aDisRAE: ADAPTIVE DISCRIMINATIVE REPRESENTATION AUTOENCODER FOR FEW-SHOT CYBERATTACK DETECTION

Manh Tuan Nguyen[1], *Le Dinh Trang Dang*[1], *Van Loi Cao*[1,*]

Abstract

Due to the scarcity of labeled anomalous data, few-shot learning has emerged as a critical paradigm for detecting novel and rare cyberattacks. The Discriminative Representation Autoencoder (DisRAE) framework learns a latent space where anomalies are pushed away from a central cluster of normal data, but it struggles with advanced attacks that closely mimic benign behavior. These subtle anomalies are often mapped too close to the normal cluster, leading to detection evasion. To address this limitation, this paper proposes the Adaptive Discriminative Representation Autoencoder (aDisRAE). The framework enhances the training objective by incorporating a prior outlier score that quantifies the subtlety of each anomaly. This score guides an adaptive repulsion mechanism, applying a stronger force to anomalies that most resemble normal data, ensuring a more effective separation in the latent space. The experiments evaluate aDisRAE on three public benchmark datasets: NSL-KDD, CIC-IDS2017 and UNSW-NB15. The results show a notable improvement, raising AUC by up to 10% and boosting robustness, especially against evasive attacks.

Index terms

Cyberattack detection; anomaly detection; few-shot learning; discriminative autoencoder.

1. Introduction

Cyberattacks pose an escalating threat to data privacy and information security, growing in sophistication over decades [1]. To combat these threats, a defense-in-depth security posture is often established, employing network monitoring systems such as Firewalls, Intrusion Detection and Prevention Systems and Security Information and Event Management solutions. These systems traditionally rely on signature-based detection and conventional machine learning approaches to identify malicious activities [2]. However, these methods struggle to keep pace with rapidly evolving cyberattacks. Signature-based systems depend on known attack patterns, requiring labor-intensive updates to incorporate

[1]Institute of Information and Communication Technology, Le Quy Don Technical University
[*]Corresponding author, email: loi.cao@lqdtu.edu.com DOI: 10.56651/lqdtu.jst.v14.n02.1108.ict

new threat signatures, which often lag behind the dynamic threat landscape. Similarly, traditional supervised machine learning demands large, balanced and labeled datasets to train effective classifiers — a requirement often unmet due to the scarcity of labeled anomaly data [3]. These limitations highlight the need for more adaptive and efficient detection methods to address novel and complex cyber threats.

To overcome the shortcomings of traditional approaches, advanced machine learning paradigms, including semi-supervised and few-shot learning, offer promising solutions for enhancing cyberattack detection. Semi-supervised approach builds detection models from uncompleted labeled network traffic, typically labeled normal data, reducing dependency on labeled anomaly samples [4]. In real-world scenarios, known labeled anomalous data and limited labeled samples of new cyberattacks can be obtained through human expertise or collaborative knowledge sharing. This makes the few-shot learning approach particularly effective for the scenario [5]. Inspired by human learning, few-shot learning enables models to generalize accurately from few labeled samples by leveraging prior knowledge during training and adapting swiftly to new tasks during testing [6], [7]. Unlike traditional supervised learning, few-shot learning excels in handling limited data, making it ideal for detecting novel cyberattacks with minimal prior examples.

The ability to develop effective cyberattack detection models from a limited set of anomalous examples is paramount for robust cybersecurity. A prominent strategy in this domain is representation learning, which seeks to learn a discriminative embedding space where normal and anomalous instances are clearly separable. The DisRAE framework [8], for example, exemplifies this approach. It functions as a powerful feature extractor, $f(x)$, that maps high-dimensional inputs into a structured latent space. Within this space, normal instances are engineered to cluster near the origin, while anomalies are projected to distant locations, thereby enabling effective detection. However, a key limitation of DisRAE and similar methods is their weakness against attacks that closely mimic normal behavior. These advanced anomalies have features that look very much like those of benign data. As a result, they are mapped into areas of the embedding space that are hard to distinguish from normal regions, allowing them to evade detection.

To address this limitation, the study introduces aDisRAE, integrating prior knowledge of the input data into DisRAE's learning process. This is achieved by introducing a prior outlier score, which quantifies the inherent anomalousness of each instance based on the original data distribution. This score can be derived through a distance-based metric or a machine learning model. Regardless of its origin, this pre-computed score is then integrated into a composite loss function to guide the main model's training. This approach encourages the model to achieve two goals simultaneously: to learn a discriminative representation and to create a more uniform distribution between the attack patterns and the benign traffic. Crucially, this mechanism applies a stronger repulsive force to anomalies that most closely resemble normal data, ensuring they are pushed further away in the embedding space than more obvious attacks. This targeted separation significantly enhances the model's ability to distinguish subtle attacks from normal behavior, thereby improving overall detection performance and robustness.

The rest of this paper is organized as follows: Sections 2 and 3 briefly discuss recent anomaly detection approaches and provide the background knowledge of few-shot/discriminative learning. Section 4 describes the proposed method with the outlier score mechanism. Section 5 and 6 present experiments and discussion results. Finally, Section 7 summarizes the findings and draws future directions.

2. Related work

Many studies have been designed to train anomaly detection models based on small-sample datasets. However, traditional supervised methods have failed to achieve high detection rates on such datasets. To address this issue, some studies have explored transfer learning to transfer knowledge learned from related source domains to target domains with limited data. Chen *et al.* [9] proposed a cross-domain intrusion detection model for imbalanced data by enhancing the information transmission link in adversarial domain adaptation. They introduced an information-enhanced adversarial DA (IADA) method. The domain adaptation approach can transfer cyberattack detection knowledge from the IoT domain to the network domain by utilizing common attack features shared across both domains [10]. Hashemi *et al.* [11] proposed an improved domain adaptation technique for traffic data, which incorporated pseudo labels for the target domain. This method can be leveraged to detect unseen anomalies in target network systems with limited labeled data.

Several studies have also advanced few-shot learning approaches for network anomaly detection. Yu *et al.* [12] proposed a metric-based approach that integrated a softmax function with center loss to address the few-shot problem in intrusion detection. However, their method assumes that the attack classes in the testing phase are also present in the training data, limiting its applicability to novel attacks. In contrast, Lu *et al.* [13] utilized the Model-Agnostic Meta-Learning (MAML) framework, converting numerical network data into images and optimizing model parameters to effectively handle limited training samples. Their approach demonstrates improved adaptability to new attack types.

Cao *et al.* [8] introduced a few-shot learning framework that combines a DisRAE with a classifier trained on representations of normal samples and a small set of labeled anomalies. This method leverages the autoencoder's ability to create a discriminative embedding space for enhanced detection. Similarly, Rustam *et al.* [14] developed a real-time methodology for network attack detection, achieving high accuracy using a meta-RF-GNB model that integrated random forests and Gaussian naive Bayes classifiers. Ye *et al.* [15] addressed data scarcity using Latent Dirichlet Allocation for data expansion and proposed a semantic-aware generative learning scheme to improve detection of contextually complex attacks.

He *et al.* [16] proposed an intrusion detection system combining generative adversarial networks with MAML to improve few-shot detection of rare attacks. Additionally, Matching Networks [17] and Prototypical Networks [18] employ distinct embedding functions and prototype-based representations to improve anomaly detection accuracy.

Ding *et al.* [19] introduced Graph Deviation Networks, leveraging graph-based few-shot learning for network anomaly detection, while Xu *et al.* [20] proposed a meta-learning based approach tailored for few-shot scenarios. Moon *et al.* [21] combined MAML with variational autoencoders to enhance time-series anomaly detection, demonstrating robustness in dynamic environments.

Collectively, these works illuminate few-shot learning's pivotal role in surmounting labeled data bottlenecks in cybersecurity, from metric refinements to meta-adaptive and generative enhancements. However, a common limitation remains: most methods apply the same regularization to all anomalies, failing to effectively handle the range of anomaly types—where obvious attacks are easily distinguished, but subtle ones that mimic normal traffic often go undetected. Our work addresses this gap by enhancing the DisRAE framework [8] with an adaptive outlier score mechanism that adjusts the separation of anomalies in the latent space based on their inherent distinctiveness, achieving more balanced and effective detection of diverse threats in few-shot scenarios.

3. Background

3.1. Few-shot cyberattack detection

Few-shot learning has attracted significant attention from researchers due to its potential to address new or rare attack types in cyber attack detection. The core idea of few-shot learning is to leverage knowledge from previous learning tasks to quickly adapt to a new task. Most approaches share a common two-phase structure. In the training phase, the model learns from prior tasks, aiming to acquire useful assumptions and a strong generalization ability. Then in the testing phase, a small number of labeled samples from new or rare attack types (support set) are provided to the trained model, enabling it to classify or predict these novel attacks within a query set. In the context of cyber attack detection, each task T can be seen as a binary classification problem. Let T_m represent a set of tasks derived from the training data, which involve classifying normal versus anomalous data. The anomalous instances may belong to multiple categories $C_1, C_2, \ldots, C_k$. During the testing phase, the framework constructs a new set of tasks T_n from the support set, which contains normal traffic and a few labeled samples from a new anomaly category C_{new}. The query set is then utilized to comprehensively evaluate the detection performance of the trained model on unseen samples.

In this study, the objective is to learn a robust feature representation from the training set during the training phase. The model is trained on a large number of known attack types to capture the most discriminative features that separate normal and malicious traffic. During the testing phase, the pre-trained feature extractor is used to process an upsampled version of the limited samples from new or rare attack classes. The resulting embeddings are then employed to train a binary classifier to distinguish between normal and anomalous traffic.

3.2. Discriminative Representation Autoencoder

AutoEncoder is well-known for anomaly detection in unsupervised setting. It aims to achieve good reconstruction for normal instances, thereby making the reconstruction error (RE) large for anomaly instances:

$$\mathcal{L}_{RE} = ||x - \hat{x}||^2 \tag{1}$$

In contrast, DisAE emerges as an innovation that leverages supervised learning to carve out distinct representation spaces for both normal and anomaly classes [22]. These approaches aim to push anomaly instances far away from the manifold while concurrently minimizing the RE of normal data. The objective function of DisAEs presented as:

$$\mathcal{L}(X^+ \cup X^-) = max(0, l(x) \times (||x - \hat{x}||^2 - m)) \tag{2}$$

where, X^+ and X^- symbolize the normal and anomaly classes respectively, with $l(x)$ assigned as 1 for normal data and -1 for anomaly instances and m is the margin. For the scenario, equation 2 aims to minimize RE to belong within the range $(0, m)$ for normal instances, while accentuating RE greater than m for anomaly instances.

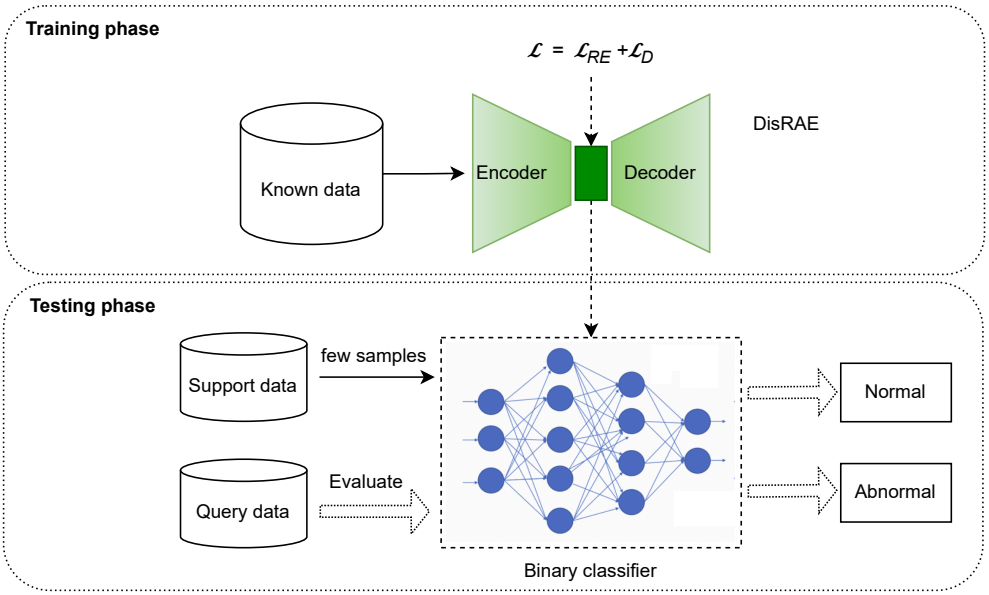


Fig. 1. The DisRAE Framework for cyberattack detection.

DisAE relies on a margin m to define the separation in RE scores between normal and anomalous data, without directly enforcing separation in the latent space. In contrast, Cao *et al.* [8] introduced DisRAE, which incorporates a novel regularized loss into the standard autoencoder framework. The total loss function is expressed as follows:

$$\mathcal{L} = \mathcal{L}_{RE} + \mathcal{L}_D = ||x - \hat{x}||^2 + |z(x)|^{l(x)} \tag{3}$$

where, $\mathcal{L}_D$ is the discriminative loss and $z(x)$ is the latent vector. The term $|z(x)|^{l(x)}$ promotes latent space separation by enlarging $|z(x)|$ for anomalies (yielding $1/|z(x)|$) and reducing it for normal data (yielding $|z(x)|$). This formulation ensures that normal data cluster near the origin and anomalies are pushed farther away, creating distinct latent distributions. The objective function trains $\mathcal{L}_{RE}$ using only normal data to maintain low RE for normal instances. Inspired by this, the study proposes a method to control the regularization term of DisRAE more effectively, as described in Section 4.

The overview of DisRAE framework is illustrated in Fig. 1 with two phase: the training phase and the testing phase. In the training phase, the model follows supervised learning with known labeled data, which contains normal and large sample of known attack categories. In the testing phase, we use the bottleneck part of DisRAE to capture the representation of few samples of new/rare attack in the hidden space. The embedding of them will be oversampled to make balanced data with normal data for constructing a binary classifier. The query data, which contains normal data and the remain of the new/rare attack category in the dataset, will be used for evaluation.

4. Proposed method

This section outlines proposed framework, namely aDisRAE, which addresses the key limitation of DisRAE by enhancing its training objective to better handle advanced anomalies that closely mimic normal behavior. In the DisRAE approach, anomalous instances are uniformly repelled from the latent-space origin using a fixed regularization strength, which hampers the detection of subtle attacks with features similar to benign data. As a result, these attacks are often mapped to regions hard to distinguish from normal clusters, allowing evasion.

To overcome this, the study incorporates knowledge about the raw data properties directly into the learning process by introducing a prior outlier score r. The outlier score quantifies the inherent anomalousness of each instance based on the original data distribution (e.g., via distance-based metrics or an anomaly learning model). This pre-computed score is integrated into a composite loss function, encouraging the model to simultaneously learn a discriminative representation and apply a targeted repulsive force. Crucially, anomalies that most closely resemble normal data receive a stronger repulsive force, ensuring they are pushed further away in the embedding space than more obvious attacks (i.e., a lower r_i leads to stronger repulsion, while a higher r results in weaker repulsion). This mechanism creates a more uniform distribution between attack patterns and benign traffic, enabling targeted separation that significantly enhances the model's ability to distinguish subtle attacks from normal behavior and improves overall detection performance and robustness. The core idea of the proposed approach is illustrated in Fig. 2. The formulation of the anomaly score and the integration of the loss function are discussed in detail in the following subsections.

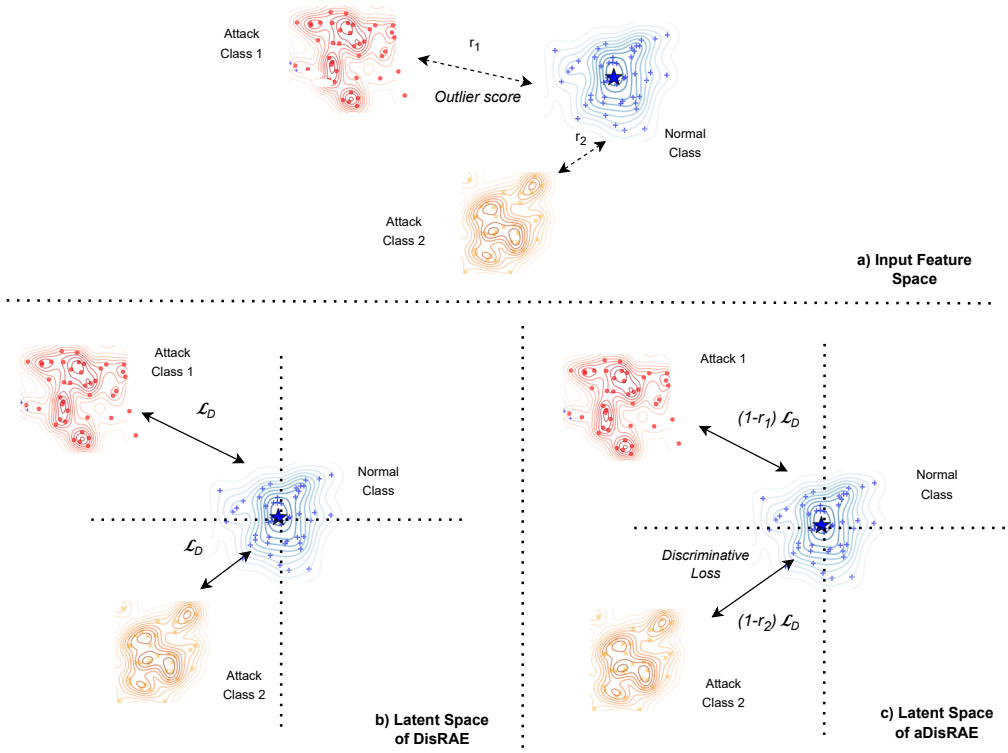


Fig. 2. Distributions of normal samples and cyberattack samples in: a) the input feature space, b) the latent space of DisRAE, and c) the latent space of aDisRAE. Attack class 2 is more sophisticated than attack class 1, which leads to $r_2 > r_1$.

4.1. Outlier score estimation

As previously discussed, the outlier score r_i of a given anomalous example x_i can be generated using either a distance-based method or a machine learning model. For this task, the framework employs the one-time sampling method (proposed in [23]) and an AutoEncoder.

4.1.1. Distance-based method

One-time sampling [23], a distance-based method, is used to estimate the outlier score of anomalous samples relative to normal data. Following [23], the outlier score of a sample x_i with respect to a set X is defined as:

$$r_i = \min_{x_j \in S(X)} \text{dis}(x_i, x_j), \tag{4}$$

where, $S(X)$ is a subset randomly and independently sampled from $X$and $\text{dis}(.,.)$ denotes the pairwise Euclidean distance between x_i and each object $x_j \in S(X)$.

The proposed framework aims to estimate the outlier scores of cyberattack data points against a baseline of normal data. Therefore, each x_i is drawn from the cyberattack

classes, while X represents the normal training data. The size of the subset $S(X)$ is set to 2000. It is important to note that a subset $S(X)$ is sampled only once for evaluation point anomalous examples. This approach makes one-time sampling highly efficient, as it avoids computing distances to the entire dataset and sampling multiple times.

4.1.2. Reconstruction error

Alternatively, the outlier score can be generated using an Autoencoder, a well-known one-class classification method (OCC). As introduced in Section 3, a standard AutoEncoder trained solely on normal data can be used for anomaly detection. This corresponds to a vanilla one-class AutoEncoder, not the DisRAE used in proposed model. Its RE serves as a measure of a data point's abnormality. Specifically, given a data point x_i and its reconstructed version $\hat{x_i}$, the outlier score r_i is calculated as the squared Euclidean norm of the RE:

$$r_i = ||x_i - \hat{x_i}||^2, \tag{5}$$

Equation (5) indicates that a higher score r_i signifies that the sample x_i is more distinct from the normal data on which the model was trained.

4.2. Objective function

The core of the proposed aDisRAE framework is a novel objective function that incorporates the pre-computed outlier score r_i to adaptively regularize the latent space. After estimating r_i for each anomalous sample x_i using the method in Section 4.1, the scores are normalized to $(0, 1)$ via Min–Max scaling.

The overall objective function $\mathcal{L}$ for aDisRAE is a composite loss that combines a reconstruction term $\mathcal{L}_{RE}$ with an adaptive discriminative term $\mathcal{L}_D$. This function is formally defined as:

$$\mathcal{L} = \mathcal{L}_{RE} + \mathcal{L}_D = \frac{1}{n}\sum_{i=1}^{n}(||x_i - \hat{x_i}||^2 + (1 - r_i)|z(x_i)|^{l(x_i)}),$$

where, $z(x_i)$ is the latent representation of the input x_i and $l(x_i)$ is its corresponding label. The behavior of this loss function can be analyzed in two distinct cases:

- *For a normal sample ($l(x_i) = +1$)*: As r_i represents the outlier score capturing how far a sample departs from normal patterns, a normal sample is assigned $r_i = 0$. The loss for a normal sample x_i simplifies to:

  $$\mathcal{L}(x_i) = \|x_i - \hat{x}_i\|^2 + |z(x_i)|. \tag{6}$$

  This objective encourages the model to learn compact representations by minimizing both the RE and the magnitude of the latent vector, effectively pulling normal samples toward the origin. This is consistent with the original DisRAE formulation.

- *For an anomalous sample ($l(x_i) = -1$)*: The pre-computed outlier score $r_i \in (0, 1]$ is used. The loss for an anomalous sample x_i becomes:

$$\mathcal{L}(x_i) = \|x_i - \hat{x}_i\|^2 + (1 - r_i)\frac{1}{|z(x_i)|}. \tag{7}$$

  In this case, the term $(1 - r_i)$ acts as an adaptive regularization weight. Anomalies that are subtle and closely resemble normal data (i.e., having a low outlier score r_i) receive a stronger repulsive force, as the weight $(1 - r_i)$ is larger. This forces their latent representations $z(x_i)$ further from the origin to minimize the loss. Conversely, more obvious anomalies (with a high r_i) receive a weaker repulsive force. This targeted mechanism ensures a more effective separation of hard-to-detect anomalies from the normal cluster.

By training the model with this objective function, aDisRAE learns to create a latent space where the separation between normal and anomalous data is not uniform but is instead adapted to the inherent difficulty of detecting each specific anomaly. To distinguish between the two methods for estimating the outlier score, the model using the distance-based score is hereafter referred to as aDisRAE$_{dis}$, while the model using the reconstruction-based score is referred to as aDisRAE$_{occ}$.

5. Experiments

This section details the experiments designed to evaluate the effectiveness of proposed aDisRAE framework. The experiments are structured to validate the core hypotheses presented in Section 4: (1) that the pre-computed outlier score, r, can effectively quantify the subtlety of anomalous instances and (2) that integrating this score into the objective function leads to superior detection performance, particularly for advanced attacks that mimic normal behavior. The series of experiments are described as follows:

- *Outlier score analysis*: The experiment aims to empirically demonstrate that subtle, hard-to-detect attacks indeed yield lower outlier scores compared to more obvious ones. Thus, this justifies the need for an adaptive repulsion mechanism.
- *Performance evaluation for detection*: This main experiment evaluates the detection performance of aDisRAE$_{dis}$ and aDisRAE$_{occ}$ against key baselines and state-of-the-art methods on three public datasets, including OCC$_{AE}$, DisAE, and DisRAE.
- *Robustness and sensitivity analysis*: Finally, this investigates the robustness of aDisRAE by analyzing its sensitivity to key hyperparameters (k-shot) that represents the number of available anomalies for supporting. This aims to show that aDisRAE maintains its effectiveness even with limited knowledge of attack patterns.

The remainder of this section will introduce the validating datasets and provide detailed experimental settings.

Table 1. The NSL-KDD and UNSW-NB15 datasets

NSL-KDD			UNSW-NB15					
Class	Training	Testing	Class	Training	Testing	Class	Training	Testing
Normal	67,343	9,711	Normal	56,000	37,000	Reconnaissance	10,491	3,496
Probe	11,656	2,421	Generic	40,000	18,871	Analysis	2,000	677
DoS	45,927	7,458	Exploits	33,393	11,132	Backdoor	1,746	583
R2L	995	2,887	Fuzzers	18,184	6,062	ShellCode	1,133	378
*U2R**	52	67	DoS	12,264	4,089	*Worms**	130	44

Table 2. The CICIDS-2017 datasets

Class	Records	Class	Records	Class	Records
Benign	2,359,087	FTP-Patator	7,938	Web Attack - Brute Force	1,507
DoS Hulk	231,072	*SSH-Patator**	5,897	Web Attack - SQL Injection	1,507
Port Scan	158,930	*DoS slowloris**	5,796	Web Attack - XSS	652
DDoS	41,835	DoS Slow-httptest	5,499	Infiltration	36
DoS GoldenEye	10,293	*Botnet**	1,966	Heartbleed	11

5.1. Public datasets

The evaluation employs three widely-used public datasets to ensure a comprehensive assessment: NSL-KDD [24], UNSW-NB15 [25] and CIC-IDS2017 [26]. The study begins with the NSL-KDD dataset, a widely used benchmark containing four primary attack categories: DoS (Denial of Service; e.g., Teardrop, Smurf), Probe (surveillance and scanning attacks; e.g., Satan, Portsweep), R2L (Remote to Local attacks; e.g., snoop, Httptunnel), and U2R (User to Root attacks; e.g., Rootkit, Buffer Overflow). The distribution of these attack categories is detailed in Table 1. Next, the UNSW-NB15 dataset is leveraged, which serves as a benchmark for Network Intrusion Detection Systems. It provides a comprehensive mix of contemporary attack types and updated packet data, thereby offering a more realistic evaluation environment (Table 1). To further test the scalability and robustness of aDisRAE, the CIC-IDS2017 dataset is employed, which consists of extensive network traffic captures with 80 statistical features and covers a diverse range of cyberattacks like DDoS, PortScan and Web attacks as shown in Table 2. The cyberattack groups with "*" are selected as few-shot classes.

5.2. Experimental settings

The architecture of the autoencoders is configured in accordance with the guidelines described in [27]. Following this rule of thumb, the number of hidden neurons in the central hidden layer, denoted as m, is determined by $m = [1 + \sqrt{n}]$, where n denotes the number of original input features.

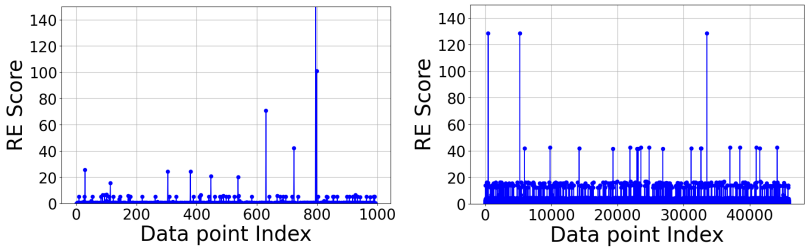
For each dataset, specific attack classes were designated as "few-shot" targets for the testing phase (support and query processes). These were selected based on their low sample counts, which simulates the challenge of detecting rare attacks. The selected classes are: U2R in NSL-KDD; DoS Slowloris, Botnetand SSH-Patator in CIC-IDS2017;

and Worms in UNSW-NB15. During training, these designated few-shot classes were entirely excluded from the training set for the DisAE, DisRAE and aDisRAE models. For instance, with NSL-KDD, the models were trained on normal data and all other attack classes (Probe, DoS, R2L), but not U2R. The testing phase then exclusively evaluated the model's ability to distinguish between normal test data and the unseen U2R attacks. This same protocol was applied to the CIC-IDS2017 and UNSW-NB15 datasets.

The testing phase consists of the supporting and querying processes. The experiment simulates few-shot scenario by constructing a support set containing $k = 20$ samples from the new/rare anomaly class and 200 normal samples. FSL classifiers are then trained on the latent representations generated by DisAE, DisRAE and aDisRAE using this support set. During the training process, batches are formed by randomly selecting 10 samples from the few-shot collection and 10 from the normal data, with the number of epochs set to ensure all anomaly samples were seen at least once. To ensure robust evaluation, the experiment repeats this process 10 times with different, randomly sampled support sets reported by the average Area Under the Curve (AUC) across all runs on query sets.

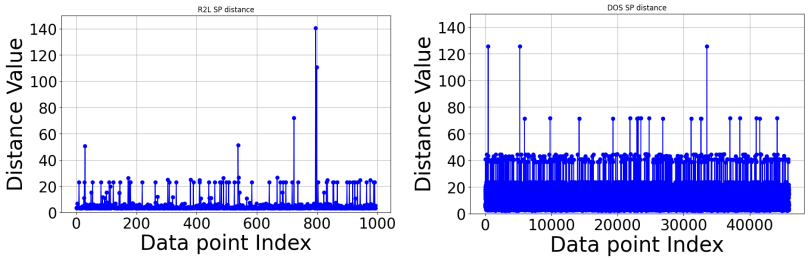
6. Results and discussions

6.1. Outlier score analysis



a) Outlier score of R2L *b) Outlier score of DoS*

Fig. 3. Outlier score of the R2L and DoS attacks estimated by AE-based OCC.



a) Outlier score of R2L *b) Outlier score of DoS*

Fig. 4. Outlier score of the R2L and DoS attacks estimated by One-time sampling.

Table 3. AUCs of aDisRAE in comparison to other methods

Datasets	**NSL-KDD**	**CICIDS-2017**			**UNSW-NB15**
Attack	**U2R**	**DoS Slowloris**	**SSH Patator**	**Botnet**	**Worms**
OCC$_{AE}$	0.834	0.756	0.856	0.859	0.764
DisAE	0.875	0.814	0.869	0.816	0.857
DisRAE	0.888	0.825	0.927	0.920	0.898
aDisRAE$_{dis}$	0.940	0.833	0.932	0.933	0.930
aDisRAE$_{occ}$	0.938	0.873	0.938	0.929	0.963

Fig. 3 and Fig. 4 present the outlier score distributions estimated by AE-based OCC and One-time sampling for two attack categories in the NSL-KDD dataset: DoS and R2L. The fundamental nature of these attacks differs significantly; DoS attacks generate a high volume of anomalous traffic, whereas R2L attacks are more subtle and designed to mimic legitimate access patterns. Consequently, the distributions in Figs. 3 and 4 demonstrate that DoS attacks produce substantially higher outlier scores than R2L attacks under both estimation methods. This indicates that the anomalous behavior of DoS is more readily identifiable. The performance results in Table 3 for aDisRAE$_{dis}$ and aDisRAE$_{occ}$ corroborate this finding, showing distinct efficacy levels for different attack types.

6.2. Performance evaluation

This section analyzes the experimental results presented in Table 3 and Fig. 5, evaluating the performance of the aDisRAE method against the three other methods. The findings empirically validate core hypothesis: the integration of a prior outlier score into the training objective significantly enhances detection capabilities, particularly for advanced anomalies that closely resemble benign data.

The results in Table 3 demonstrate a clear and consistent performance improvement when comparing aDisRAE$_{dis}$ and aDisRAE$_{occ}$, against their foundational models (DisAE and DisRAE) and AE-based one-class classifier (OCC$_{AE}$). The table uses gray-scale to present the performance of these classifiers. In each few-shot attack class (column), the highest AUC is highlighted by the lightest gray. Across all evaluated attack types, both aDisRAE variants achieve higher AUC scores. For instance, on the U2R attack, aDisRAE$_{dis}$ (0.940) and aDisRAE$_{occ}$ (0.938) substantially outperform DisRAE (0.888). A similar significant leap is observed with the Worms attack, where aDisRAE$_{occ}$ achieves an AUC of 0.963, a marked improvement over DisRAE's 0.898. This consistent superiority validates the effectiveness of the adaptive repulsion mechanism. Whereas DisRAE applies a uniform repulsive force to all anomalies, aDisRAE leverages the outlier score r to apply a stronger, more targeted force on subtle attack instances (those with a lower r score). This forces the model to create a more discernible separation in the latent space precisely for the anomalies that are harder to distinguish. The result is a more robust decision boundary, leading to the observed performance gains.

The primary strength of aDisRAE lies in its ability to handle subtle attacks that mimic normal behavior—a known challenge for many anomaly detection systems. The U2R attack in the NSL-KDD dataset is a canonical example of such an attack, as it originates from a legitimate user account and exhibits behavior that is difficult to distinguish from normal activity. As shown in Table 3, it is on this very attack that aDisRAE demonstrates its most significant advantage. The performance jump from DisRAE (0.888) to aDisRAE$_{dis}$ (0.940) is an increase of over 5.2%, highlighting the model's enhanced capacity to identify these evasive threats. Similar pattern is also found on the DoS Slowloris attack. It is characterized by low-and-slow traffic patterns that can easily be mistaken for legitimate, albeit slow, connections.

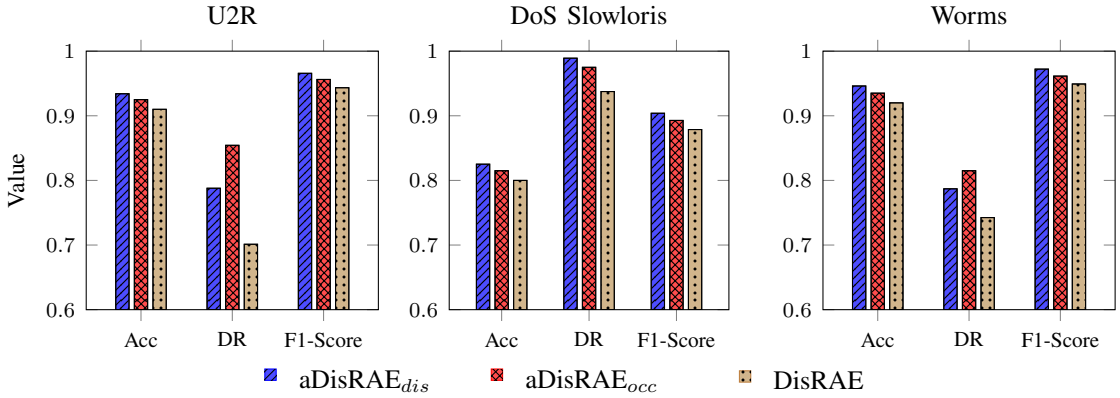


Fig. 5. aDisRAE versus other methods on different metrics.

To assess the practical applicability of the proposed method, the performance was evaluated against DisRAE using Accuracy (Acc), Detection Rate (DR), and F1-Score at a fixed classification threshold, as illustrated in Fig. 5. The results reveal a consistent and significant advantage for the proposed approach across all tested scenarios. Notably, both aDisRAE$_{dis}$ and aDisRAE$_{occ}$ outperform the baseline DisRAE across all three metrics, confirming the overall effectiveness of integrating the outlier score. In addition, the improvement is particularly pronounced in the DR metric (the proportion of anomalies correctly identified). Across all three attack types, especially U2R and Worms, the aDisRAE variants demonstrate a substantial leap in their ability to correctly identify anomalous instances compared to DisRAE. This highlights that the adaptive repulsion mechanism not only enhances general classification accuracy but, more critically, significantly boosts the model's sensitivity to detecting threats.

6.3. k-shot sensitivity

The model's sensitivity to the number of available anomaly samples (k-shot) was analyzed, with results shown in Fig. 6. The findings highlight the robustness of the proposed framework, as both aDisRAE$_{dis}$ and aDisRAE$_{occ}$ consistently outperform the DisRAE baseline across all tested k values, proving effective even with very limited anomaly data ($k = 5$). The two variants exhibit distinct behaviors. aDisRAE$_{dis}$

demonstrates remarkable stability, maintaining high and consistent AUC scores regardless of the value of k. This indicates strong robustness to variations in the support set size. In contrast, aDisRAE$_{occ}$ is more sensitive, showing greater performance fluctuations but also achieving the highest peak AUCs on certain attacks (e.g., Worms and DoS Slowloris). In overall, the aDisRAE variants perform efficiently even with small supporting sets, specifically aDisRAE$_{dis}$ demonstrating the best stability over five settings of k.

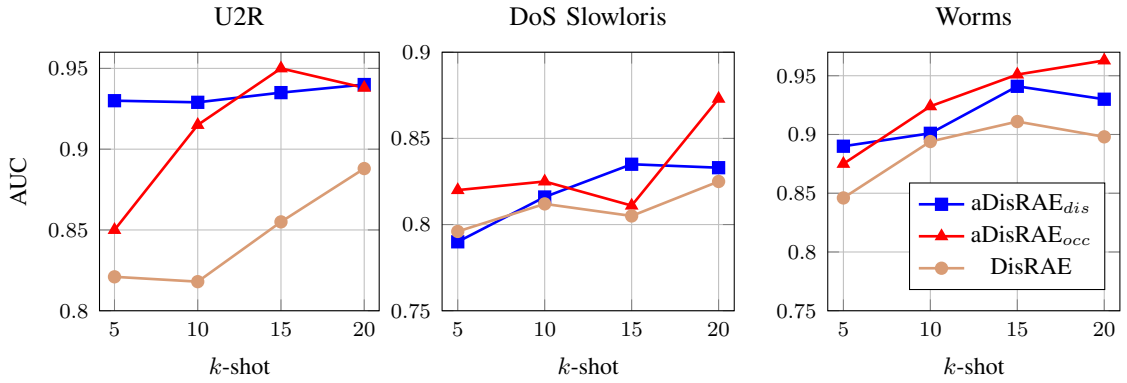


Fig. 6. k-shot sensitivity for AUC across U2R, DoS Slowloris and Worms attacks.

6.4. Complexity and training time

In this approach, the prior outlier score is computed once before training, adding no latency to real-time inference and preserving the baseline's high-speed detection. The training-time overhead depends on the estimation strategy. The One-time sampling method is lightweight, with a low complexity of $O(N \cdot S)$, as it uses a single matrix operation on a fixed-size subset S. In contrast, OCC-AE trains a small autoencoder to capture nonlinear anomaly patterns, resulting in a higher complexity of $O(E \cdot N_{all} \cdot W)$, with E, where E and W denote the number of epochs and model parameters, respectively, but yielding a more semantically informative outlier score.

In summary, the experimental results conclusively demonstrate that proposed aDisRAE framework, by integrating a prior outlier score, consistently and significantly outperforms baseline methods across all evaluated metrics. This enhancement is particularly pronounced for subtle, mimicking attacks, where the adaptive repulsion mechanism proves highly effective at creating a discernible separation for hard-to-detect anomalies. Furthermore, the framework demonstrates remarkable robustness and stability even with very limited anomaly samples, affirming its practical viability for real-world deployment.

7. Conclusions

The paper introduced aDisRAE, an adaptive framework designed to enhance few-shot cyberattack detection by addressing the limitations of the DisRAE model. By integrating a pre-computed outlier score into the training objective, the proposed method

applies a targeted, stronger repulsive force to subtle anomalies that mimic normal behavior. Experimental results on benchmark datasets (NSL-KDD, CIC-IDS2017 and UNSW-NB15) conclusively demonstrate that aDisRAE significantly outperforms baseline methods, improving AUC by up to 10% and showing marked gains in DR, especially for evasive attacks. The framework also proves robust, maintaining high performance even with very limited anomaly samples. For future work, the authors plan to evaluate the effectiveness of aDisRAE in more complex settings, such as real-time streaming environments and on diverse network datasets.

References

[1] V. V. Phoha, "Internet security dictionary," New York: Springer, 2002.

[2] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, Vol. 2, No. 1, pp. 1–22, 2019. DOI: 10.1186/s42400-019-0038-7

[3] M. Hosseini and W. Shi, "Intrusion detection in IoT network using few-shot class incremental learning," in *Future of Information and Communication Conference*. Springer, 2024, Vol. 921, pp. 617–636. DOI: 10.1007/978-3-031-54053-0_41

[4] V. L. Cao, M. Nicolau, and J. McDermott, "Learning neural representations for network anomaly detection," *IEEE Transactions on Cybernetics*, Vol. 49, No. 8, pp. 3074–3087, 2018. DOI: 10.1109/TCYB.2018.2838668

[5] R. Duan, D. Li, Q. Tong, T. Yang, X. Liu, and X. Liu, "A survey of few-shot learning: An effective method for intrusion detection," *Security and Communication Networks*, Vol. 2021, No. 1, 2021. DOI: 10.1155/2021/4259629

[6] A. Yang, C. Lu, J. Li, X. Huang, T. Ji, X. Li, and Y. Sheng, "Application of meta-learning in cyberspace security: A survey," *Digital Communications and Networks*, Vol. 9, No. 1, pp. 67–78, 2023. DOI: 10.1016/j.dcan.2022.03.007

[7] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys*, Vol. 53, No. 3, pp. 1–34, 2020. DOI: 10.1145/3386252

[8] V. L. Cao, M. T. Nguyen, and T. D. Le Dinh, "Few-shot learning with discriminative representation for cyberattack detection," in *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, pp. 1–6. DOI: 10.1109/KSE59128.2023.10299444

[9] Y. Chen, S. Su, D. Yu, H. He, X. Wang, Y. Ma, and H. Guo, "Cross-domain industrial intrusion detection deep model trained with imbalanced data," *IEEE Internet of Things Journal*, Vol. 10, No. 1, pp. 584–596, 2022. DOI: 10.1109/JIOT.2022.3201888

[10] J. Lan, X. Liu, B. Li, and J. Zhao, "A novel hierarchical attention-based triplet network with unsupervised domain adaptation for network intrusion detection," *Applied Intelligence*, Vol. 53, No. 10, pp. 11 705–11 726, 2023. DOI: 10.1007/s10489-022-04076-0

[11] M. J. Hashemi, E. Keller, and S. Tizpaz-Niari, "Detecting unseen anomalies in network systems by leveraging neural networks," *IEEE Transactions on Network and Service Management*, Vol. 20, No. 3, pp. 2515–2528, 2022. DOI: 10.1109/TNSM.2022.3220775

[12] Y. Yu and N. Bian, "An intrusion detection method using few-shot learning," *IEEE Access*, Vol. 8, pp. 49 730–49 740, 2020. DOI: 10.1109/ACCESS.2020.2980136

[13] C. Lu, X. Wang, A. Yang, Y. Liu, and Z. Dong, "A few-shot-based model-agnostic meta-learning for intrusion detection in security of internet of things," *IEEE Internet of Things Journal*, Vol. 10, No. 24, pp. 21 309–21 321, 2023. DOI: 10.1109/JIOT.2023.3283408

[14] F. Rustam, A. Raza, M. Qasim, S. K. Posa, and A. D. Jurcut, "A novel approach for real-time server-based attack detection using meta-learning," *IEEE Access*, Vol. 12, pp. 39 614–39 627, 2024. DOI: 10.1109/ACCESS.2024.3375878

[15] T. Ye, G. Li, I. Ahmad, C. Zhang, X. Lin, and J. Li, "FLAG: Few-shot latent dirichlet generative learning for semantic-aware traffic detection," *IEEE Transactions on Network and Service Management*, Vol. 19, No. 1, pp. 73–88, 2021. DOI: 10.1109/TNSM.2021.3131266

[16] J. He, L. Yao, X. Li, M. K. Khan, W. Niu, X. Zhang, and F. Li, "Model-agnostic generation-enhanced technology for few-shot intrusion detection," *Applied Intelligence*, Vol. 54, No. 4, pp. 3181–3204, 2024. DOI: 10.1007/s10489-024-05290-8

[17] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, *Barcelona, Spain*, Red Hook, NY, USA: Curran Associates Inc.," 2016, p. 3637–3645. DOI: 10.5555/3157382.3157504

[18] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, *Long Beach, California*, Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4080–4090. DOI: 10.5555/3294996.3295163

[19] K. Ding, Q. Zhou, H. Tong, and H. Liu, "Few-shot network anomaly detection via cross-network meta-learning," in *Proceedings of the Web Conference 2021*, New York, NY, USA: Association for Computing Machinery," 2021, p. 2448–2456. DOI: 10.1145/3442381.3449922

[20] C. Xu, J. Shen, and X. Du, "A method of few-shot network intrusion detection based on meta-learning framework," *IEEE Transactions on Information Forensics and Security*, Vol. 15, pp. 3540–3552, 2020. DOI: 10.1109/TIFS.2020.2991876

[21] J. Moon, Y. Noh, S. Jung, J. Lee, and E. Hwang, "Anomaly detection using a model-agnostic meta-learning-based variational auto-encoder for facility management," *Journal of Building Engineering*, Vol. 68, 2023. DOI: 10.1016/j.jobe.2023.106099

[22] S. Razakarivony and F. Jurie, "Discriminative autoencoders for small targets detection," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 3528–3533. DOI: 10.1109/ICPR.2014.607

[23] M. Sugiyama and K. Borgwardt, "Rapid distance-based outlier detection via sampling," *Advances in Neural Information Processing Systems*, Vol. 26, pp. 467–475, 2013. DOI: 10.5555/2999611.2999664

[24] L.Dhanabal and S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 4, No. 6, pp. 446–452, 2015.

[25] 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. DOI: 10.1109/MilCIS.2015.7348942

[26] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." *Proceedings of the 4th International Conference on Information Systems Security and Privacy - ICISSP*, Vol. 1, pp. 108–116, 2018. DOI: 10.5220/0006639801080116

[27] V. L. Cao, M. Nicolau, and J. McDermott, "A hybrid autoencoder and density estimation model for anomaly detection," in *Parallel Problem Solving from Nature–PPSN XIV, Edinburgh, UK, September 17-21, 2016*. Springer, 2016, pp. 717–726. DOI: 10.1007/978-3-319-45823-6_67

Manuscript received 18-10-2025; Accepted 19-12-2025.

■

Manh Tuan Nguyen received the B.Sc. and M.Sc. degree in the field of Electronics and Telecommunications from University of Engineering and Technology, Vietnam National University, Vietnam. He is currently studying the Ph.D. program in Computer Science at Le Quy Don Technical University. His current research interests include Machine Learning, Anomaly Detection and Information Security. Email: tuannm_ncs42@lqdtu.edu.vn.

Le Dinh Trang Dang received the B.Sc. degrees in Control and Automation Engineering from Hanoi University of Science and Technology, Vietnam in 2011 and the Ph.D. degree in Electrical Engineering, focused on VLSI design, from Kyung Hee University, South Korea in 2018. Currently, he works as an assistant lecturer in the Information Technology Department, Le Quy Don Technical University, Vietnam. His major research interests include memory for deep-learning processors, VLSI architectures, hardware security. Email: trangdld@lqdtu.edu.vn.

Van Loi Cao received the B.Sc. and M.Sc. degree in computer science from Le Quy Don Technical University in Vietnam, and the Ph.D degree from University College Dublin, Ireland. He is currently the Head of the Information Security Department at the Institute of Information Technology and Communication, Le Quy Don Technical Unversity. His current research interests include Deep Learning, Machine Learning, Anomaly Detection, IoT Security, and Information Security. Email: loi.cao@lqdtu.edu.vn

aDisRAE: BỘ TỰ MÃ HÓA BIỂU DIỄN PHÂN BIỆT THÍCH ỨNG CHO PHÁT HIỆN TẤN CÔNG MẠNG TRONG ĐIỀU KIỆN ÍT MẪU

Nguyễn Mạnh Tuấn, Đặng Lê Đình Trang, Cao Văn Lợi

Tóm tắt

Do sự khan hiếm dữ liệu bất thường có nhãn, học ít mẫu đã nổi lên như một hệ phương pháp quan trọng để phát hiện các cuộc tấn công mạng mới và hiếm gặp. Mô hình DisRAE học một không gian biểu diễn ẩn nơi dữ liệu bất thường được đẩy ra xa tâm cụm chứa dữ liệu bình thường, nhưng lại gặp khó khăn với các cuộc tấn công có hành vi tương tự hành vi bình thường. Vì vậy, những bất thường tinh vi này thường được ánh xạ quá gần với cụm dữ liệu bình thường, dẫn đến việc khó bị phát hiện. Để giải quyết hạn chế này, bài báo đề xuất mô hình DisRAE thích ứng, gọi là aDisRAE. Phương pháp này cải tiến hàm mục tiêu huấn luyện bằng cách tích hợp độ bất thường tiên nhiệm (*prior outlier score*) nhằm định lượng mức độ tinh vi của mỗi bất thường. Điểm số này sẽ định hướng một cơ chế lực đẩy thích ứng, tác động một lực mạnh hơn lên các bất thường giống với dữ liệu bình thường, đảm bảo sự phân tách hiệu quả hơn trong không gian ẩn. Nhóm tác giả đánh giá aDisRAE trên ba bộ dữ liệu tiêu chuẩn: NSL-KDD, CIC-IDS2017 và UNSW-NB15. Kết quả cho thấy một sự cải thiện hiệu suất đáng kể, làm tăng chỉ số AUC lên đến 10% và thể hiện độ bền vững được nâng cao, đặc biệt đối với các loại tấn công có tính lẩn tránh cao.

Từ khóa

Phát hiện tấn công mạng; phát hiện bất thường; học ít dữ liệu; bộ tự mã hoá phân biệt.