These methods also remain highly sensitive to subtle adversarial perturbations, leading to undetected intrusions in dynamic environments.

In contrast, SAGE frames defense as a per-sample decision over the state-of-the-art methods. Rather than assuming adversarial attack knowledge, it trains a second-level learner to select the most suitable defense. The resulting attack-agnostic mechanism integrates evidence across defenses during training and, at deployment, assigns an appropriate defense to each sample in real time. This shift from static defenses to data-driven, sample-aware selection improves adaptability in evolving threat environments while keeping inference lightweight and deployment practical for ML-IDS.

E. Dynamic Defense Methods

Dynamic defense strategies have been investigated across several cybersecurity domains. Zheng et al. [39] examined approaches such as Moving Target Defense and Mimic Defense to increase attacker cost through surface reconfiguration and redundancy. El Gadal and Ganti [40] proposed a reinforcement learning–based framework for dynamically selecting intrusion detection and mitigation techniques in SDN environments. Rehman et al. [41] introduced a proactive mechanism that combines OS diversity and cyber deception to deter attackers in IoT networks. Feng et al. [42] developed a deep reinforcement learning algorithm to adaptively allocate resources under adversarial conditions. To the best of our knowledge, no prior work has investigated dynamic defense for ML-based intrusion detection systems under adversarial attacks. Existing approaches, while effective in specific domains, often depend on predefined policies or simplified attacker assumptions, limiting their applicability in evolving adversarial settings. In contrast, we frame dynamic defense selection as a data-driven, attack-agnostic process that enables per-sample adaptation to both conventional cyber threats and adversarial manipulations.

F. Active Learning

Active learning (AL) aims to reduce labeling costs by selectively querying the most informative samples from a large unlabeled pool under a fixed budget [43]. This is especially valuable when annotation is costly, class distributions are imbalanced, or the data-generating process evolves, as it enables models to refine decision boundaries while covering rare or underrepresented cases. Commonly used AL strategies include uncertainty sampling (selecting points the model is least confident about), Query-by-Committee (selecting samples where multiple models disagree), density-weighted methods (prioritizing uncertain samples in dense regions of the data), and core-set or diversity-based selection (choosing batches that broadly cover the feature space). These strategies are model- and domain-agnostic and are frequently used as baselines for evaluating new active learning techniques [44].

AL has also been applied to domain-specific tasks. For example, Vaarandi and Guerra-Manzanares [45] use AL to prioritize NIDS alerts, reducing labeling costs. Alaa and Schenck [46] handle missing data by selecting samples based on imputation uncertainty. Fan et al. [47] combine AL with semi-supervised learning for fault diagnosis. Despite these advances, active learning has not been widely integrated with dynamic defense selection in ML-IDS, where the learner must choose among defenses per sample. Consequently, empirical evidence for this specific setting remains limited.

III. SAGE Framework

We design Sample-Aware Guarding Engine (SAGE), a dynamic defense selection framework designed to enhance the robustness of machine learning-based intrusion detection systems against adversarial attacks. Extending DYNAMITE [15], SAGE achieves both superior computational efficiency and enhanced robustness by leveraging targeted subsample selection and a multi-level, active learning–guided per-sample defense selection strategy. SAGE integrates several key components—adversarial sample generation, defense model training, and an active learning–guided dynamic defense selection algorithm—forming a comprehensive pipeline that evaluates and mitigates adversarial threats both effectively and efficiently.

As shown in Figure 2, the pipeline begins with data preprocessing, where raw data is cleaned, normalized, and encoded. The preprocessed data is then used to train a baseline DNN model along with multiple adversarial defense models. To simulate real-world scenarios, adversarial datasets are generated using diverse attack strategies, providing a comprehensive benchmark for evaluating framework effectiveness. Each defense model is subsequently assessed to determine its performance across different adversarial scenarios, identifying the most effective strategies for each case. To enable dynamic defense selection, a multi-level learner examines dataset patterns to predict the most effective defense for unseen adversarial

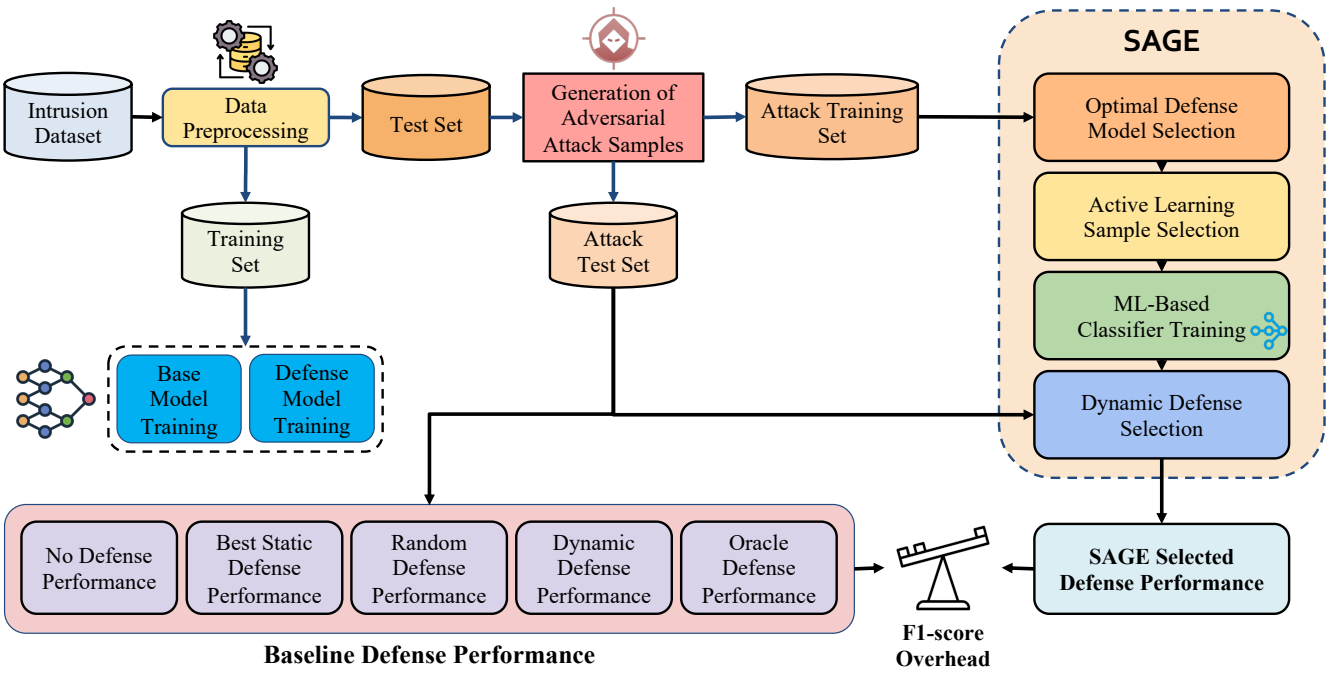

Fig. 2. Overview of the Sample-Aware Guarding Engine (SAGE) pipeline. Raw data is preprocessed and used to train a baseline ML-IDS model alongside multiple specialized defense models. Adversarial samples are generated across diverse attack strategies and intensities, and each defense model is evaluated to assign optimal defense labels. Entropic Open-set Active Learning (EOAL) identifies the most informative and diverse samples, which are then used to train a second-level learner that dynamically selects the most effective defense for each incoming adversarial sample, ensuring robust and efficient protection.

samples. This learner combines machine learning with active learning in a multi-layer framework to accurately select the optimal defense model for each input. By integrating optimal defense labels derived from the performance matrix—which evaluates the effectiveness of each defense model against various adversarial attacks—the SAGE dynamically selects the most suitable defense for each attack scenario. Finally, the SAGE's performance is compared to that of Oracle, the best static defense models (state-of-the-art defenses), random defense selection, and recommendation-based dynamic selection. In the following subsections, we describe each SAGE module in sequence, beginning with data preprocessing, followed by adversarial attack generation, defense training and best defense labeling, the dynamic defense selection algorithm, and concluding with the inference evaluation protocol.

A. Data Preprocessing

This module performs data cleaning to remove redundant or irrelevant features, standardization to normalize numerical features for consistent scaling, and categorical encoding to convert classification features into numerical representations suitable for ML models. After preprocessing, the data is split into training and test sets. The training set is used to train the baseline and defense models, while the test set is reserved for generating adversarial attacks and conducting final evaluations.

B. Generation of Adversarial Attack Samples

1) Selected Adversarial Attacks: We select nine state-of-the-art adversarial attacks (BIM [23], FGSM [22], PGD [7], DF [25], AutoPGD [24], ZOO [26], SINIFGSM [27], VNIFGSM [28], and CaFA [29]) to generate adversarial samples. Perturbation magnitude is controlled via the epsilon (ε) parameter, allowing tests under a range of attack intensities. These attacks with varying perturbation magnitudes provide a comprehensive benchmark for assessing the robustness of SAGE across diverse adversarial scenarios.

2) Adversarial Dataset Generation: The generation process involves applying each attack model to the dataset, with ε values adjusted to simulate varying levels of adversarial intensity. A unique adversarial dataset is generated for each combination of nine attack methods and four epsilon values $\{0.01, 0.1, 0.2, 0.3\}$, resulting in a total of 36 distinct datasets. Each attack is applied to the test dataset, maintaining the same sample size as the original. This ensures consistent evaluation while introducing adversarial perturbations based on attack type and intensity. After generating adversarial attack samples, we split them into two sets: *attack training* and *attack test*. The training portion is used to train our dynamic defense selection model, while the test portion is used for final evaluation.

C. Baseline Model Training

To establish a performance baseline, a Deep Neural Network (DNN) [48] is trained on the original, unperturbed dataset. The model is then evaluated under different adversarial attack configurations, providing a reference for assessing the effectiveness of defense strategies. This baseline serves as a crucial benchmark, illustrating the impact of adversarial attacks on model performance and emphasizing the importance of robust defense mechanisms and dynamic selection approaches.

D. Defense Model Training

We leverage ten state-of-the-art defenses against adversarial attacks: Projected Gradient Descent Adversarial Training [7], Interpolated Adversarial Training [31], Tradeoff-inspired Adversarial Defense via Surrogate-loss Minimization (TRADES) [8], Free Adversarial Training [32], Gaussian Augmenter [33], Defensive Distillation [9], Robust Soft Label Adversarial Distillation (RSLAD) [10], Feature Squeezing [11], Gaussian Noise [35], and Robust Generative Adaptation Network (RGAN) [34]. To address varying defense requirements, we introduce multiple parameter configurations for certain defense models. For RSLAD, configurations like RSLAD10 and RSLAD100 adjust optimization strength to evaluate robustness trade-offs. This approach systematically assesses adaptability to different adversarial perturbation levels. Applying these defense methods to diverse adversarial attacks enables the framework to evaluate model adaptability and performance across attack scenarios. These defenses form the basis of the dynamic selection mechanism, allowing the framework to deploy the most effective strategy for each adversarial sample.

E. Optimal Defense Identification

1) Constructing Attack Training and Attack Test Data: The attack training and attack test data are created using a subset of the 36 adversarial datasets—generated using different attack methods and epsilon values from the test set—ensuring a distinction between known and unknown data during model evaluation. Specifically, the datasets with an epsilon value of 0.1 (9 datasets) are used as attack training data, representing the known data. The remaining datasets, with remaining epsilon values (27 datasets), serve as attack test data, representing the unknown data. Beyond varying only the perturbation strength, we further evaluate robustness under a more challenging scenario: excluding certain adversarial attack types entirely from the training phase and introducing them only during testing. This extension allows us to assess how well our method generalizes to previously unseen adversarial threats, and we introduce this setup in detail in Section IV-D.

2) Optimal Defense Selection: To assess the defense models, we process attack training data through all ten defenses and record key metrics, such as the macro F1-score. This generates a "performance matrix", where each entry represents a defense model's effectiveness against a specific adversarial sample. The matrix serves as a basis for comparing defenses and identifying best strategies, offering insights into how each model addresses adversarial perturbations. To determine the most effective defense for each adversarial sample, we analyze the performance metrics of all ten defense models and select the highest-performing defense for each sample. This selected defense is then used as the label, which forms the ground truth for training our dynamic defense selection mechanism.

F. Dynamic Defense Selection Algorithm

To achieve robust and efficient defense allocation, SAGE incorporates a two-layer dynamic defense selection algorithm (Figure 3). In the first layer, active learning is employed

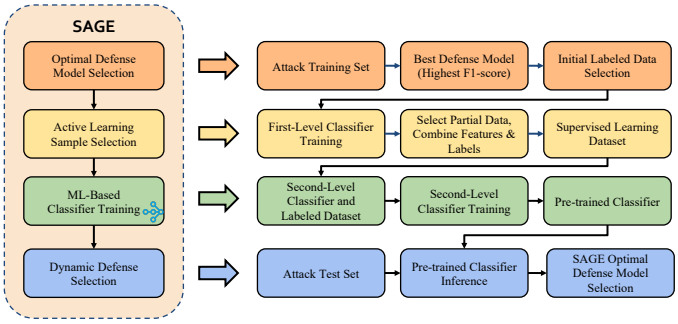

Fig. 3. Dynamic Defense Selection Algortihm

to identify the most informative samples while minimizing labeling costs. Specifically, we adopt Entropic Open-set Active Learning (EOAL) [16] to prioritize uncertain or underrepresented inputs, thereby improving the handling of previously unseen adversarial attacks. In the second layer, a lightweight ensemble-based classifier maps adversarial inputs to their corresponding optimal defenses, enabling efficient and accurate per-sample defense selection.

1) Active Learning Sample Selection (First Layer):

a) Motivation and Challenges: A central challenge in our dynamic defense selection approach is training the second-level learner to assign the most effective defenses to adversarial samples. Obtaining the optimal defense labels for this training is *computationally expensive*, as it requires evaluating each defense against multiple attack variants. Active learning offers a potential remedy by selecting and labeling the most informative samples; however, standard active learning methods, while effective at reducing labeling effort, typically assume a fixed input distribution [49]. This assumption breaks down in realistic deployments, where *an adversary can introduce previously unseen attacks by varying attack strength or creating entirely new attack types.* This limitation can lead to a dynamic defense policy that overfits to the training distribution and performs poorly on novel adversarial inputs, highlighting the need for a sample selection strategy that explicitly accounts for distributional shifts.

b) EOAL for Dynamic Defense Selection: To overcome the limitations of standard active learning, SAGE leverages Entropic Open-set Active Learning (EOAL) [16], which explicitly targets samples that are both uncertain and sparsely represented. In our context, these "unknowns" arise in two complementary ways: (i) *variation in attack strength*, where training covers only a single perturbation level of a given attack type but deployment may encounter a broader range, and (ii) *unseen attack types*, resulting from the exclusion of certain attack models during training. To handle (i), we include the same attacks at varying perturbation levels during evaluation, while for (ii) we introduce entirely new attack types at test time. EOAL addresses both types of "unknowns" by selecting high-uncertainty samples near decision boundaries affected by perturbation-strength shifts and within sparsely represented regions corresponding to missing attack models [16]. By prioritizing these inputs, the second-level learner develops a defense selection policy that improves generalization to previously

unseen attacks while minimizing labeling cost, supporting our framework's goals of scalable and robust ML-IDS defense against adversarial attacks.

c) EOAL Workflow: The EOAL process begins with the attack training set and their per-sample optimal defense labels derived from the "performance matrix". From this set, 10% of the data is selected as the initial labeled dataset using *stratified random sampling* (i.e., randomly sampling within each defense class) to ensure balanced representation of attack patterns. This step corresponds to the first-level classifier training in Fig. 3 (yellow boxes). We train a lightweight classifier (implemented as a random forest in our experiments) to provide a preliminary mapping from inputs to defense labels, which is used exclusively to guide the selection of unlabeled samples for querying. By combining EOAL with the first-level random forest classifier, we iteratively apply the active learning process to the remaining unlabeled pool over multiple acquisition rounds, gradually increasing the size of the labeled dataset in each round (see Section V-D for an ablation study). Specifically, we rank the unlabeled samples based on the entropy predicted by the random forest and select a diverse batch to minimize redundancy and enable efficient batch processing. Each selected batch is paired with its optimal-defense label to form supervised subsets of different sizes, which are then fed to the second-level learner.

d) EOAL Methodology and Mathematical Formulation: EOAL selects samples that provide the most useful information for training the second-level defense selector. For each unlabeled input x, EOAL computes an acquisition score that balances: (i) uncertainty with respect to known defense labels, and (ii) uncertainty arising from potentially unseen (open-set) patterns. Let $F(\cdot)$ denote the first-level feature extractor, and let the first-level classifier produce one-vs-rest posteriors $\{p_i(x)\}_{i=1}^{K}$ over the K defense labels.

▷ **Closed-set entropy:** This term measures how uncertain the model is about the known defense labels. High uncertainty indicates that the sample lies near the decision boundary of the current classifier. It is computed as the normalized average of one-vs-rest binary entropies:

$$S_c(x) = \frac{1}{K}\sum_{i=1}^{K} H(p_i), \quad p_i := p_i(x) \tag{1}$$

where $S_c(x) \in [0,1]$, $H(p_i)$ denotes the binary entropy function.

▷ **Distance-based (open-set) entropy:** To account for unseen or underrepresented attack patterns, EOAL estimates uncertainty based on the sample's proximity to clusters in feature space that represent open-set regions. Let $\{c_i\}_{i=1}^{K}$ be the cluster centers obtained from a working set of unlabeled samples. A soft assignment with temperature $T > 0$ gives

$$q_i(x) = \frac{\exp\big(-\|F(x) - c_i\|/T\big)}{\sum_{j=1}^{K}\exp\big(-\|F(x) - c_j\|/T\big)}, \tag{2}$$

$$S_d(x) = -\frac{1}{\log K}\sum_{i=1}^{K} q_i(x)\,\log q_i(x), \tag{3}$$

where $S_d(x) \in [0,1]$ measures how scattered the sample is across open-set regions. High values indicate the sample is near sparsely represented regions, which could correspond to unseen attacks.

▷ **Acquisition score:** The final score combines closed-set and open-set uncertainties:

$$S(x) = S_c(x) - S_d(x) \tag{4}$$

Samples with the largest $S(x)$ (high closed-set uncertainty but low open-set proximity) are selected for labeling. This ensures that EOAL focuses on inputs that are informative for the second-level learner, avoiding hard-but-uninformative outliers. In our framework, the first-level classifier outputs $p_i(x)$ and features $F(x)$. We implement clustering with k-means on $F(x)$ to obtain K centers $\{c_i\}_{i=1}^{K}$. EOAL operates in batches; to reduce redundancy, we apply a *farthest-first* diversity filter in feature space: starting from the top-scoring candidate, we iteratively add the sample that maximizes its minimum distance to the current batch until the batch quota is reached. This ensures that selected queries are informative.

2) ML-Based Classifier Training (Second Layer): The second layer of our dynamic defense selection algorithm trains an ML-based classifier to assign the most suitable defense model to each incoming sample, ensuring robust performance across varying attack conditions. We adopt models that balance resilience to heterogeneous attack patterns with the capacity to model nonlinear feature–defense relationships. Their computational efficiency enables low-latency per-sample inference, making them well-suited for dynamic defense selection.

During training, adversarial samples selected through EOAL are paired with their corresponding optimal defense labels to train the second-level learner. This process enables the model to learn the mapping between input features and effective defense strategies, thereby identifying patterns that inform per-sample defense allocation. At inference time, the trained second-level learner predicts the optimal defense for each test sample, including adversarial strengths/attacks not encountered during training. As a result, the framework dynamically selects defenses on a per-sample basis, generalizing to novel adversarial inputs and adapting to varying attack types and intensities. By moving beyond static defense mechanisms, this dynamic approach enhances resilience against previously unseen threats while maintaining computational efficiency, ensuring robust and practical performance.

G. Final Evaluation Protocol

During the final evaluation phase, each test sample is assigned to a corresponding defense model, ensuring that the selected strategy effectively mitigates the adversarial attack. To comprehensively assess SAGE's effectiveness, we compute the Macro F1-Score for each selected defense model, offering a holistic measure of overall performance. This final metric is then compared with other baseline approaches, such as Oracle, random defense, the best static defense (state-of-the-art), and recommendation-based dynamic defense, highlighting SAGE's robustness across diverse adversarial attack scenarios.

TABLE I
SELECTED INTRUSION DATASETS

Dataset	Year	Number of Features	Number of Attacks	Number of Samples
WUSTL-IIOT [50]	2021	41	6	1M
UNSW-NB15 [51]	2015	43	9	278K
X-IIoTID [48]	2021	68	18	800K
Edge-IIoTID [52]	2022	61	14	2.2M

IV. EXPERIMENTAL ANALYSIS

A. Selected Datasets

Table I summarizes the four intrusion datasets used in our paper. **WUSTL-IIoT** [50] captures IIoT testbed traffic across realistic attack scenarios (41 features; 1M samples). **UNSW-NB15** [51] mixes real and synthetic flows covering nine attack types (43 features; 278K samples). **X-IIoTID** [48] is a device- and connectivity-agnostic IIoT corpus with multi-view features from network, host, logs, and alerts (68 features; 800K instances). **Edge-IIoTset** [52] spans cloud/fog/edge IoT traffic with 14 attack types (61 features; 2.2M instances).

B. Baselines

To evaluate the effectiveness of SAGE, we compare its performance against multiple baseline approaches, enabling a systematic assessment of the benefits of dynamic defense selection. Each baseline serves as a reference point for quantifying SAGE's improvements in robustness and efficiency under diverse adversarial attack scenarios.

No Defense: This baseline evaluates the performance of a standard Deep Neural Network (DNN) model under adversarial attacks without any defense. It helps us establish the lower performance bound, emphasizing the vulnerability of unprotected models to adversarial perturbations.

Random Defense: This baseline selects a defense for each test sample uniformly at random from the candidate portfolio, without any knowledge of the attack. To control for randomness, the evaluation is repeated over 100 independent runs, and the reported results correspond to the mean macro-F1 across runs, providing a stable estimate of expected performance while keeping computation tractable.

Best Static (State-of-the-Art) Defense: This baseline identifies the most effective defense model by evaluating all candidates on the attack training data. The selected model achieves the highest average performance across all adversarial attack types and is then applied to the attack test set. As a carefully tuned but non-adaptive approach, it represents the state-of-the-art ML-IDS defense. While it provides the optimal static configuration achievable via train-time selection, it cannot adapt to variations in test-time attacks, highlighting the limitations of non-dynamic strategies. As our best static defense baseline, we select, for each dataset, the most effective static defense identified through our own analysis. Specifically, we use *PGD-Adversarial Training (PGD-AT)* for UNSW-NB15 [7], *TRADES* for WUSTL-IIoT [8], and the *Gaussian Augmenter (GA)* for both X-IIoTID and Edge-IIoTset [33].

Recommendation-Based Dynamic Defense: The recommendation defense baseline uses a recommendation system to select the most appropriate defense model for each adversarial sample. This system measures the similarity between attack patterns and the training data using the Manhattan distance and identifies the optimal defense model based on this measure. The selected defense is then applied to the adversarial sample, providing a semi-informed approach to defense selection.

Oracle Defense: The Oracle represents the *theoretical upper bound* of defense performance, achieved by selecting, for each test sample, the defense that performs best in hindsight. Such per-sample selection is infeasible in practice, as it would require evaluating all defenses at inference time. Comparing SAGE's performance to the Oracle demonstrates how closely SAGE approximates this ideal, which is computationally infeasible for real-world applications.

C. Experimental Setup

Hardware. We conduct our experiments on a Linux virtual machine server equipped with a 16-core CPU, 32 GB of RAM, and an NVIDIA A100 GPU with 80 GB of memory.

Evaluation Metric. We select the Macro F1 score as our evaluation metric because it offers a balanced assessment of model performance across all classes, independent of class distribution. This metric is especially pertinent for datasets with imbalanced attack types, as it ensures that minority classes are appropriately represented in the evaluation.

SAGE Performance Scoring. The final defense performance is evaluated using a weighted scoring formula, which combines the number of samples handled by each defense model and its performance:

$$\text{Score} = \sum_{i=1}^{N} \left(\frac{\text{Sample Count}_i}{\text{Total Samples}} \times \text{Model Performance}_i \right) \quad (5)$$

where Sample Count_i represents the number of adversarial samples assigned to the i-th defense model, and $\text{Model Performance}_i$ denotes the Macro F1 score of the i-th defense model. This formula calculates a weighted average of the defense models' performances, with the weight determined by the proportion of samples managed by each model. By doing so, it provides a holistic view of how well the dynamic algorithm selection performs across all assigned samples. A higher score reflects both the framework's ability to assign the most suitable defense models and the overall effectiveness of those models in mitigating adversarial impacts.

D. Dynamic Defense Selection Setup

Model Selection: We evaluate two ML models—XGBoost (XGB) and Random Forest (RF)—for the second-level learner. These models are chosen for their ability to effectively map attack patterns to optimal defense strategies. Each model is trained to predict the most suitable defense model for a given adversarial sample, and their performance is compared using the macro F1-score to identify the most effective approach for defense selection in IoT and IIoT environments.

Unknown Adversarial Attacks Setup: We evaluate defense selection under two types of "unknowns" adversarial conditions: (i) variations in attack strength, and (ii) previously unseen attacks (refer to Sec. III-F for details).

a) Variations in Attack Strength: To probe generalization under intensity changes, we adopt an ε-shift protocol in which the performance matrix and optimal-defense labels are constructed on adversarial samples generated at a fixed training strength (e.g., ε=0.1), while evaluation is conducted at unseen strengths ($\varepsilon \in \{0.01, 0.2, 0.3\}$). All other components—data preprocessing, the defense portfolio, and EOAL-driven acquisition for the selector—remain unchanged. Macro-F1 is computed per attack model together with the cross-attack 'Average', and results are compared against Oracle, best static, dynamic, random, and recommendation-based baselines.

b) Previously Unseen Attacks: To emulate partial coverage of the threat landscape, we further consider training regimes in which one or more attack models are withheld: (i) exclude CaFA; (ii) exclude CaFA and AutoPGD; (iii) exclude CaFA, AutoPGD, and DF. The performance matrix and optimal-defense labels are rebuilt using only the remaining attacks. EOAL acquires labels from this reduced pool, and the selector is retrained accordingly.

Active Learning: We restrict the proportion of labeled data used to train the second-level learner to a maximum of 50% of the available adversarial training pool across all datasets. In practice, we examine an extremely low-label setting with only 1% labeled data. For the ablation study, we systematically vary the cumulative label ratio at 1%, 10%, 20%, and 50%. Empirically, the best-performing label ratio (highest macro-F1) is dataset-specific: WUSTL-IIoT at 1%, UNSW-NB15 at 50%, Edge-IIoTest at 10%, and X-IIoTID at 50%.

V. Results

A. SAGE *Defense Performance*

Table II presents a comparative analysis of SAGE and the selected baselines (Oracle, dynamic, best static, random, and no defense) under the scenario where the perturbation amount is varied for the same adversarial attack types, i.e., *variations in attack strength*. The table reports the macro F1 score for each adversarial attack, as well as the average score across all perturbation levels and attack scenarios. It is evident that SAGE substantially outperforms the dynamic, best static, random, and no defense baselines, achieving performance nearly equivalent to that of the Oracle. Table III presents both the maximum and average performance improvements of SAGE compared to these baselines. These results highlight SAGE's effectiveness in maintaining robustness as the strength of adversarial perturbations changes.

Comparison with Random and Best Static Defenses: SAGE exhibits the most significant improvement in performance for the DF attack among all considered adversarial attacks. It achieves substantial improvements over both random and best static defenses, with performance gains of 199.4% and 68.8% on the UNSW-NB15 dataset, 1097.3% and 1800.1% on the WUSTL-IIoT dataset, 80.2% and 16.8% on the X-IIoTID dataset, and 173.3% and 293.4% on the Edge-IIoTest dataset, respectively, for the DF attack. This demonstrates that, especially in the case of stronger attacks, SAGE substantially outperforms random and best static, highlighting its effectiveness and reliability in optimizing defense model allocation. As shown in Table II, static defenses are more vulnerable to white-box attacks, as attackers can exploit their weaknesses, whereas in black-box attacks like ZOO, static defenses maintain relatively stable performance. However, SAGE still achieves a notable performance gain, improving 'Average' up to 12% over the best static defense. Our method also shows average improvements over both baselines, highlighting its effectiveness and adaptability in dynamically assigning optimal defense strategies.

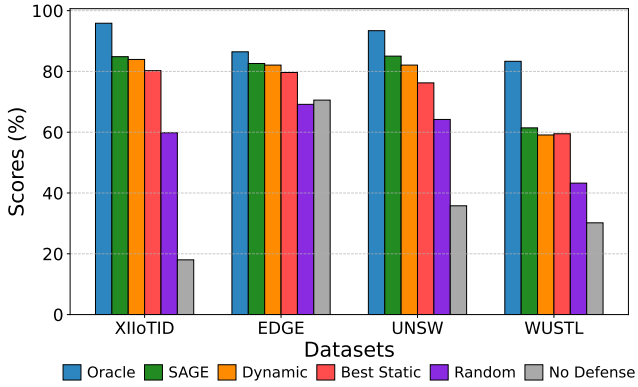

Fig. 4. Average performance (Macro-F1) across the five baselines

Comparison with Dynamic Defense: When compared to the recommendation-based dynamic defense strategy, Table II highlights SAGE's superior robustness and adaptability across a variety of adversarial scenarios and datasets. For instance, SAGE achieves up to 41.3% improvement in macro F1 scores, as seen in DF (40.47% vs. 28.64%) on WUSTL-IIoT. Unlike dynamic defenses, which rely on distance to adjust model parameters, SAGE employs an adaptive mechanism that dynamically mitigates perturbations without dependence on distance-based adjustments. This approach enables SAGE to avoid overfitting to specific attack patterns, enhancing its resilience against a wide range of adversarial threats. Another key aspect of SAGE's superiority is its balanced performance across both white-box and black-box attacks. Under the black-box ZOO attack, SAGE maintains high F1 scores (e.g., 95.34% on UNSW-NB15 and 98.14% on X-IIoTID), closely matching or exceeding dynamic defenses (94.81% and 97.83%, respectively). These results highlight that SAGE achieves a superior trade-off, significantly enhancing robustness in the challenging white-box setting while remaining competitive against black-box threats. As a result, as shown in Table III, SAGE's average F1 scores reflect improvements of up to 5.0% over dynamic defenses across all attack scenarios.

Comparison with Oracle: SAGE demonstrates a remarkably small performance gap with the Oracle, underscoring its effectiveness in dynamically selecting defenses across diverse adversarial conditions. In black-box attacks like ZOO, where it achieves 95.34% vs. 98.95% on UNSW-NB15, 97.50% vs. 99.51% on WUSTL-IIoT, and 98.14% vs. 98.35% on X-IIoTID, demonstrating near-Oracle performance without requiring prior knowledge of attack mechanisms. Even in the most challenging case, the DF attack, SAGE's 97.53% closely rivals the Oracle's 99.16%, suggesting that its adaptive mod-

TABLE II
FINAL PERFORMANCE (MACRO F1-SCORE) COMPARISON

(%)	Oracle	SAGE	Dynamic	Best Static	Random	No Defense	Oracle	SAGE	Dynamic	Best Static	Random	No Defense
	UNSW-NB15						WUSTL-IIoT					
Clean Data	99.12	94.79	96.82	88.65	89.37	90.70	99.51	98.84	98.84	94.81	89.08	94.24
BIM	95.09	84.15	76.59	74.06	65.78	30.78	89.25	60.85	61.64	67.17	46.61	28.89
FGSM	95.24	86.55	86.32	83.20	67.78	43.20	85.06	62.57	57.09	68.68	46.62	28.83
PGD	95.09	84.15	76.59	74.06	65.78	30.78	89.25	60.85	61.64	67.17	46.61	28.89
DF	71.76	62.40	59.76	36.97	20.84	10.81	64.13	40.47	28.64	2.13	3.38	1.41
AutoPGD	96.32	87.66	80.22	75.99	67.36	29.52	93.88	66.00	64.20	65.31	46.27	31.01
ZOO	98.95	95.34	94.81	88.64	88.55	83.51	99.51	97.50	99.59	94.81	88.88	90.47
CaFA	98.71	95.28	94.04	87.66	67.08	28.71	43.32	32.23	28.37	20.85	16.68	3.07
SINIFGSM	94.23	83.97	82.51	80.55	65.56	31.27	90.22	61.49	71.04	67.63	44.43	30.45
VNIFGSM	95.19	85.80	87.85	84.89	69.04	33.56	95.34	70.87	74.26	81.67	49.81	28.68
Average	93.40	**85.03**	82.08	76.23	64.20	35.79	83.33	**61.43**	60.72	59.49	43.25	30.19
	X-IIoTID						Edge-IIoTest					
Clean Data	98.71	98.17	97.88	93.11	96.98	98.40	100.00	97.80	100.00	94.45	91.24	94.41
BIM	95.91	80.48	82.20	77.77	55.59	10.80	98.87	94.80	91.80	94.18	79.34	84.29
FGSM	94.81	78.65	71.96	75.87	54.20	22.77	99.17	93.47	93.28	94.40	80.62	85.55
PGD	95.91	80.48	82.20	77.77	55.59	10.80	98.87	94.80	91.80	94.18	79.34	84.29
DF	99.16	97.53	97.74	89.47	54.14	1.79	10.44	5.90	5.61	1.50	2.16	1.54
AutoPGD	95.87	81.41	83.80	79.11	56.69	10.22	99.79	97.94	97.30	94.30	85.30	84.11
ZOO	98.35	98.14	97.83	93.11	94.79	82.53	100.00	98.59	100.00	94.56	91.29	94.41
CaFA	89.33	71.34	72.24	68.58	54.29	1.51	72.61	70.51	72.36	55.20	41.57	29.97
SINIFGSM	97.06	90.15	82.43	77.19	57.28	12.73	98.90	92.70	92.19	94.20	79.59	84.29
VNIFGSM	96.29	85.46	85.08	83.34	55.72	8.93	99.38	94.80	94.41	94.32	83.31	86.59
Average	95.85	**84.85**	83.94	80.25	59.81	18.01	86.45	**82.61**	82.08	79.65	69.17	70.56

TABLE III
SAGE F1-SCORE IMPROVEMENT RATE

Improvement Rate (%)		Dynamic	Best Static	Random
UNSW-NB15	Max	9.9	68.8	**199.4**
	Average	3.9	15.2	**46.1**
WUSTL-IIoT	Max	41.3	**1800.1**	1097.3
	Average	5.0	**201.0**	157.7
X-IIoTID	Max	9.4	16.8	**80.2**
	Average	1.2	5.7	**44.9**
Edge-IIoTest	Max	5.2	**293.4**	173.3
	Average	1.1	36.5	**39.0**

eling enables it to effectively resist the smallest perturbation attacks. These results highlight SAGE's ability to allocate defenses effectively without requiring exhaustive model evaluations, making it both efficient and overhead. On average, the performance difference remains minimal, with a 3.84% gap in Edge-IIoTest and 8.1% in UNSW-NB15, reinforcing SAGE's adaptability even in more complex adversarial scenarios.

Clean Data Performance: Table II presents a comparison of clean data performance for SAGE against baseline methods. The results are evaluated using macro F1 scores under clean (no adversarial attack) conditions. Specifically, SAGE achieves F1 scores of 94.79% on UNSW-NB15, 98.84% on WUSTL-IIoT, 98.17% on X-IIoTID, and 97.80% on Edge-IIoTest, closely approaching the Oracle's near-optimal scores of 99.12%, 99.51%, 98.71%, and 100.00%, respectively. Compared to Dynamic and No Defense baselines, SAGE consistently delivers superior or comparable performance, while significantly outperforming Random and Best Static in most cases. As illustrated in Table II (Clean Data), SAGE maintains top-tier performance on clean data, ranking at the first or second level relative to all baselines beside the Oracle. These results highlight that SAGE achieves enhanced robustness without compromising performance under benign conditions, underscoring its effectiveness across diverse datasets.

Insights: The results reflect SAGE's ability to approach the theoretical upper bound established by the Oracle while demonstrating its flexibility in handling diverse adversarial scenarios. Since the Oracle determines the best performance by testing every dataset across all models, achieving this level of optimality in practice would require significant effort and resources, making it impractical for real-world applications. In contrast, the SAGE dynamically allocates defense strategies using a machine learning-based approach. This method eliminates the need for manually selecting the optimal defense for each attack, showcasing the SAGE's adaptability in addressing complex adversarial scenarios. Moreover, SAGE's effectiveness lies in its robust generalization across both white-box and black-box attack scenarios, a flexibility that the Oracle's idealized approach cannot replicate in practical settings. For instance, SAGE's performance in complex attacks like CaFA and DF, highlights its ability to mitigate sophisticated perturbations effectively, even without the Oracle's perfect knowledge. This adaptability stems from SAGE's integrated modeling of spatial and temporal dependencies, which allows it to capture nuanced patterns in data distributions, unlike the Oracle's reliance on exhaustive testing.

TABLE IV
UNSEEN ADVERSARIAL ATTACK PERFORMANCE (MACRO F1, AVERAGES)

Average (%)	UNSW-NB15	WUSTL-IIoT	X-IIoTID	Edge-IIoTset
No Exclusion	**85.03**	61.43	**84.85**	**82.61**
Exclude CaFA	84.08	57.26	83.74	79.72
Exclude CaFA&AutoPGD	83.99	**61.71**	83.25	77.98
Exclude CaFA&AutoPGD&DF	79.50	57.78	80.85	81.40

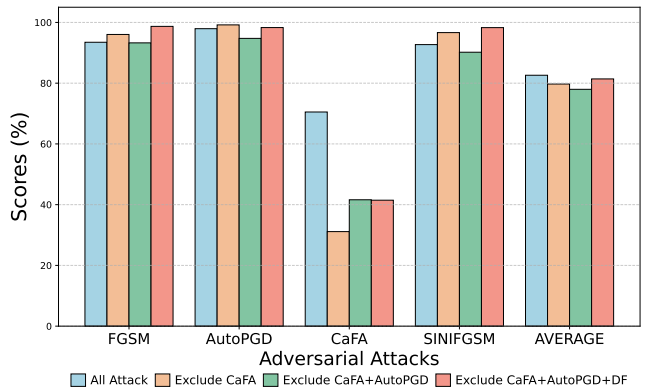

Fig. 5. Edge-IIoTest Unseen Adversarial Attack Performance Comparison

B. Unseen Adversarial Attacks Performance

To evaluate robustness to unknown attack models, i.e., *previously unseen attack types*, we remove one or more attack models from the training set and then evaluate on the full suite, as detailed in Sec. IV-D. Table IV reports 'Average' macro-F1. The four evaluation conditions are: no exclusion (all attacks included); exclude CaFA; exclude CaFA and AutoPGD; exclude CaFA, AutoPGD, and DF.

First, we analyze individual attack performances on a per-attack basis. Withholding a given attack model primarily affects the withheld model while leaving other models largely unaffected. For instance, on WUSTL-IIoT, DF decreases from 36.24% (when DF is included in training) to 0.99% when DF is withheld, corresponding to a drop of 35.25 points; on Edge-IIoTset, CaFA declines from 70.51% to 31.13% when CaFA is withheld, a reduction of 39.38 points. By contrast, ZOO remains near ceiling across conditions, with performance ranging from 97% to 99% on all four datasets. This pattern suggests that the selector establishes distinct decision regions for each attack model: removing a model leaves its decision region under-represented without causing systemic degradation in the performance of other models.

Second, despite such targeted drops, the system-level 'Average' remains comparatively stable across exclusion regimes, evidencing cross-attack generalization. From 'No Exclusion' to 'Exclude CaFA+AutoPGD+DF', the average decreases by only 5.53 points on UNSW-NB15 (85.03% to 79.50%), 3.65 points on WUSTL-IIoT (61.43% to 57.78%), 4.00 points on X-IIoTID (84.85% to 80.85%), and 1.21 points on Edge-IIoTest (82.61% to 81.40%). Clean traffic also stays high across conditions (e.g., UNSW 94.79–96.03%, Edge 97.68–98.69%), reinforcing that robustness is preserved at the system level even when specific attacks are unknown at training time.

Overall, the SAGE exhibits localized sensitivity (strong drops on the held-out model) and global stability (modest change in the cross-attack average), which together characterize reliable defense selection under distribution shift.

C. Overhead Analysis

The overhead analysis evaluates the computational efficiency of SAGE by comparing the per-sample defense selection time against Oracle, Dynamic, and Best-static baselines. Table V reports the processing time in ms/sample, illustrating the substantial efficiency advantage of SAGE over the competing methods. SAGE maintains low inference overhead, requiring less than 1.01 ms per sample across datasets. By contrast, the Oracle requires approximately 24–27 ms per sample, making SAGE up to $29\times$ faster. The efficiency gain comes from executing a single learned selector and dispatching one defense per input, rather than exhaustively evaluating the entire defense portfolio as the Oracle does. These dramatic reductions highlight SAGE's active learning and machine learning-driven adaptive mechanism that enables rapid defense selection without the resource-intensive enumeration required by the Oracle, making it far more practical for real-world applications where computational efficiency is critical.

Compared to the Dynamic baseline, which leverages distance for real-time parameter adjustments, SAGE exhibits slightly higher processing times, with a gap of up to 0.21 ms/sample across the datasets. However, this marginal increase is offset by SAGE's superior robustness, as demonstrated in prior sections, particularly against gradient-based attacks (e.g., up to 41% F1 score improvement for one attack and up to 5% on average). The Dynamic approach, while computationally efficient, often sacrifices generalization due to its reliance on real-time distance calculations, which can lead to suboptimal defense strategies in complex scenarios.

Against the Best Static baseline, SAGE offers significant efficiency gains, achieving speedups of $2.48\times$ (0.89 vs. 2.21 ms/sample) on UNSW-NB15, $2.64\times$ (0.92 vs. 2.43 ms/sample) on WUSTL-IIoT, $2.51\times$ (0.85 vs. 2.13 ms/sample) on X-IIoTID, and $2.40\times$ (1.01 vs. 2.42 ms/sample) on Edge-IIoTest. While Best Static is more efficient than the Oracle, it still lacks the adaptability of SAGE, underscoring SAGE's efficiency in accelerating defense selection while maintaining strong performance.

TABLE V
PROCESSING TIME (MS/SAMPLE) COMPARISON

ms/Sample	UNSW-NB15	WUSTL-IIoT	X-IIoTID	Edge-IIoTest
SAGE	0.89	0.92	0.85	1.01
Oracle	24.28	26.70	23.44	26.58
Dynamic	**0.68**	**0.69**	**0.63**	**0.72**
Best Static	2.21	2.43	2.13	2.42

D. Ablation Study

1) Active Learning Methods: We evaluate several active learning (AL) sampling strategies for their effect on the robustness of the second-level defense selector, including conventional uncertainty sampling, density weight sampling, batch mode sampling [44], and Entropic Open-set Active Learning (EOAL). As shown in Fig. 6, across representative white-box gradient attacks, e.g., FGSM, AutoPGD, and composite attacks, e.g., CaFA, EOAL attains the best or tied-best macro-F1 on most datasets and exhibits reduced variance across attacks, indicating superior stability and generalization. Three empirical insights emerge from our evaluation. First,

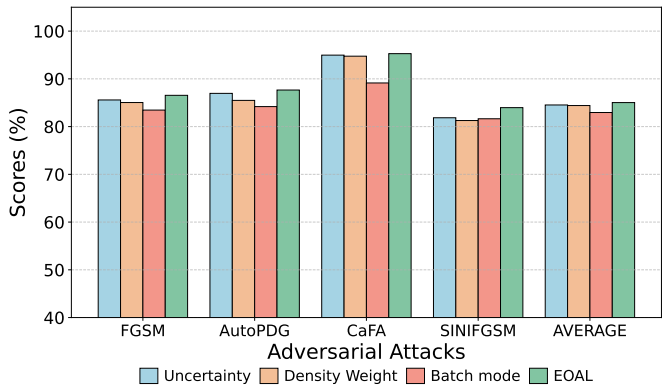

Fig. 6. Active Learning Performance Comparison (UNSW-NB15)

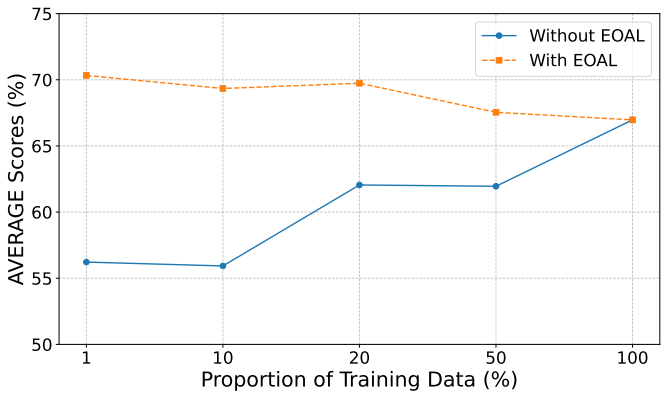

Fig. 7. Different Proportion of Training Data Comparison (WUSTL-IIoT)

EOAL uses entropy-based acquisition with explicit open-set modeling to consistently query the most informative samples, focusing annotation on decision regions associated with novel or rare attack behaviors and avoiding adversarial hard-but-uninformative cases. Second, its batch-quota acquisition under comparable labeling budgets (50%, 20%, 10%, 1%) preserves diversity and limits redundancy, yielding higher information per label and pushing the selector toward near full-data performance at substantially lower annotation rates. Third, under our ε-shifted protocol—training with $\varepsilon = 0.1$ and testing on unseen $\varepsilon \in \{0.01, 0.2, 0.3\}$—EOAL maintains a stable mapping from inputs to optimal defenses, demonstrating robustness to distribution shift. Together with the comparative results, these insights justify our adoption of EOAL as the active learning component of SAGE.

2) Different Proportion of Training Data: We examine SAGE's label-efficiency under EOAL by varying the proportion of labeled attack training data. As shown in Fig. 7, macro-F1 increases with larger budgets, yet under EOAL the curve is steep at low label fractions: strong performance already emerges at 1% of the pool, approaching the performance achieved with substantially larger budgets (e.g., 10%–50%). By explicitly targeting sparsely covered regions of the attack space, EOAL can match—or even exceed—full-label training with only 1% supervision, delivering gains of up to 3.4 points. By contrast, without EOAL, using stratified random sampling, performance improves more slowly and requires substantially larger labeled fractions to reach comparable results. This advantage stems from EOAL's entropy-ranked, diversity-aware acquisition, which concentrates annotation on the most informative regions and reduces redundancy; as a result, the second-level learner approaches its full-data quality with a small fraction of labels, cutting annotation cost and training time while preserving robustness.

VI. Conclusion

In high-stakes operational networks, ML-based intrusion detection systems (ML-IDS) must contend with evolving attack geometries and distribution shifts. To address this, our framework, SAGE, formulates adversarially robust ML-IDS as a per-sample defense selection problem. It couples a curated defense portfolio with an EOAL-trained selector, enabling the system to choose the most effective defense for each input at inference while maintaining label- and compute-efficiency and preserving accuracy on clean data. SAGE achieves near-Oracle performance, attaining a best-case macro-F1 within 3.8% of the Oracle while accelerating per-sample computation by up to $29\times$. It delivers the largest gains over baselines, with up to a 5.0% improvement relative to a recommendation-based dynamic baseline, and remains robust under unseen attacks, with the cross-attack average decreasing by only 1.21 points.

Acknowledgments

This work has been funded in part by NSF, with award numbers #1826967, #1911095, #2003279, #2052809, #2100237, #2112167, #2112665, and in part by PRISM and CoCoSys, centers in JUMP 2.0, an SRC program sponsored by DARPA.

References

[1] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. De Alvarenga, "A survey of intrusion detection in internet of things," *Journal of Network and Computer Applications*, vol. 84, pp. 25–37, 2017.

[2] O. I. Abiodun, E. O. Abiodun, M. Alawida, R. S. Alkhawaldeh, and H. Arshad, "A review on the security of the internet of things: Challenges and solutions," *Wireless Personal Communications*, vol. 119, pp. 2603–2637, 2021.

[3] K. A. Da Costa, J. P. Papa, C. O. Lisboa, R. Munoz, and V. H. C. de Albuquerque, "Internet of things: A survey on machine learning-based intrusion detection approaches," *Computer Networks*, vol. 151, pp. 147–157, 2019.

[4] O. Gungor, E. Li, Z. Shang, Y. Guo, J. Chen, J. Davis, and T. Rosing, "Rigorous evaluation of machine learning-based intrusion detection against adversarial attacks," in *2024 IEEE International Conference on Cyber Security and Resilience (CSR)*. IEEE, 2024, pp. 152–158.

[5] N. Mishra and S. Pandya, "Internet of things applications, security challenges, attacks, intrusion detection, and future visions: A systematic review," *IEEE Access*, vol. 9, pp. 59 353–59 377, 2021.

[6] A. Alotaibi and M. A. Rassam, "Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense," *Future Internet*, vol. 15, no. 2, p. 62, 2023.

[7] A. Madry, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[8] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.

[9] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[10] B. Zi, S. Zhao, X. Ma, and Y.-G. Jiang, "Revisiting adversarial robustness distillation: Robust soft labels make student better," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 443–16 452.

[11] W. Xu, "Feature squeezing: Detecting adversarial exa mples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[12] I. Debicha, B. Cochez, T. Kenaza, T. Debatty, J.-M. Dricot, and W. Mees, "Adv-bot: Realistic adversarial botnet attacks against network intrusion detection systems," *Computers & Security*, vol. 129, p. 103176, 2023.

[13] I. Debicha, R. Bauwens, T. Debatty, J.-M. Dricot, T. Kenaza, and W. Mees, "Tad: Transfer learning-based multi-adversarial detection of evasion attacks against network intrusion detection systems," *Future Generation Computer Systems*, vol. 138, pp. 185–197, 2023.

[14] C. Wang, J. Wang, and Q. Lin, "Adversarial attacks and defenses in deep learning: A survey," in *Intelligent Computing Theories and Application: 17th International Conference, ICIC 2021, Shenzhen, China, August 12–15, 2021, Proceedings, Part I 17*. Springer, 2021, pp. 450–461.

[15] J. Chen, O. Gungor, Z. Shang, E. Li, and T. Rosing, "Dynamite: Dynamic defense selection for enhancing machine learning-based intrusion detection against adversarial attacks," in *2025 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2025, pp. 213–219.

[16] B. Safaei, V. Vibashan, C. M. De Melo, and V. M. Patel, "Entropic open-set active learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 5, 2024, pp. 4686–4694.

[17] H. Sebestyen, D. E. Popescu, and R. D. Zmaranda, "A literature review on security in the internet of things: Identifying and analysing critical categories," *Computers*, vol. 14, no. 2, 2025.

[18] O. Gungor, T. Rosing, and B. Aksanli, "Roldef: Robust layered defense for intrusion detection against adversarial attacks," in *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2024, pp. 1–6.

[19] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," *applied sciences*, vol. 9, no. 20, p. 4396, 2019.

[20] O. Gungor, T. Rosing, and B. Aksanli, "Stewart: Stacking ensemble for white-box adversarial attacks towards more resilient data-driven predictive maintenance," *Computers in Industry*, vol. 140, p. 103660, 2022.

[21] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[22] I. J. Goodfellow, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[23] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[24] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International conference on machine learning*. PMLR, 2020, pp. 2206–2216.

[25] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[26] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[27] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," 2020.

[28] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," 2021.

[29] M. Ben-Tov, D. Deutch, N. Frost, and M. Sharif, "Cafa: Cost-aware, feasible attacks with database constraints against neural tabular classifiers," 2025.

[30] S. Goyal, S. Doddapaneni, M. M. Khapra, and B. Ravindran, "A survey of adversarial defenses and robustness in nlp," *ACM Computing Surveys*, vol. 55, no. 14s, pp. 1–39, 2023.

[31] A. Lamb, V. Verma, J. Kannala, and Y. Bengio, "Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy," in *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, 2019, pp. 95–103.

[32] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!" *Advances in neural information processing systems*, vol. 32, 2019.

[33] V. Zantedeschi, M.-I. Nicolae, and A. Rawat, "Efficient defenses against adversarial attacks," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 39–49.

[34] Y. Li, L. Huang, S. Tian, H. Liu, and Z. Li, "Robust generative adaptation network for open-set adversarial defense," *IEEE Transactions on Information Forensics and Security*, 2025.

[35] S. A. Kassam, *Signal detection in non-Gaussian noise*. Springer Science & Business Media, 2012.

[36] D. Han, Z. Wang, Y. Zhong, W. Chen, J. Yang, S. Lu, X. Shi, and X. Yin, "Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2632–2647, 2021.

[37] Y. Alslman, M. Alkasassbeh, and M. J. Abdel-Rahman, "Breaking and healing: Gan-based adversarial attacks and post-adversarial recovery for 5g idss," *IEEE Access*, pp. 1–1, 2025.

[38] H. Holla, S. R. Polepalli, and A. A. Sasikumar, "Adversarial threats to cloud ids: Robust defense with adversarial training and feature selection," *IEEE Access*, vol. 13, pp. 84 992–85 003, 2025.

[39] Y. Zheng, Z. Li, X. Xu, and Q. Zhao, "Dynamic defenses in cyber security: Techniques, methods and challenges," *Digital Communications and Networks*, vol. 8, no. 4, pp. 422–435, 2022.

[40] W. El Gadal and S. Ganti, "Dynamic defense framework: A unified approach for intrusion detection and mitigation in sdn," in *2024 8th Cyber Security in Networking Conference (CSNet)*, 2024, pp. 22–27.

[41] Z. Rehman, I. Gondal, M. Ge, H. Dong, M. Gregory, and Z. Tari, "Proactive defense mechanism: Enhancing iot security through diversity-based moving target defense and cyber deception," *Computers & Security*, vol. 139, p. 103685, 2024.

[42] X. Feng, J. Han, R. Zhang, S. Xu, and H. Xia, "Security defense strategy algorithm for internet of things based on deep reinforcement learning," *High-Confidence Computing*, vol. 4, no. 1, p. 100167, 2024.

[43] A. Tharwat and W. Schenck, "A survey on active learning: State-of-the-art, practical challenges and research directions," *Mathematics*, vol. 11, no. 4, p. 820, 2023.

[44] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM computing surveys (CSUR)*, vol. 54, no. 9, pp. 1–40, 2021.

[45] R. Vaarandi and A. Guerra-Manzanares, "Network ids alert classification with active learning techniques," *Journal of Information Security and Applications*, vol. 81, p. 103687, 2024.

[46] A. Tharwat and W. Schenck, "Active learning for handling missing data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 3273–3287, 2025.

[47] C. Fan, Q. Wu, Y. Zhao, and L. Mo, "Integrating active learning and semi-supervised learning for improved data-driven hvac fault diagnosis performance," *Applied Energy*, vol. 356, p. 122356, 2024.

[48] M. Al-Hawawreh, E. Sitnikova, and N. Aboutorab, "X-iiotid: A connectivity-agnostic and device-agnostic intrusion data set for industrial internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 5, pp. 3962–3977, 2022.

[49] D. Cacciarelli and M. Kulahci, "Active learning for data streams: a survey," *Machine Learning*, vol. 113, no. 1, pp. 185–239, 2024.

[50] M. Zolanvari *et al.*, "Wustl-iiot-2021 dataset for iiot cybersecurity research," *Washington University in St. Louis, USA*, 2021.

[51] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6.

[52] M. A. Ferrag *et al.*, "Edge-iiotset: A new comprehensive realistic cyber security dataset of iot and iiot applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40 281–40 306, 2022.