Regular Article

ProtoKNN: A Hybrid ProtoNets-KNN for Few-Shot Cyberattack Detection

Le Hoan Hoang, Manh Tuan Nguyen, Van Loi Cao

Le Quy Don Technical University, Hanoi, Vietnam

Correspondence: Van Loi Cao, loi.cao@lqdtu.edu.vn

Communication: received 24 February 2026, revised 17 March 2026, accepted 24 March 2026
Online publication: 26 March 2026, Digital Object Identifier: 10.21553/rev-jec.438

Abstract– **Modern network security systems have faced significant challenges from novel attacks with extreme data scarcity, known as few-shot learning problem (FSL). Meta-learning, particularly Prototypical Networks (ProtoNets), has emerged as a promising solution to this problem. However, ProtoNets rely on Euclidean distance to a single prototype, assuming isotropic and spherical class distributions. We argue that network traffic is too diverse for simple clusters; its complex feature distributions cause "centroid misalignment", where a single center cannot accurately represent the attack. To address this, we propose a Hybrid ProtoKNN method. By integrating a local KNN metric with the global prototypical objective, we relax the spherical constraint and effectively recover misaligned outliers. We evaluate our approach on the NSL-KDD and CIC-IDS2017 datasets. Experimental results in various few-shot scenarios demonstrate that our model significantly improves detection performance on rare and complex attack categories, such as U2R, R2L, and Heartbleed, compared to standard meta-learning methods.**

Keywords– **Intrusion detection, few-shot learning, prototypical letworks, *k*-nearest neighbors, meta-learning.**

1 Introduction

Cybersecurity is facing a growing challenge from sophisticated zero-day exploits that attempt to bypass the current security systems [1–5]. To counter these evolving threats, network security systems, such as Network Intrusion Detection Systems (NIDS), should be capable of recognizing novel attack patterns from only a few labeled examples. This task can be known as Few-Shot Learning problem (FSL) [6]. In practice, network traffic often follows a heavy-tailed distribution: while normal connections and volumetric attacks (e.g., DDoS) are abundant, highly dangerous exploits (e.g., User-to-Root, Remote-to-Local, Shell) are extremely rare [7–9]. Consequently, enabling NIDS to generalize from limited data has become a critical research priority.

Recent literature has explored Metric-based FSL approaches, such as Prototypical Networks (ProtoNets) [10] and Matching Networks (MatchingNets) [11], as potential solutions to this data scarcity [12]. The core idea is to learn a shared embedding space where similar samples cluster together, allowing a classifier to categorize query instances based on their distance to known support samples. However, applying these generic algorithms to NIDS presents significant challenges. Unlike visual data (e.g., images), network traffic features exhibit extreme intra-class variance and high dimensionality [3, 5, 9]. Existing methods frequently struggle to construct a discriminant metric space that is robust against the noise and complexity of network behaviors, leading to suboptimal performance on minority classes [8].

A fundamental limitation lies in the geometric assumptions of these models. ProtoNets rely on the Euclidean distance between query instances and a single class prototype for classification, while other variants employ parametric density methods, such as Gaussian-based approaches, to model the support samples [10]. These techniques implicitly assume that class distributions are unimodal and spherical (Gaussian-like) [12]. We argue that such rigid assumptions fundamentally contradict the nature of network traffic, where attack classes typically form arbitrary, anisotropic, or elongated distributions. When a spherical constraint is imposed on such complex geometries, the decision boundary often fails to encompass the scattered "tails" of the data, leading to a critical failure known as "centroid misalignment".

This raises a critical research question: How to capture the arbitrary shapes of attack manifolds from limited data without relying on the restrictive spherical assumption? To address this, we propose a hybrid ProtoNets and *k*-Nearest Neighbors method (called ProtoKNN) that relaxes the global spherical constraint by integrating local neighborhood affinity [13]. This work addresses the research question along three axes. First, we identify Centroid Misalignment as a primary bottleneck for minority-class performance through a detailed geometric analysis. Second, we introduce a hybrid objective that balances the stability of global clustering (prototypical loss) with the flexibility of local manifold learning (KNN loss), allowing the model to adapt to irregular distributions of attacks. Finally, extensive experiments are carried out on the NSL-KDD [8] and CIC-IDS2017 [14] datasets to evaluate ProtoKNN

in comparison to ProtoNets and MatchingNets. This demonstrates that our method often out-performs ProtoNets and MatchingNets on rare, high-variance attack classes including U2R, R2L, Infiltration, Web Attack SQL Injection, and Heartbleed across various few-shot learning scenarios.

The remainder of this paper is organized as follows. Sections 2 and 3 review recent studies on few-shot learning paradigms in cybersecurity and present foundational knowledge on metric-based few-shot learning. Following these, the proposed ProtoKNN is detailed in Section 4. Section 5 provides an extensive evaluation and discussion of ProtoKNN to existing methods on benchmark datasets. Finally, Section 6 concludes the paper and outlines potential directions for future research.

2 Related Work

Network attack identification has traditionally relied on supervised learning, where traffic patterns are analyzed to detect malicious activities. Extensive surveys [5, 9] have categorized these techniques, highlighting their dependence on large-scale, labeled datasets. However, as noted by Sommer and Paxson [3], the "closed world" assumption of machine learning often conflicts with the dynamic nature of network security, where novel attacks constantly emerge. Furthermore, traditional anomaly detection often suffers from high false alarm rates [4]. Analysis of benchmark datasets also reveals that real-world traffic exhibits complex, heavy-tailed distributions that are difficult to model using standard statistical approaches [8].

To address the scarcity of labeled samples for novel attacks, researchers have turned to Few-Shot Learning (FSL). Metric-based meta-learning has emerged as a dominant paradigm, aiming to learn a transferable similarity measure [12]. Foundational works in this area include Matching Networks [11], which use attention-based embeddings; Relation Networks [15], which learn non-linear metric via neural networks, and Model-Agnostic Meta-Learning (MAML) [16], which focuses on fast optimization-based adaptation. Among these, ProtoNets [10] have become a standard benchmark due to their computational efficiency.

In cybersecurity, Xu *et al.* [17] successfully adapted the meta-learning to traffic classification, demonstrating superior performance to deep learning methods in data-scarce scenarios. Recently, Yu *et al.* [18] applied FSL to mitigate class imbalance in intrusion detection datasets; and Cao *et al.* [19] proposed a discriminative representation method for few-shot cyberattack detection. Despite these advancements, centroid-based methods suffer from a rigid geometric assumptions. By utilizing Euclidean distance to a single prototype, they implicitly assume that class distributions are unimodal and spherical (Gaussian-like) [10], which rarely holds true for complex network traffic.

To overcome the limitations of Euclidean metrics, recent research has explored manifold-aware approaches that consider the topological structure of data. Graph Neural Networks (GNNs) have gained traction for their ability to model complex dependencies in non-Euclidean domains [20]. For instance, Dynamic Graph CNNs learn directly from the local structure of feature spaces [21]. Lo *et al.* [22] proposed E-GraphSAGE that utilizes GNNs to capture edge-based intrusion patterns in IoT networks. While GNNs are powerful, non-parametric methods like KNN offer a robust alternative for local density estimation. KNN decision boundaries naturally adapt to the local geometry of a data manifold, regardless of its global shape [13]. Our work builds upon this insight, integrating the local flexibility of KNN with the global stability of ProtoNets to address geometric misalignment in FSL problems.

3 Background

This section provides the theoretical foundation for understanding our hybrid approach. It first defines the FSL problem and the meta-learning paradigm used to address it. This section then details the specific mechanisms of ProtoNets and KNN, which serve as the primary constituents of our manifold-aware architecture.

3.1 Few-Shot Learning

Few-Shot Learning (FSL) is a paradigm designed to enable models to generalize to novel classes using only a limited number of labeled examples [12]. To address the challenges of data scarcity, the meta-learning (or "learning-to-learn") approach is commonly adopted. The core objective is to train the model across a diverse of tasks to acquire a transferable metric space, allowing it to generalize to unseen classes with minimal supervision.

To implement meta-learning, the episodic training strategy introduced by Vinyals *et al.* [11] is utilized. This strategy mimics the few-shot evaluation environment during the training phase by organizing data into discrete "episodes" rather than mini-batches. Each training iteration is structured as an N-way K-shot classification task, where an episode comprises two distinct subsets:

- Support Set ($\mathcal{S}$): $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N\times K}$ serves as the reference knowledge, containing K labeled examples ($\mathbf{x}_i$, y_i) for each of the N classes.
- Query Set ($\mathcal{Q}$): $\mathcal{Q} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{M}$ consists of M unseen samples from the same classes used to calculate the loss and evaluate the model's adaptation performance.

Let f_ϕ be an embedding function (typically a neural network) parameterized by ϕ, which maps the input features into an embedding space. The meta-learning objective is to optimize ϕ to minimize the prediction error on $\mathcal{Q}$ conditioned on $\mathcal{S}$. This is typically achieved by minimizing the negative log-likelihood over a distribution of episodes

$$\mathcal{L}_{\text{meta}} = \frac{1}{|\mathcal{Q}|} \sum_{(\mathbf{x},y)\in\mathcal{Q}} -\log p(y|\mathbf{x}, \mathcal{S}; \phi). \tag{1}$$

3.2 Prototypical Networks

Prototypical Networks [10] represent a state-of-the-art (SOTA) metric-based meta-learning algorithm. They operate on the premise that there exists an embedding space where points cluster around a single prototype representation for each class. For given class j, the prototype $\mathbf{c}_j$ is computed as the mean vector of the embedded support samples

$$\mathbf{c}_j = \frac{1}{|\mathcal{S}_j|} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}_j} f_\phi(\mathbf{x}_i), \tag{2}$$

where $\mathcal{S}_j \subset \mathcal{S}$ is the subset of support samples belonging to class j.

For a query $\mathbf{x}_q$, the model produces a probability distribution over classes based on the squared Euclidean distance $d(\cdot,\cdot)$ between the query embedding $f_\phi(\mathbf{x}_q)$ and each prototype $\mathbf{c}_j$

$$p(y = j|\mathbf{x}_q, \mathcal{S}; \phi) = \frac{\exp\left(-d(f_\phi(\mathbf{x}_q), \mathbf{c}_j)\right)}{\sum_{l=1}^{N} \exp\left(-d(f_\phi(\mathbf{x}_q), \mathbf{c}_l)\right)}, \tag{3}$$

where $d(f_\phi(\mathbf{x}_q), \mathbf{c}_j) = \left\|f_\phi(\mathbf{x}_q) - \mathbf{c}_j\right\|_2^2$.

During the meta-training phase, the parameters ϕ are optimized by minimizing the negative log-likelihood (prototypical loss) for the ground-truth class j

$$\mathcal{L}_{\text{proto}} = -\log p(y = j|\mathbf{x}_q, \mathcal{S}; \phi). \tag{4}$$

In the meta-testing (inference) stage, the class with the highest probability is assigned to the query instance. While effective for unimodal distributions, this global centroid-based approach may struggle with the complex, non-spherical manifolds often found in network traffic data.

3.3 k-Nearest Neighbors

The k-Nearest Neighbors (KNN) algorithm is a non-parametric decision rule for pattern classification, introduced by Cover and Hart [13]. KNN classifies a query $\mathbf{x}_q$ based on the local topology of the feature space without assuming a specific geometric distribution for the data. Let $\mathcal{N}_k(\mathbf{x}_q, \mathcal{S}_j)$ be the set of k nearest neighbors of the query $\mathbf{x}_q$ within the support subset $\mathcal{S}_j$ of class j

$$\mathcal{N}_k(\mathbf{x}_q, \mathcal{S}_j) = \underset{\substack{\mathcal{S}' \subset \mathcal{S}_j \\ |\mathcal{S}'| = k}}{\arg\min} \sum_{\mathbf{x}_i \in \mathcal{S}'} \|f_\phi(\mathbf{x}_q) - f_\phi(\mathbf{x}_i)\|_2, \tag{5}$$

where k is the neighborhood size and $\|\cdot\|_2$ denotes the standard L_2 norm. The local affinity, or representative distance from the query to class j, is defined as the average distance to these k neighbors

$$d_{\text{knn}}(\mathbf{x}_q, j) = \frac{1}{k} \sum_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_q, \mathcal{S}_j)} \|f_\phi(\mathbf{x}_q) - f_\phi(\mathbf{x}_i)\|_2. \tag{6}$$

By focusing on local density rather than a global mean, KNN can naturally adapt to arbitrary, non-spherical manifolds. This provides a robust mechanism to handle the outliers and scattered data patterns inherent in network intrusions.

4 Proposed Method

The proposed ProtoKNN method is designed to overcome the Centroid Misalignment problem identified in PrototNets. By integrating global class representations with local manifold affinity, ProtoKNN can remove the restrictive spherical assumption to better capture the complex, elongated geometries of attack distributions. The followed subsections present how to define the local manifold and integrate it in ProtoKNN.

4.1 Manifold-aware correction strategy

Standard ProtoNets assume that classes form compact, hyperspherical clusters. However, as illustrated in Figure 1 (left), network attack distributions are often anisotropic or elongated. In such cases, a query sample at the distribution "tail" may be geometrically closer to the normal class centroid than its own class prototype. Therefore, the reliance on a single global prototype can lead to misclassification.

To address this, we introduce a manifold-aware correction using KNN, shown in Figure 1 (right). By integrating local neighborhood proximity, the model can recognize the structural continuity between the query and adjacent attack instances. This dual-path integration ensures that the model maintains global structural stability while remaining flexible to recover tailing (isolated) points that are globally misaligned from their class centers.

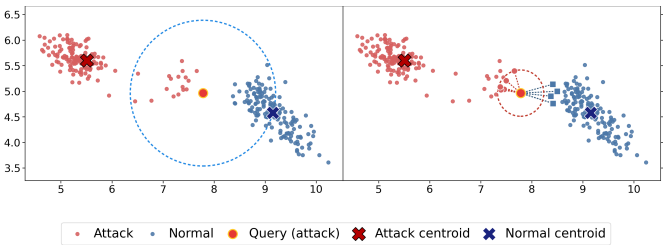


Figure 1. Centroid misalignment (left) and Manifold-aware correction via KNN (right).

4.2 Hybrid ProtoKNN

For each class j in a N-way episode, we compute a hybrid distance that views the embedding space through two terms: *Global similarity* and *Local manifold proximity*.

Global similarity: we measure the squared Euclidean distance between the query embedding z_q and the prototype $\mathbf{c}_j$ of class j

$$d_{\text{proto}}(\mathbf{z}_q, \mathbf{c}_j) = \|\mathbf{z}_q - \mathbf{c}_j\|_2^2, \tag{7}$$

where z_q is computed as $\mathbf{z}_q = f_\phi(\mathbf{x}_q)$.

Local manifold proximity: To resolve the spherical limitation, we evaluate the proximity to the local data manifold

$$d_{\text{knn}}(\mathbf{z}_q, j) = \frac{1}{k} \sum_{\mathbf{z}_s \in \mathcal{N}_k(\mathbf{z}_q, \mathcal{S}_j)} \|\mathbf{z}_q - \mathbf{z}_s\|_2, \tag{8}$$

where $\mathcal{S}_j$ is the set of support samples of class j; $\mathcal{N}_k(\mathbf{z}_q, \mathcal{S}_j)$ denotes the subset of k support

instances $\in \mathcal{S}_j$ that are closest to the query embedding $\mathbf{z}_q$. This local metric effectively traces the manifold structure of attack classes by considering individual instance proximities.

We then combine these terms into a unified hybrid distance metric, modulated by a balancing hyperparameter $\alpha \in [0, 1]$

$$D(\mathbf{z}_q, j) = \alpha \cdot d_{\text{knn}}(\mathbf{z}_q, j) + (1 - \alpha) \cdot d_{\text{proto}}(\mathbf{z}_q, \mathbf{c}_j). \quad (9)$$

The second term ensures rapid convergence for samples near the class core, while the linear KNN term recovers instances from the "tail" of the class. Therefore, this guides training stability across complex attack manifolds.

Finally, the model is trained end-to-end by minimizing the hybrid loss function $\mathcal{L}_{\text{hybrid}}$, defined as the negative log-likelihood of the true class y_q across the query set $\mathcal{Q}$

$$\mathcal{L}_{\text{hybrid}} = \frac{1}{|\mathcal{Q}|} \sum_{(\mathbf{x}_q, y_q) \in \mathcal{Q}} -\log \left(\frac{\exp(-D(\mathbf{z}_q, y_q))}{\sum_{l=1}^{N} \exp(-D(\mathbf{z}_q, l))} \right), \quad (10)$$

where y_q is the ground-truth label of $\mathbf{x}_q$, and N is the number of classes in the episode. By optimizing this objective, the embedding function f_ϕ learns representations that simultaneously encourage inter-class separability and intra-class manifold continuity. This effectively addresses the irregular distributions of rare network attacks.

5 Evaluation and Discussion

In this section, we design two experiments to evaluate the effectiveness of ProtoKNN: (1) compare the performance to ProtoNets and the MatchingNets; (2) investigate the influence of the trade-off hyperparameter α. The performance of these models are measured using the Area Under the ROC Curve (AUC) and the F1-Score on two benchmark datasets.

5.1 Datasets

In this study, two benchmark network security datasets are employed for experiments, namely the NSL-KDD [8] and CIC-IDS2017 datasets [14]. The NSL-KDD dataset is a refined version of the original KDD'99, designed to eliminate redundant records and improve the complexity of the classification task. It includes four attack categories: DoS, Probe, R2L, and U2R. The statistical distribution is presented in Table I.

Table I
Statistics of the NSL-KDD Dataset (KDDTrain+)

No.	Attack Type	Samples	Percentage (%)
1	Normal	67,343	53.46%
2	DoS	45,927	36.46%
3	Probe	11,656	9.25%
4	*R2L*	995	0.79%
5	*U2R*	52	0.04%
	Total	**125,973**	**100.00%**

The CIC-IDS2017 dataset captures modern network traffic and a wide variety of contemporary attack scenarios. It comprises over 2.8 million samples distributed across 15 attack categories. As shown in Table II, the dataset exhibits extreme class imbalance, which is ideal for evaluating few-shot learning. From this dataset, we select infrequent and hard-to-detect attacks to serve as our few-shot classes such as R2L and U2R in NSL-KDD, and Heartbleed and Infiltration in CIC-IDS2017. No data augmentation is applied to the minority classes. The extreme class imbalance is inherently addressed by the few-shot learning paradigm.

Table II
Statistics of the CIC-IDS2017 Dataset

No.	Attack Type	Samples	Percentage (%)
1	Benign	2,272,688	80.324452%
2	DoS Hulk	230,124	8.133358%
3	PortScan	158,930	5.617122%
4	DDoS	128,027	4.524906%
5	DoS GoldenEye	10,293	0.363789%
6	FTP-Patator	7,938	0.280556%
7	SSH-Patator	5,897	0.208420%
8	DoS slowloris	5,796	0.204850%
9	DoS Slowhttptest	5,499	0.194353%
10	Bot	1,966	0.069485%
11	Web Attack Brute Force	1,507	0.053262%
12	Web Attack XSS	652	0.023044%
13	Web Attack SQL Injection	21	0.000742%
14	*Infiltration*	36	0.001272%
15	*Heartbleed*	11	0.000389%
	Total	**2,829,385**	**100.00%**

5.2 Experimental settings

This subsection details the episodic configuration for ProtoKNN, MatchingNets and ProtoNets. The KNN parameter is specifically chosen to align with this episodic configuration. The trade-off hyperparameter α is set to a balanced weight of 0.5. Following the episodic training strategy, datasets are split into disjoint class sets. Abundant classes are utilized during the meta-training phase, while rare attack categories are strictly reserved as unseen classes for the meta-testing phase.

Meta-training phase: Each episode is structured as an N-way task, specifically 4-way for NSL-KDD and 14-way for CIC-IDS2017. We employ a substantial support set size of $n_{support} = 100$ and $n_{query} = 50$ samples per class. This dense configuration ensures a stable and statistically representative estimation of class distributions. By leveraging the abundant traffic data available in the training classes, the hybrid objective effectively captures both global class prototypes and intricate local manifold structures. This provides a sufficiently dense neighborhood for the embedding function to learn and represent complex attack geometries. To optimize the embedding function, the model is trained end-to-end for 1,000 episodes using the Adam optimizer with a learning rate of 0.001. During this procedure, the hybrid loss function is iteratively minimized based on the prediction error across the query sets.

Meta-testing phase: The model is evaluated via a 2-way binary classification task, distinguishing "Normal" traffic from a specific rare attack category. To rigorously test generalization under extreme scarcity, we measure performance across two distinct scenarios: 1-shot ($n_{support} = 1$) and 5-shot ($n_{support} = 5$). This setup ensures a comprehensive assessment of the model's generalization capability.

KNN configuration: Within ProtoKNN, the local neighborhood size is set to a default value of $K = 10$. For meta-testing, the neighborhood size is adjusted to $K = \min(K, n_{support})$. This constrain ensures that the local manifold estimation remains valid even when the support set is extremely small.

5.3 Result discussion

This subsection details our experimental findings. While Section 5.3.1 discusses the main performance results, the subsequent parts explore the impact of the hyperparameter α and the characteristics of the embedding spaces, providing a deeper explanation for the effectiveness of ProtoKNN.

5.3.1 Performance of ProtoKNN: The experimental results, summarized in Table III and Table IV, demonstrate the robustness and generalization capability of ProtoKNN across diverse few-shot scenarios. In the 1-shot scenario, ProtoKNN consistently outperforms ProtoNets across nearly all attack categories. Specifically, on the NSL-KDD dataset, our method achieves peak AUC values of 93.60% for R2L and 96.73% for U2R. This improvement suggests that by relaxing the rigid spherical constraint inherent in centroid-based ProtoNets, ProtoKNN objective better accommodates the irregular, non-convex manifolds often associated with minority attack classes.

As the number of examples increases to the 5-shot setting, ProtoKNN achieves near-perfect AUC for Heartbleed (99.65%) and maintains superior performance for U2R (97.20%). While MatchingNets exhibit competitive AUC for the scattered Infiltration class, ProtoKNN offers a more adaptable alternative to ProtoNets. It demonstrates a notable capacity to encompass misaligned outliers: samples that typically deviate from a single class centroid but remain identifiable within a dense local neighborhood. The F1-score results in Table IV confirm the superior detection effectiveness of ProtoKNN under extreme data scarcity. In the challenging 1-shot setting, while MatchingNets suffer from significant performance degradation due to noise sensitivity, ProtoKNN maintains a robust F1-score of 88.53% for R2L attacks. This stability demonstrates that integrating local manifold structures allows the model to capture precise attack signatures from only a single labeled example. As the support set increases to 5-shot, ProtoKNN consistently achieves peak F1-scores, notably reaching 93.52% for U2R and 97.53% for Heartbleed. Such high F1-scores emphasize the model's ability to minimize false positives while ensuring high recall, making it a reliable solution for identifying rare and evolving network threats.

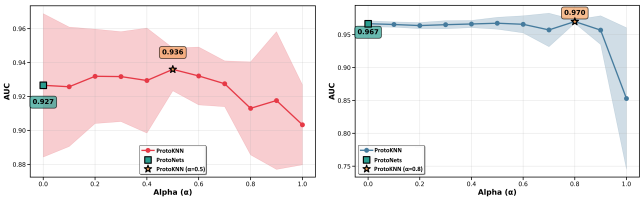
Table III
AUCs of FSL-based models in 1-Shot and 5-Shot Scenarios

Dataset (Attack Class)	Shot	MatchingNets	ProtoNets	ProtoKNN
NSL-KDD (R2L)	1	56.86 ± 4.06	92.66 ± 4.18	**93.60 ± 1.26**
NSL-KDD (U2R)	1	72.28 ± 8.95	96.66 ± 0.44	**96.73 ± 0.87**
CIC-IDS 2017 (Heartbleed)	1	84.11 ± 4.41	97.50 ± 3.92	**98.52 ± 1.56**
CIC-IDS 2017 (Infiltration)	1	**81.06 ± 1.57**	62.78 ± 8.20	65.26 ± 7.87
NSL-KDD (R2L)	5	80.48 ± 6.17	**93.85 ± 1.75**	93.54 ± 1.67
NSL-KDD (U2R)	5	86.95 ± 4.28	96.99 ± 0.36	**97.20 ± 0.58**
CIC-IDS 2017 (Heartbleed)	5	92.47 ± 2.45	99.09 ± 1.63	**99.65 ± 0.50**
CIC-IDS 2017 (Infiltration)	5	**90.65 ± 1.18**	73.88 ± 8.63	77.02 ± 9.27

Table IV
F1-Score of FSL-based models in 1-Shot and 5-Shot Scenarios

Dataset (Attack Class)	Shot	MatchingNets	ProtoNets	ProtoKNN
NSL-KDD (R2L)	1	53.37 ± 1.45	87.66 ± 4.02	**88.53 ± 1.29**
NSL-KDD (U2R)	1	62.35 ± 4.65	91.79 ± 1.05	**92.67 ± 1.29**
CIC-IDS 2017 (Heartbleed)	1	70.69 ± 5.12	91.70 ± 7.43	**93.57 ± 4.48**
CIC-IDS 2017 (Infiltration)	1	**69.93 ± 3.65**	61.23 ± 5.24	60.88 ± 4.15
NSL-KDD (R2L)	5	63.58 ± 5.52	89.62 ± 1.07	**89.97 ± 1.06**
NSL-KDD (U2R)	5	74.23 ± 5.30	92.78 ± 0.65	**93.52 ± 0.88**
CIC-IDS 2017 (Heartbleed)	5	73.49 ± 9.42	96.34 ± 3.53	**97.53 ± 1.64**
CIC-IDS 2017 (Infiltration)	5	**82.35 ± 3.01**	72.03 ± 6.38	72.61 ± 6.58

5.3.2 Influence of Hyperparameter α: The hyperparameter α is a core component of ProtoKNN, governing the equilibrium between global class statistics and local manifold learning. As illustrated in Figure 2, the AUC trajectories for both R2L and U2R achieve their respective peaks at $\alpha = 0.5(0.936)$ and $\alpha = 0.8(0.970)$. These optimal points demonstrate that integrating local neighborhood structures with global prototypes is essential for mitigating centroid misalignment in minority classes.



(a) R2L AUC Trend (b) U2R AUC Trend

Figure 2. Influence of the hyperparameter α on R2L and U2R

Notably, the performance degrades significantly as α approaches 1.0, where the model relies exclusively on local manifolds. This sharp decline reinforces the necessity of the global prototype as a crucial regularizing reference, preventing the model from overfitting to local noise or outliers within the embedding space. Thus, a balanced α ensures that ProtoKNN captures both the general distribution and the intricate local geometry of complex attack patterns. While a fixed alpha demonstrates the effectiveness of our hybrid approach, these results suggest that a dynamic tuning strategy, which could adapt the balance based on class-specific geometries, remains a promising avenue for future research.

5.3.3 Visualization of Embedding spaces: The impact of ProtoKNN on the learned representation of the R2L class is illustrated in Figure 3. In ProtoNets (left), the

rigid spherical assumption often leads to overlapping regions where outliers at the tails of the distribution are misclassified. In contrast, ProtoKNN (right) leverages the local KNN component to allow the feature points to form elongated chains and clusters. This flexibility enables the decision boundary to adapt to the anisotropic distribution of the R2L manifold, resulting in a more discernible separation between normal traffic and attack instances, which directly reduces the misclassification of outliers and yields higher detection performance.

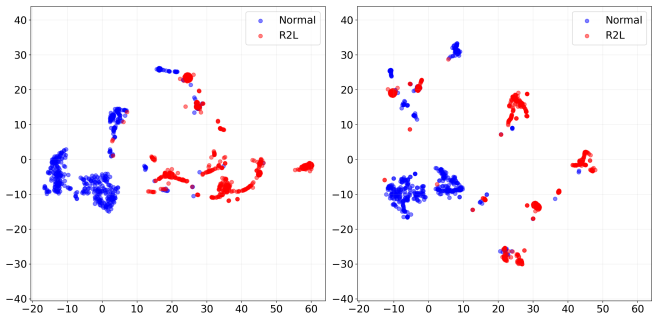


Figure 3. Embedding space of ProtoNets (left) vs. ProtoKNN at $\alpha = 0.5$ (right) on R2L.

Similarly, this refinement is further evidenced in the extremely scarce U2R class (Figure 4). With a higher local weight ($\alpha = 0.8$), ProtoKNN effectively traces the fragmented densities of the attack manifold, whereas ProtoNets struggles to encompass scattered samples within a single centroid. By capturing misaligned outliers that are globally distant but locally consistent, the model confirms its ability to relax spherical constraints in favor of more complex, arbitrary attack geometries.

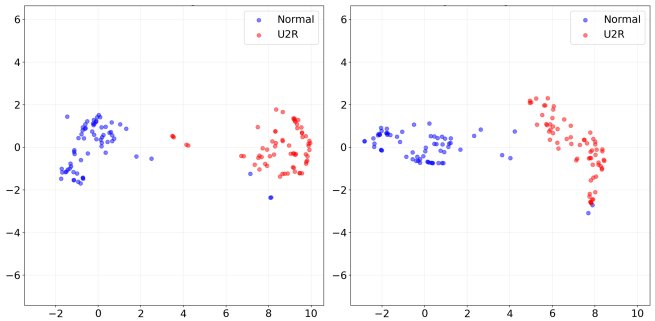


Figure 4. Embedding space of ProtoNets (left) vs. ProtoKNN at $\alpha = 0.8$ (right) on U2R.

In summary, the experimental results confirm that ProtoKNN effectively overcomes the limitations of rigid centroid-based representations by integrating global prototypes with local geometric structures. By using a balanced weight of $\alpha = 0.5$, the model consistently yields superior AUC and F1-scores, particularly in 1-shot scenarios. The relaxation of spherical constraints allows ProtoKNN to successfully recover misaligned outliers and adapt to the anisotropic distributions of rare attack classes. Furthermore, while the method requires distance calculations to all support instances, the computational overhead remains minimal due to the extremely small support sets, ensuring its efficiency under extreme data scarcity.

6 Conclusion

In this study, we identified and addressed the Centroid Misalignment problem in few-shot attack detection. While traditional ProtoNets rely on the restrictive assumption that attack classes form spherical clusters, our proposed model overcomes this limitation by integrating local KNN-based metrics with global class representations. This dual-metric strategy enables the model to learn a flexible embedding space that simultaneously optimizes inter-class separability and intra-class manifold continuity.

Empirical evaluations on the NSL-KDD and CIC-IDS2017 datasets demonstrate that our model consistently outperforms standard meta-learning methods. By employing a balanced configuration ($\alpha = 0.5$), the method successfully recovers misaligned outliers in minority classes like R2L and U2R, while significantly improving detection for high-variance intrusions such as Infiltration. These results confirm that relaxing rigid geometric constraints in favor of local neighborhood affinity provides a more robust and reliable decision boundary. Ultimately, our method strategy offers a superior solution for identifying rare and sophisticated cyberattacks within complex, data-scarce network environments.

References

[1] S. K. K. Nandiraju, S. K. Chundru, M. S. V. Tyagadurgam, V. N. Gangineni, S. Pabbineedi, and A. B. Kakani, "Enhancing cybersecurity: Zero-day attack detection in network traffic with deep learning model," *Asian Journal of Research in Computer Science*, vol. 18, no. 7, pp. 262–273, 2025.

[2] R. Ahmad, I. Alsmadi, W. Alhamdani, and L. Tawalbeh, "Zero-day attack detection: a systematic literature review," *Artificial Intelligence Review*, vol. 56, no. 10, pp. 10 733–10 811, 2023.

[3] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2010, pp. 305–316.

[4] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & Security*, vol. 28, no. 1-2, pp. 18–28, 2009.

[5] A. Alshamrani, Y.-W. Chow, W. Susilo, J. Rosli, and K. Brewer, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 120, pp. 1–13, 2018.

[6] J. Chen, C. Wang, Y. Hong, R. Mi, L.-J. Zhang, Y. Wu, H. Wang, and Y. Zhou, "A survey on anomaly detection with few-shot learning," in *Proceedings of the International Conference on Cognitive Computing*. Springer, 2024, pp. 34–50.

[7] Y. A. Jerusha, S. S. Ibrahim, and V. Varadharajan, "A Novel Semantic Driven Meta-Learning Model for Rare Attack Detection," *IEEE Access*, 2025.

[8] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proceedings of the IEEE Symposium on Computational*

Intelligence for Security and Defense Applications (CISDA), 2009, pp. 1–6.
[9] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: methods, systems and tools," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.
[10] 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 (NIPS)*, 2017, pp. 4080–4090.
[11] 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 (NIPS)*, 2016, pp. 3630–3638.
[12] 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 (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.
[13] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[14] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new dataset for cyber security in the cloud," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, 2018, pp. 403–412.
[15] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1199–1208.
[16] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 1126–1135.
[17] C. Xu, J. Shen, X. Du, and F. Zhang, "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.
[18] Y. Yu and N. Bian, "An intrusion detection method using few-shot learning," *IEEE Access*, vol. 8, pp. 49 730–49 740, 2020.
[19] V. L. Cao, T. M. 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.
[20] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.
[21] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 5, pp. 1–12, 2019.
[22] W. W. Lo, S. Layeghy, M. Sarhan, M. Gallagher, and M. Portmann, "E-GraphSAGE: A graph neural network based intrusion detection system for IoT," in *Proceedings of the IEEE/IFIP Network Operations and Management Symposium (NOMS)*, 2022, pp. 1–9.

Le Hoan Hoang is currently pursuing his Master's degree at Le Quy Don Technical University. His research interests include cybersecurity, few-shot learning, and network intrusion detection systems. Email: hoanglehoan95@gmail.com

Manh Tuan Nguyen received the B.Sc. and M.Sc. degree in 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.

Van Loi Cao received the B.Sc. and M.Sc. degree in computer science from Le Quy Don Technical University (LQDTU), Hanoi, Vietnam, and the Ph.D degree from University College Dublin (UCD), Dublin, Ireland. He is currently the Deputy Head of Information Security Department, the Faculty of Information Technology, LQDTU. His current research interests include Deep Learning, Meta-learning, Anomaly Detection, IoT Security, and Information Security. Email: loi.cao@lqdtu.edu.vn.