Regular Article

Investigate Discriminative AutoEncoder in Few-shot Learning-based Anomaly Detection

Van Loi Cao

Institute of Information and Communication Technology, Le Quy Don Technical University, Hanoi, Vietnam

Correspondence: Van Loi Cao, loi.cao@lqdtu.edu.vn
Communication: received 26 May 2024, revised 07 June 2024, accepted 12 June 2024
Online publication: 15 June 2024, Digital Object Identifier: 10.21553/rev-jec.375

Abstract– **Discriminative AutoEncoder (DisAE) plays a crucial role in enhancing the adaptability and generalization of few-shot learning methods (DisAEFL) for detecting rare anomalies. DisAE captures meta-knowledge from multiple known tasks, facilitating rapid adaptation in DisAEFL. Key factors like the discriminative parameter (a) and the normal proportion parameter (p_n) significantly impact DisAEFL performance. However, their influence on the DisAE manifold and DisAEFL's efficacy in rare cyberattack detection remains understudied in cybersecurity. This study presents an investigative approach to probe DisAE's influence on DisAEFL's performance in addressing rare/unseen cyberattacks, aiming to gain insight into the DisAE manifold and outline future research directions. Through intensive analysis, we focus on parameters a and p_n, detailing how to examine them to observe DisAE's effects on DisAEFL. Two main experiments are conducted to investigate their influences. Experimental results on the NSL-KDD dataset reveal a strong correlation between these parameters and both the DisAE manifold and DisAEFL performance. These findings suggest strategies for more efficiently constructing the DisAE manifold to enhance DisAEFL's adaptability and generalization. Overall, this study contributes to advancing anomaly detection methodologies in cybersecurity by shedding light on the interplay between DisAE, DisAEFL, and crucial parameters.**

Keywords– **Anomaly detection, meta-learning, few-short learning, fiscriminative autoencoder.**

1 Introduction

The detection of anomalies has been a topic of considerable interest within the research community in recent years, particularly in fields such as network intrusion detection, IoT botnet detection as well as cancer detection and fraud transaction detection [1–5]. In cybersecurity, anomaly detection faces significant challenges arising from the scarcity of cyberattack data, the evolving nature of cyberattack methods, and new malicious code [6, 7]. This can often result in very limited or no data being available to construct generalized models for identifying these types of cyberattacks. Recently, deep learning and machine learning techniques have been developed to address the aforementioned challenges. For instance, approaches such as semi-supervised learning, weakly-supervised learning as well as meta-learning techniques (i.e. few-shot, one-shot, and zero-shot learning) have been employed [8–10]. These approaches enable the construction of detection models from incomplete data (i.e., lacking labels, very little or no data from certain cyberattack types) to detect emerging cyberattacks in the future.

Meta-learning is a training process aimed at creating models with high adaptability to new tasks or domains from limited data [11, 12]. Meta-learning is also known as a learning strategy: learning to learn. By simultaneously learning from multiple tasks during meta-training process, meta-learning methods can extract higher-level knowledge from these tasks, enabling to construct models with high adaptability and generalization for unseen/new tasks in the querying phase. In this regard, few-shot learning is considered a specific case of meta-learning and has found numerous applications in cybersecurity in recent years [6, 7]. Few-shot learning is designed to address problems in contexts where there is very little data available for each class (i.e., few, or even no samples) for training [13–15]. It often employs meta-learning techniques to configure the training process. This allows few-shot learning methods to leverage meta-knowledge from similar tasks to effectively learn unseen/new tasks with few instances per class. In other words, few-shot learning consists of two-stage training: (1) meta-training that attempts to learn higher-level knowledge from similar tasks with available data; (2) supporting that adjusts models by using few samples from the unseen/new task. Few-shot learning is often used to detect types of cyberattacks that often produce limited data (i.e., few or no samples) such as novel and zero-day cyberattacks. This method will use some well-known cyberattacks with abundant data as tasks for the meta-training process, and few samples from unseen/zero-day cyberattacks for the supporting process. By applying the meta-learning strategy, few-shot learning methods can effectively address unseen/zero-day attacks.

Recently, few-shot learning has become a highly effective approach for rare/small cyberattak groups [6–8, 16, 17]. The meta-training phase plays a crucial role in enhancing the generalization and rapid adaptation capabilities of few-shot learning models to new tasks. In meta-learning phase, feature

representations are often constructed for facilitating few-shot learning models to learn new tasks [16, 17]. Discriminative AutoEncoder (DisAE) is recognized as a powerful method for simultaneously learning meta-knowledge for multiple tasks during the meta-training phase. It was introduced by Razakarivony et al. [18] to learn a latent representation (manifold) that supports a classifier in detecting tiny target image classes.

Regarding anomaly detection, DisAE can learn a latent representation from normal and anomalous classes in which the latent representation emphasizes accurate reconstruction of normal instances while pushing anomalies away from the latent space [16–18]. This is achieved by enforcing the reconstruction error (RE) to be smaller than a discriminative parameter a for normal instances, while the RE for anomalies is required to be larger. This approach enables meta-testing to leverage DisAE for adapting models to unseen/new tasks. Typically, the discriminative parameter a is often set to 1.0 in practice [17]. Additionally, the ratio of normal to anomalous data, particularly within meta-training batches, is often aligned with that of the original dataset. The choices of these parameters could significantly influence DisAE and few-shot learning methods. To our best knowledge, the influence of DisAE w.r.t these parameters on the performance of few-shot learning methods has not been examined extensively.

Therefore, this study aims to investigate the influences of DisAE on the performance of DisAE-based few-shot learning methods (DisAEFL) to rare/small unseen cyberattacks. In other words, the behavior of the DisAE manifold is observed w.r.t the discriminative parameter (a) and the normal proportion (p_n) in the meta-training batches. Essentially, the discriminative parameter regulates how effectively the manifold can represent normal examples versus anomalies. Specifically, as the value of parameter a decreases, DisAE becomes better at reconstructing normal data while provides looser constraints on reproducing anomalies. The behavior of the DisAE manifold is reversed as a increases. Regarding the normal proportion, a higher ratio of normal data in meta-training batches may lead to a better representation of normal data. Understanding the characteristics of DisAE w.r.t these parameters can potentially unveil innovative approaches to enhance DisAEFL in the future. The details of how to get insight into the DisAE behavior is presented in Section 4. The key contributions of this study include:

1) Introduce an approach to extensively investigate the characteristics of DisAE in enhancing the performance of DisAEFL.
2) Conduct a series of experiments aimed at analyzing the behavior of DisAE and evaluating the effectiveness of DisAEFL in identifying rare or small cyberattack groups. This study provides insights that inform future avenues for enhancing DisAEFL.

The subsequent sections of the paper are structured as follows: Sections 2 and 3 provide backgrounds of meta-learning, few-shot learning and Discriminative AutoEncode for understanding the followed sections. The investigation approach on the behavior of DisAE is presented in Section 4. Experiments and result discussion are provided in Section 5. Following this is conclusion that highlights remark results and draw future research directions.

2 Backgrounds

This section presents the backgrounds of meta-learning, few-shot learning techniques, as well as Discriminative AutoEncoders. This is fundamental knowledge for understanding the following sections in this paper.

2.1 Meta-Learning and Few-Shot Learning

Meta-learning, also known as learning-to-learn, emerged initially in the educational science community with the first concept by Maudsley et al. [11] before its incorporation into machine learning. Unlike traditional machine learning approaches, meta-learning diverges in its treatment of both the sample set and the query set, which are derived from the labeled dataset [19]. This implies the capacity to construct a task set, $\mathcal{T}_{train}$, consisting of multiple tasks $\mathcal{T}_1, \ldots, \mathcal{T}_m$. Similarly, the task set $\mathcal{T}_{test}$ encompasses tasks like $\mathcal{T}_1, \ldots, \mathcal{T}_q$, specifically designated for testing purposes. Within the realm of meta-learning, $\mathcal{T}_{train}$ and $\mathcal{T}_{test}$ function as the training and test sets, respectively, within the meta-task framework. Consequently, they can be referred to as the meta-training set and meta-testing set, respectively.

Few-shot learning was introduced to solved the challenge of constructing models with limited labeled data, a common scenario in many real-world applications, such as intrusion detection [20]. One powerful approach to few-shot learning involves leveraging meta-learning to facilitate quickly adaptation to new tasks. Few-shot learning uses the concept of *n-way, k-shot* to refer to the number of classes and the number of examples available per class in the training phase (meta-training as well as supporting stages). During meta-training, the model learns to generalize from the meta-training task $\mathcal{T}_{train}$. This process exposes the model to a diverse range of tasks, enabling it to learn a generalized representation of the underlying data distribution. During the meta-testing phase, the model encounters new tasks. In each task, it is provided with a *support set* containing a few examples (*k-shot*) from each of the classes (*n-way*). Additionally, there is a *query set* containing the remaining examples from these classes for evaluating the performance of the resulting model.

By employing meta-learning techniques, few-shot learning methods can create generalization models that are robust and adaptable [14]. These models excel at tasks where labeled data is scarce or costly to obtain, making them highly valuable in domains such as computer vision, natural language processing, and cybersecurity [12].

2.2 Discriminative AutoEncoder

AutoEncoders [21] are renowned for their ability in feature extraction and dimensionality reduction.

The ordinary AE serves as a stalwart in these task, comprising the Encoder, Decoder, and a bottleneck layer (also called latent representation). Operating on unlabeled data x, the reconstruction loss function (RE) aims to minimize the dissonance between the input x and its corresponding output $\hat{x}$. This loss function also quantifies the proximity of the input to the underlying manifold. Depending specific problems, either squared or absolute loss error can be used to defined the RE loss. Equation 1 is the RE loss in form of an squared value.

$$L(x, \hat{x}) = \|x - \hat{x}\|^2 \tag{1}$$

AutoEncoders excel in unsupervised representation learning. Yet, with the rise of sophisticated attacks, distinguishing cyberattack traffic from normal traffic becomes challenging, which presents a problem for classifiers solely reliant on normal data.

In contrast, Discriminative AutoEncoders (DisAEs) emerge as an innovation that leverages supervised learning to carve out distinctive representation spaces for both normal and anomaly classes [17, 18]. These pioneers in anomaly detection endeavor to push anomaly instances far away from the manifold while concurrently mitigating the reconstruction error of normal data. Central to the ability of DisAEs lies its discriminative loss function, as delineated by Raza et al. [18]:

$$L(X^+ \cup X^-) = \max(0, l(x) \times (d(x) - a)) \tag{2}$$

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, $d(x)$ is the reconstruction error (RE) as defined in equation 1, and a is the discriminative parameter. Note that, the authors set a is equal to 1 in the study [18]. For the scenario, equation 2 aims to minimize the distance $d(x)$ to belong within the range $[0, 1]$ for normal instances, while accentuating distances greater than 1 for anomaly instances.

3 Related Work

The application of few-shot learning in cybersecurity has garnered significant attention, with several studies exploring various methodologies to address the challenges of limited labeled data and the need for efficient anomaly detection. Yu et al. [22] proposed a strategy that utilizes a metric-based approach with a traditional softmax function and center loss to tackle the few-shot problem in network anomaly detection. However, their experiments did not evaluate scenarios where the attack class during testing was not included in the training phase. Chaomeng Lu et al. [23] implemented the Model-Agnostic Meta-Learning (MAML) algorithm to address scenarios with scarce trainable samples, transforming numerical network data into images and optimizing parameters using the MAML framework. Cao et al. [16] proposed a few-shot framework consisting of training a discriminative autoencoder and building a classifier trained by representations of normal samples and few labeled anomalies. Furqan Rustam et al. [24] introduced a methodology for real-time collection and detection of network attacks, achieving impressive accuracy using the meta-RF-GNB model. Ye et al. [25] expanded the training dataset using the Latent Dirichlet Allocation algorithm and introduced the Latent Dirichlet Generative Learning scheme for semantic-aware traffic detection.

Xiong Li et al. [26] presented a novel intrusion detection system that enhances few-shot attack detection using generative adversarial networks (GAN) and MAML, demonstrating superior performance in identifying few-shot attacks compared to other methods. Matching Networks [15] and Prototypical Networks [14] utilize distinct embedding functions and prototype representations for anomaly detection. In this context, Ding et al. [27] proposed Graph Deviation Networks (GDN), leveraging few-shot learning for network anomaly detection. Xu et al. [19] introduced a few-shot detection approach based on a meta-learning framework, while Moon et al. [28] combined MAML with variational autoencoder for time-series anomaly detection.

In summary, these studies highlight the promising applications of few-shot learning within cybersecurity, presenting novel approaches to bolster anomaly detection accuracy even when faced with a scarcity of labeled data. Nevertheless, there remains a gap in research regarding the comprehensive exploration of feature representation models during the meta-training phase of few-shot learning techniques. This study seeks to address this gap by delving into the behavior of the Discriminative AutoEncoder introduced in [18], when utilized as a latent representation within few-shot learning frameworks.

4 Investigations on Discriminative AutoEncoders

This section presents how to investigate the behavior of the DisAE manifold to the performance of DisAEFL for cyberattack detection. The overview of DisAEFL is illustrated in Figure 1 with two phases: meta-training and meta-testing. Meta-training is utilized to learn a feature representation from normal data and large cyberattack groups to enhance the generalization ability of few-shot learning models. On another hand, meta-testing adjusts the models to quickly adapt to new tasks using a supporting set. The supporting set contains only few examples from new tasks in meta-testing. Our approach is to get insight into the behavior of DisAE by examining the influence of the discriminative parameter a and the normal proportion p_n on the DisAE manifold as well as the DisAEFL performance. Details of our approach is presented as follows.

Firstly, the discriminative parameter a is moved from a small to a large values for observing the DisAE manifold and the DisAEFL performance. Figure 2 shows the normal and anomalous losses in the loss function of DisAE at $a = 1$. Note that the normal and anomaly losses refer to the reconstruction errors ($d(x)$ in equation 2) of normal data and anomalies, respectively. It can be seen from equation 2 that the smaller the parameter value a is, the more the loss function forces DisAE to learn to

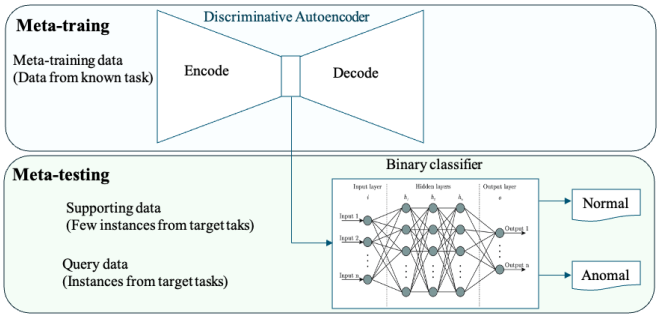


Figure 1. Overview of DisAEFL.

reconstruct normal data better, while allowing anomalies to be distributed in a larger space freely. When the value of parameter a is larger, the loss function can provide a larger space for normal loss while also inhibiting DisAE from learning to recover anomalies. Normal examples often share some common characteristics [29]; a well-trained DisAE often represents them well on meta-training data and performs well on meta-testing data. However, anomalies tend to differ from each other. Therefore, if DisAE represents known anomalies well, it may reduce adaptability to new anomalies in meta-testing. This suggests that a DisAE with a small parameter a can help DisAEFL learn meta-knowledge from tasks in meta-training to adapt and generalize well to rare tasks in meta-testing.

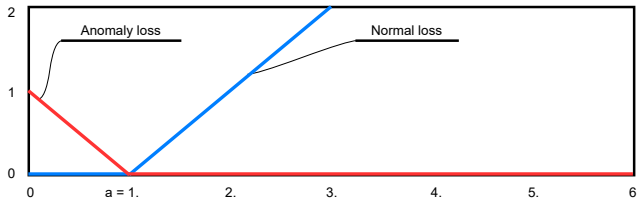


Figure 2. Discriminative loss in the case $a = 1$.

As discussed above, a well-trained DisAE that effectively represents normal data while reconstructing anomalies poorly can help improve the rapid adaptation and generalization of DisAEFL to new tasks. Therefore, the proportion of normal data in the meta-training batch (p_n) may be another influencing factor on the performance of DisAEFL. Thus, this study examines the DisAE manifold and the performance of DisAEFL as the proportion varies from zero (no normal data) to 1.0 (no anomalies). A larger value of p_n indicates that DisAE can represent normal data better, leading to better performance of DisAEFL, and vice versa. To evaluate our investigation approach, we design two main experiments for examining the parameters a and p_n on DisAE and DisAEFL as presented in Section 5.

5 Experiments and Result Analysis

In this section, we present two main experiments to investigate the characteristics of DisAE through the performance of the DisAEFL method w.r.t various settings of the parameters p_n and a. The first experiment aims to examine the influence of a on the latent representation of DisAE. The value of p_n is set equal to 50% for examining a. The parameter a is varied in a range of {0.01, 0.1, 0.5, 1.0, 5.0}. Secondly, we explore the behavior of DisAE w.r.t the normal proportion, $p_n \in \{0\%, 5\%, 10\%, 25\%, 50\%, 75\%, 90\%, 95\%, 100\%\}$. In this case, a is set equal to 1 for investigating p_n. A proportion of zero indicates the absence of any normal points being utilized, whereas a percentage of 100% denotes that exclusively normal data is employed during training. Furthermore, this section provides analysis and discussion to get insight into the DisAE behavior in facilitating DisAEFL to identify unseen/new task of cyberattacks.

The two experiments are carried out on the NSL-KDD dataset. The rare/small cyberattack categories, namely *R2L* and *U2R*, are chosen to create meta-testing tasks $\mathcal{T}_{R2L}$ and $\mathcal{T}_{U2R}$ respectively. Let DisAEFL$_{R2L}$ and DisAEFL$_{U2R}$ denote the DisAEFL model when applying for $\mathcal{T}_{R2L}$ and $\mathcal{T}_{U2R}$, respectively. Other intrusion datasets with diverse types of cyberattacks, such as the CIC-IDS2017 dataset, will be utilized for a comprehensive investigation in our future work. Details of experimental settings and result discussion are presented in Subsection 5.1 and 5.2.

5.1 Experimental Settings

5.1.1 Datasets: In this study, the "NSL-KDD" dataset [30] is utilized to investigate the characteristics of DisAE (i.e. also the performance of DisAEFL) w.r.t various settings of a and the normal proportion p_n. The "NSL-KDD" dataset is generated from the KDD cup 99 dataset [31] eliminating duplicate records within the training dataset as well as within the testing dataset. Thus, it can reduces negative influences of duplicate records on the performance of evaluation models. The number of instances in the training and testing sets is also considered suitable for validating anomaly detection methods. Apart from normal data, the types of cyberattacks present in the training set and testing set of NSL-KDD belong to the following four main groups:

- DoS (Denial of Service): This attack type exhausts the network and system resources of the target computer, such as Back, Land, Smurf, Apache2, Worm and Neptune.
- Probe: This type of attack is aimed at gathering information about servers and networks, such as Satan, Saint and Portsweep.
- Remote-to-Local (R2L): This type of attacks involves remote access to computer systems through vulnerabilities, after which weak account credentials of the target computer can be used to access the target server, such as Guess Password, Warezmaster, and Snmpguess.
- User-to-Root (U2R): These attacks aim to gain root privileges through vulnerabilities or unauthorized actions, such as Rootkit and Sqlattack.

The dataset was designed in two separated set for two stages of constructing detection models such as the NSL-KDD training set and testing set. Each record comprises 42 distinct features in which the first 41

features represent network connection (also known as network flow), and the last attribute refers to its label (i.e. either "normal" or a specific type of cyberattacks). The details of the data distribution is presented in Table I. To prepare NSL-KDD, the categorical features such as *protocol_type*, *service*, and *flag* are convert to numeric values and then encoded by using one-hot encoding. Therefore, the resulting dataset comprises of 122 features after preprocessing. The data set is normalized into the range of $[-1, 1]$ by using the Min-max Scaler from Sklearn[1].

Table I
Normal Data and Cyberattacks in the NSL-KDD Dataset

Category	Training data	Testing data	Total data
Normal	67343	9711	77054
DoS	45927	7458	53385
Probe	11656	2421	14077
R2L	995	2887	3882
U2R	52	67	119
Total	125973	22544	148517

For meta-training, the normal data as well as DoS and Probe in the NSL-KDD training set are employed. This forms the task $\mathcal{T}_{DoS}$ and $\mathcal{T}_{Probe}$ as two task in the meta-training task $\mathcal{T}_{Train}$. In meta-testing, all normal data, R2L and U2R from the NSL-KDD testing set are utilized, resulting in two separated task, $\mathcal{T}_{R2L}$ and $\mathcal{T}_{U2R}$, for the meta-testing task $\mathcal{T}_{Test}$. Specifically, for the supporting stage, a small proportion of normal and each type of cyberattacks are randomly sampled to create supporting sets for $\mathcal{T}_{R2L}$ and $\mathcal{T}_{U2R}$, respectively. This means that we randomly sample 20 normal instances and 20 R2L instances for the supporting set of $\mathcal{T}_{R2L}$ and 20 normal instances and 20 U2R instances for the supporting set of $\mathcal{T}_{U2R}$. The rest of normal data and R2L and U2R are reserved for querying sets of the task $\mathcal{T}_{R2L}$ and $\mathcal{T}_{U2R}$. Note that, R2L and U2R are discarded from the meta-training, while DoS and Probe are absented from the meta-testing.

5.1.2 Metrics: The performance of the DisAEFL method is evaluated using various metrics, including the False Alarm Rate (*FAR*), Missed Detection Rate (*MDR*), Precision, Recall, F1-score, Accuracy (ACC) and the Area Under the Curve (*AUC*). In this evaluation, we denote the normal class as Negative and the anomaly class as Positive. The final outcome of detection models can be categorized into four components using a detection threshold: True Positives (*TP*), False Negatives (*FN*), True Negatives (*TN*), and False Positives (*FP*). Here, *TP* and *TN* represent the number of correctly classified anomaly instances and normal instances, respectively. Conversely, *FN* denotes the number of anomaly examples incorrectly identified as normal data, while *FP* represents the number of normal samples wrongly classified as anomaly. The metrics, such as *FAR*, *MDR*, *Precision*, *Recall*, *F*1-score, and ACC can be defined in equation 3, 4 and 5.

$$FAR = \frac{FP}{FP + TN},\ MDR\ = \frac{FN}{FN + TP}. \tag{3}$$

$$Precision = \frac{TP}{TP + FP},\ Recall\ = \frac{TP}{TP + FN}. \tag{4}$$

$$\begin{aligned} F1\text{-Score} &= \frac{2 \times (Precision \times Recall)}{Precision + Recall}, \\ \text{Accuracy (ACC)} &= \frac{TP + TN}{TP + TN + FP + FN}. \end{aligned} \tag{5}$$

Note that, the *FAR* and *MDR* measure the rate of misclassification, so the smaller these values are, the better the model performs.

On the other hand, AUC provides a comprehensive assessment of model performance across various classification thresholds. AUC quantifies the total area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is constructed by plotting True Positive Rate (TPR) against False Positive Rate (FPR) for different classification thresholds (as depicted in equation 6). TPR represents the ratio of true positives to the sum of true positives and false negatives, while FPR denotes the ratio of false positives to the sum of false positives and true negatives. The AUC metric is often considered more robust than accuracy (ACC) for evaluating model performance, especially in scenarios with highly imbalanced data or few-shot tasks [29].

$$TPR = \frac{TP}{TP + FN},\quad FPR = \frac{FP}{FP + TN}, \tag{6}$$

5.1.3 Parameter Settings: The hyperparameters for DisAEFL are configured based on established conventions and in alignment with prior study. The DisAE component comprises five hidden layers with sizes of 85, 49, 12, 49, and 85, as outlined in [29]. Both input and output sizes are set to 122, matching the data dimensions. Subsequently, the few-shot learning-based classifier (FL) is implemented as a three-layer Multi-layer Perceptron (MLP) network with layer sizes of 12, 5, and 2 respectively. Activation functions across all layers in DisAEFL utilize the hyperbolic tangent function (tanh), with the final layer of the MLP employing the softmax activation function. Training of DisAEFL is facilitated through back-propagation [32] coupled with the Adam optimization algorithm [33].

During the meta-training phase, the model undergoes 50 epochs with a batch size of 1000 to construct a latent representation. In the subsequent support stage, the FL classifier is trained with $n\_way = 2$ (representing two classes) and $k\_shot = 10$ (indicating 10 instances per class). This entails each batch containing 10 normal examples and 10 anomalous instances from the supporting sets detailed in Subsection 5.1.1. To regulate the learning process effectively, an early-stopping mechanism with a patience of 5 is employed. The discriminative parameter a and the proportion of normal data p_n in the meta-training batch are set as described in the preceding paragraph Section 5.

[1]https://scikit-learn.org/stable/

Table II
PERFORMANCE OF DISAEFL$_{R2L}$ W.R.T THE PARAMETER a

Metrics	Discriminative parameter a					
	0.01	**0.1**	**0.5**	**1**	**2**	**5**
FAR	0.227	0.180	0.162	**0.124**	0.284	0.214
MDR	0.215	**0.182**	0.360	0.222	0.410	0.325
Precision	0.505	0.573	0.539	**0.649**	0.380	0.483
Recall	0.785	**0.818**	0.640	0.778	0.590	0.675
F1-score	0.615	0.674	0.585	**0.708**	0.463	0.563
Accuracy	0.776	0.819	0.793	**0.853**	0.687	0.761
AUC	0.779	0.819	0.739	**0.827**	0.653	0.731

Table III
PERFORMANCE OF DISAEFL$_{U2R}$ W.R.T THE PARAMETER a

Metrics	Discriminative parameter a					
	0.01	**0.1**	**0.5**	**1**	**2**	**5**
FAR	**0.078**	0.101	0.104	0.130	0.157	0.104
MDR	0.063	0.106	0.092	**0.032**	0.075	0.100
Precision	**0.055**	0.041	0.041	0.035	0.028	0.040
Recall	0.937	0.894	0.908	**0.968**	0.925	0.900
F1-score	**0.104**	0.078	0.078	0.067	0.054	0.077
Accuracy	**0.922**	0.899	0.896	0.870	0.843	0.896
AUC	**0.929**	0.896	0.902	0.919	0.884	0.898

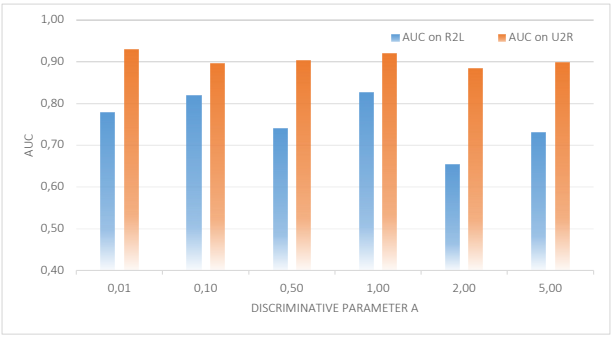


Figure 3. AUC of DisAEFL$_{R2L}$ and DisAEFL$_{U2R}$ w.r.t the parameter a.

Table IV
INFLUENCE OF p_n ON THE PERFORMANCE OF DISAEFL$_{R2L}$

Metrics	Normal proportion (p_n) in meta-training batch (%)								
	0	**5**	**10**	**25**	**50**	**75**	**90**	**95**	**100**
FAR	0.480	**0.027**	0.391	0.240	0.124	0.185	0.121	0.164	0.168
MDR	0.520	0.611	0.526	0.245	0.222	0.215	**0.130**	0.217	0.184
Precision	0.228	**0.809**	0.264	0.482	0.649	0.556	0.680	0.586	0.590
Recall	0.480	0.389	0.474	0.755	0.778	0.785	**0.870**	0.783	0.816
F1-score	0.309	0.526	0.339	0.588	0.708	0.651	**0.764**	0.670	0.685
Accuracy	0.511	0.840	0.578	0.759	0.853	0.808	**0.877**	0.824	0.828
AUC	0.500	0.681	0.541	0.757	0.827	0.800	**0.874**	0.810	0.824

5.2 Results and Discussion

The experimental results are presented in Tables II, III, Figures 3, 4 and 5 for the first experiment, and Tables IV, V, Figures 6, 7 and 8 for the second experiment. Gray-scale is used to highlight the performance of DisAEFL on the metrics. In each row, the best performance is highlighted by the lightest gray. The values in bold indicate the best performance of DisAEFL on its corresponding metrics. In each result tables, we intentionally divide them into 3 groups of metrics to help readers focus on the characteristics of each metric group when measuring the performance of DisAEFL. Specifically, the lower the values of *FAR* and *MDR* are, the better DisAEFL performs, while the *Preision*, *Recall*, $F1-Score$, *ACC* and *AUC* metrics show an opposite direction. In addition, the last row in these tables show the perofmance of DisAEFL in terms of AUC. The AUC metric is more reliable for evaluating anomaly detection methods because it is estimated over a number of classification thresholds [29].

5.2.1 Influence of Discriminative Parameter a on the Performance of DiSAEFL: Tables II and III show the performance of DisAEFL measuring by different metrics w.r.t the parameter a ranging from 0.01 to 5. Table II presents the best performance of DisAEFL$_{R2L}$ on five out of seven metrics at $a = 1$. At smaller values of a (i.e., 0.01, 0.1, and 0.5), DisAEFL$_{R2L}$ also produces competitive performance to that at $a = 1$. In table III, DisAEFL$_{U2R}$ can perform very consistently at $a = 0.01$ as five out of seven metrics also showing the best results. Two other best results are measured by *MDR* and *Recall* at $a = 1$. More importantly, both Table II and III demonstrate that DisAEFL tends to prefer small value of a (1 or smaller values), while poorly identifying the *R2L* and *U2R* attacks on large values of a (i.e., 2 and 5).

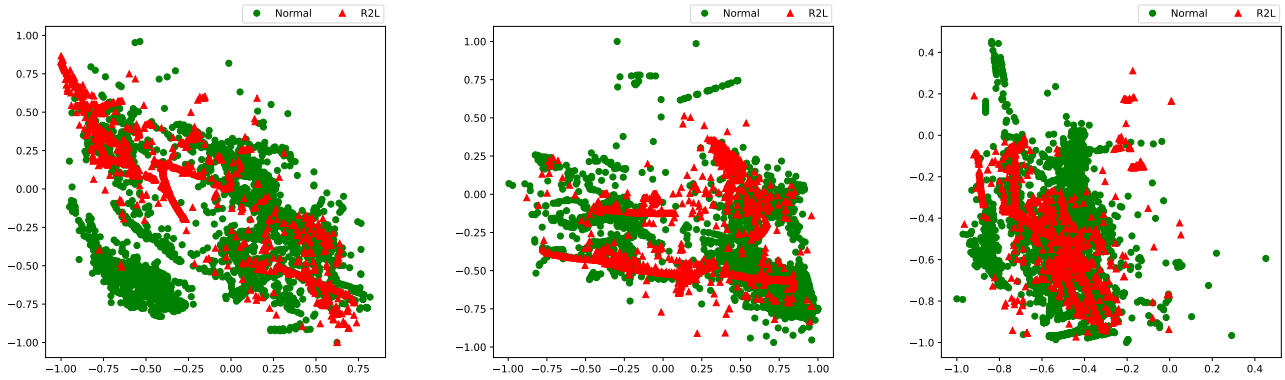
The AUC of DisAEFL$_{R2L}$ and DisAEFL$_{U2R}$ w.r.t to a is illustrated in Figure 2. It also draws a trend that the smaller value a is, the higher AUC values these models often produce. The latent representation of normal data and anomalies (R2L and U2R) from the meta-testing sets are visualized in Figures 4 and 5. It can be observed from the figures that the smaller the value of a, the better the separation between normal data and anomalies. The visualizations can confirm to the results from Tables II and III and our above discussion.

5.2.2 Investigation the Proportion of Normal Data p_n: Tables IV and V present the performance of DisAEFL$_{R2L}$ and DisAEFL$_{U2R}$ w.r.t different values of p_n ranging from 0% to 100%. Note that the discriminative parameter a is fixed as 1. It can be seen that DisAEFL$_{R2L}$ and DisAEFL$_{U2R}$ tend to perform efficiently on large values of p_n, such as $p_n \geq 50\%$, while producing poor performance with $p_n < 50\%$. When $p_n \geq 50\%$, DisAEFL often produces the best measurements at $p_n = 90\%$ on *R2L* and $p_n = 50\%$ for *U2R* with five out of seven metrics on the both cases.

Moreover, Figure 6 illustrates a sharp upward trend of the AUC yielded by DisAEFL$_{R2L}$ and DisAEFL$_{U2R}$ from 0% to 50%, and a slightly increase thereafter. Similarly to the first experiment, we also draw the normal and anomalies (R2L and U2R) from the meta-testing set in the latent representation of DisAE as shown in Figures 7 and 8. These figures illustrate that the separation of

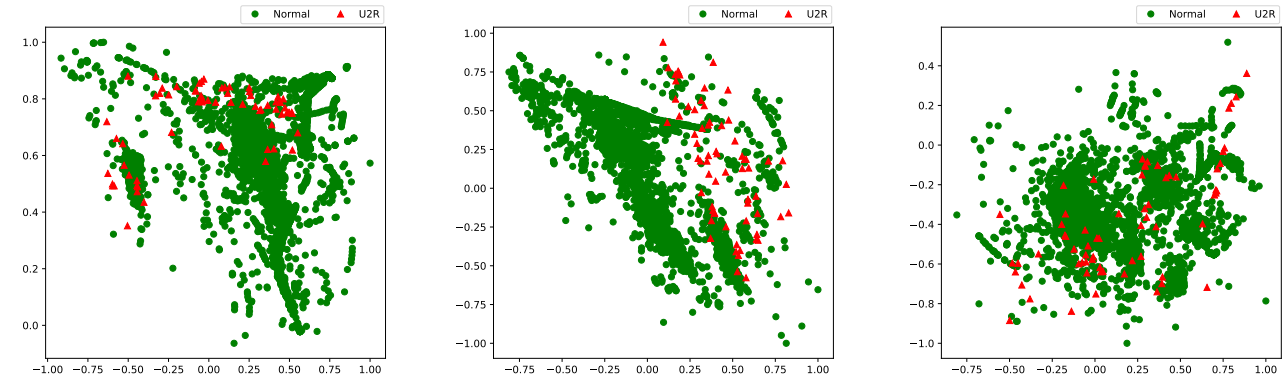
Table V
INFLUENCE OF p_n ON THE PERFORMANCE OF DISAEFL$_{U2R}$

Metrics	Normal proportion (p_n) in meta-training batch (%)								
	0	**5**	**10**	**25**	**50**	**75**	**90**	**95**	**100**
FAR	0.412	0.547	0.460	0.388	0.130	0.082	0.124	0.111	**0.066**
MDR	0.524	0.439	0.506	0.392	**0.032**	0.063	0.134	0.146	0.044
Precision	0.006	0.005	0.005	0.008	0.035	0.053	0.033	0.036	**0.066**
Recall	0.476	0.561	0.494	0.608	**0.968**	0.937	0.866	0.854	0.956
F1-score	0.011	0.01	0.01	0.015	0.067	0.100	0.063	0.069	**0.123**
Accuracy	0.588	0.453	0.54	0.612	0.870	0.918	0.876	0.889	**0.934**
AUC	0.532	0.507	0.517	0.610	0.919	0.928	0.871	0.871	**0.945**



(a) $a = 0.01$ (b) $a = 1.00$ (c) $a = 5.00$

Figure 4. Latent representation of DisAE w.r.t the parameter a on R2L.



(a) $a = 0.01$ (b) $a = 1.00$ (c) $a = 5.00$

Figure 5. Latent representation of DisAE w.r.t the parameter a on U2R.

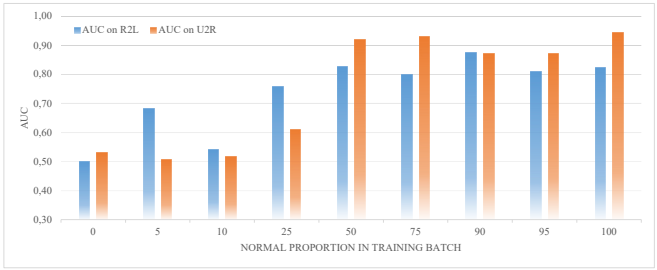


Figure 6. AUC of DisAEFL$_{R2L}$ and DisAEFL$_{U2R}$ w.r.t the parameter p_n.

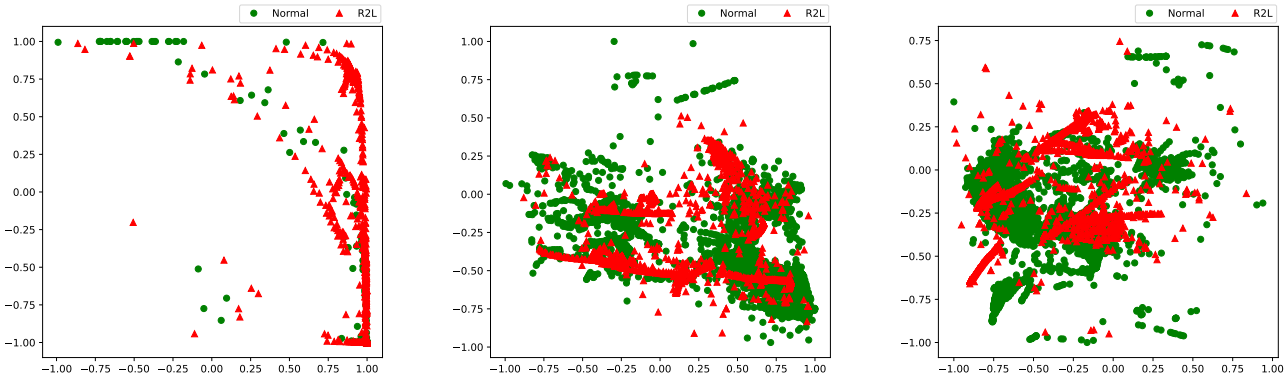
normal data and anomalies (R2L and U2R) prefers large values of p_n such as 50% and 95%. Again, the above results can suggest that the larger the normal proportion in meta-training batch is, the better DisAE can learn normal behavior, and resulting in higher performance of DisAEFL.

In summary, analysis of the experimental results shows that two parameters a and p_n have significant impacts on the DisAE manifold and the performance of DisAEFL in the two task $\mathcal{T}_{R2L}$ and $\mathcal{T}_{U2R}$. Specifically, DisAEFL tends to perform very effectively on small values of a ($a \leq 1$), while DisAEFL prefers meta-training batches with a large proportion of normal data ($p_n \geq 50\%$). Therefore, these results can confirm our investigation approach and discussion presented in Section 4. These results and analyses help open up avenues for future research in enhancing few-shot learning methods for detecting rare/new cyberattack types.

6 Conclusion

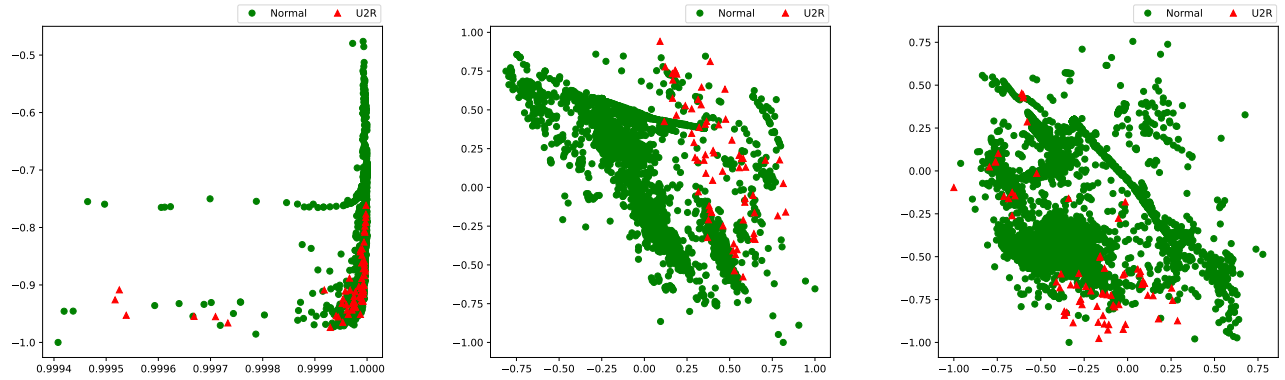
In conclusion, this study introduces an investigative approach to probe the impact of Discriminative AutoEncoder (DisAE) on DisAEFL's performance in detecting rare, unseen cyberattacks. Through extensive analysis, we closely examined DisAE's influence on the adaptability and generalization of DisAEFL, focusing on two critical parameters, a and p_n. Our experiments on the NSL-KDD dataset unveil a strong correlation between these parameters and both the DisAE manifold and DisAEFL. These findings offer valuable insights into constructing the DisAE manifold more efficiently to bolster DisAEFL's adaptability and generalization.

Moving forward, future research should explore DisAEFL's performance across diverse datasets and cyberattack scenarios, and investigate integration with complementary machine learning techniques. Based on this work, a method for optimizing parameters of DisAE will be carried out in the near future.



(a) $p_n = 5\%$ (b) $p_n = 50\%$ (c) $p_n = 95\%$

Figure 7. Latent representation of DisAE w.r.t the parameter p_n on R2L in the query set.



(a) $p_n = 5\%$ (b) $p_n = 50\%$ (c) $p_n = 95\%$

Figure 8. Latent representation of DisAE w.r.t the parameter p_n on U2R in the query set.

References

[1] T. Nishio, M. Nakahara, N. Okui, A. Kubota, Y. Kobayashi, K. Sugiyama, and R. Shinkuma, "Anomaly traffic detection with federated learning toward network-based malware detection in IoT," in *Proceedings of the GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 299–304.

[2] G. Fernandes, J. J. Rodrigues, L. F. Carvalho, J. F. Al-Muhtadi, and M. L. Proença, "A comprehensive survey on network anomaly detection," *Telecommunication Systems*, vol. 70, pp. 447–489, 2019.

[3] S. Liu, B. Cui, and W. Hou, "A survey on blockchain abnormal transaction detection," in *Proceedings of the International Conference on Blockchain and Trustworthy Systems*. Springer, 2023, pp. 211–225.

[4] M. Evangelou and N. M. Adams, "An anomaly detection framework for cyber-security data," *Computers & Security*, vol. 97, p. 101941, 2020.

[5] D. Kumar, C. Verma, Z. Illes, A. Mittal, B. Bakariya, and S. Goyal, "Anomaly detection in chest X-Ray images using variational autoencoder," in *Proceedings of the 6th International Conference on Contemporary Computing and Informatics (IC3I)*, vol. 6, 2023, pp. 216–221.

[6] M. Hosseini and W. Shi, "Intrusion detection in iot network using few-shot class incremental learning," in *Proceedings of the Future of Information and Communication Conference*. Springer, 2024, pp. 617–636.

[7] X. Zhang, Q. Wang, M. Qin, Y. Wang, T. Ohtsuki, B. Adebisi, H. Sari, and G. Gui, "Enhanced few-shot malware traffic classification via integrating knowledge transfer with neural architecture search," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 5245–5256, 2024.

[8] 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.

[9] S. D. A. Rihan, M. Anbar, and B. A. Alabsi, "Meta-learner-based approach for detecting attacks on Internet of Things networks," *Sensors*, vol. 23, no. 19, p. 8191, 2023.

[10] Y. Yan, Y. Yang, F. Shen, M. Gao, and Y. Gu, "Meta learning-based few-shot intrusion detection for 5G-enabled industrial internet," *Complex & Intelligent Systems*, pp. 1–20, 2024.

[11] D. Maudsley, *A Theory of Meta-learning and Principles of Facilitation : an Organismic Perspective*. Thesis (Ed.D.)–University of Toronto, 1979.

[12] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[13] E. Triantafillou, H. Larochelle, J. Snell, J. Tenenbaum, K. J. Swersky, M. Ren, R. Zemel, and S. Ravi, "Meta-learning for semi-supervised few-shot classification," in *Proceedings of the International Conference on Learning Representations*, 2018.

[14] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[15] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.
[16] V. L. Cao, M. T. Nguyen, and T. D. Le Dinh, "Few-shot learning with discriminative representation for cyberattack detection," in *Proceedings of the 2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, 2023, pp. 1–6.
[17] A. S. Iliyasu, U. A. Abdurrahman, and L. Zheng, "Few-shot network intrusion detection using discriminative representation learning with supervised autoencoder," *Applied Sciences*, vol. 12, no. 5, p. 2351, 2022.
[18] S. Razakarivony and F. Jurie, "Discriminative autoencoders for small targets detection," in *Proceedings of the 2014 22nd International conference on pattern recognition*. IEEE, 2014, pp. 3528–3533.
[19] C. Xu, J. Shen, and X. Du, "A method of few-shot network intrusion detection based on meta-learning framework," *TIFS*, vol. 15, pp. 3540–3552, 2020.
[20] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *Proceedings of the International conference on learning representations*, 2016.
[21] H. Bourlard and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological cybernetics*, vol. 59, no. 4, pp. 291–294, 1988.
[22] Y. Yu and N. Bian, "An intrusion detection method using few-shot learning," *IEEE Access*, vol. 8, pp. 49 730–49 740, 2020.
[23] 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.
[24] F. Rustam, A. Raza, M. Qasim, S. Posa, and A. Jurcut, "A novel approach for real-time server-based attack detection using meta-learning," *IEEE Access*, vol. PP, 03 2024.
[25] 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, 2022.
[26] J. He, L. Yao, X. Li, M. Khan, W. Niu, X. Zhang, and F. Li, "Model-agnostic generation-enhanced technology for few-shot intrusion detection," *Applied Intelligence*, pp. 1–24, 02 2024.
[27] 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*, 2021, pp. 2448–2456.
[28] 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, p. 106099, 2023.
[29] 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, 2019.
[30] S. Revathi and A. Malathi, "A detailed analysis on nsl-kdd dataset using various machine learning techniques for intrusion detection," *International Journal of Engineering Research & Technology (IJERT)*, vol. 2, no. 12, pp. 1848–1853, 2013.
[31] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *Proceedings of the 2009 IEEE symposium on computational intelligence for security and defense applications*. Ieee, 2009, pp. 1–6.
[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representations by error propagation*. Cambridge, MA, USA: MIT Press, 1986, p. 318–362.
[33] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of machine learning research*, vol. 12, no. 7, 2011.

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