APMTH 220 Geometric Methods for Machine Learning **Spring 2024**

Malicious Traffic Detection Using Manifold Learning

Name: Rama Edlabadkar *Submission Date: 3rd May 2024*

1 Abstract

Recent advancements in cybersecurity underscore the critical role of intrusion detection systems (IDS) in safeguarding networks against evolving threats. While machine learning holds promise for enhancing IDS capabilities, existing methods often struggle in unsupervised settings where labeled attack samples are scarce. This research proposes a novel approach grounded in interpretable manifold learning techniques to address this challenge. By leveraging manifold learning, lowdimensional representations of data are constructed to help clustering of normal samples, enabling robust identification of attack traffic detection. Experimental results on CICIDS 2017 data demonstrate significant improvements in clustering performance with manifold learning, particularly with Latent Map Gaussian Processes (LMGP) and Autoencoders, indicating their efficacy in enhancing cyber attack detection.

2 Introduction

In the digital age, the relentless evolution of cyber threats necessitates continuous advancements in cybersecurity measures. Intrusion Detection Systems (IDS) have emerged as a cornerstone in the architecture of network security, effectively monitoring and responding to potential threats within a network. Despite significant progress, the detection capabilities of IDS remain challenged by the growing complexity and sophistication of cyber attacks. A critical limitation arises in scenarios lacking sufficient labeled data, which are essential for training supervised machine learning models.

The field of attack detection in network traffic has predominantly relied on supervised and semi-supervised methods, which perform well when labeled data is abundant. Supervised techniques, such as neural networks [1], Gaussian mixture models [2], and kernel support vector machines [3], have demonstrated efficacy in distinguishing between normal and malicious activities by learning from labeled examples of both classes. Semi-supervised methods, while requiring fewer labels, still depend on some prior knowledge of the normal data to function effectively [4-5]. However, these approaches are less viable in environments where labeling data is impractical or impossible, such as in the detection of novel or evolving threats. This gap has spurred interest in unsupervised techniques, which do not require labeled data and are adaptable to varied and changing environments. Traditional unsupervised methods, such as clustering and nearest-neighbor based algorithms, often struggle with high-dimensional data, facing issues of scalability and computational efficiency [6-8]. For instance, the Neighbour-based methods used in attack detection rely on spatial information but face limitations with high-dimensional data due to quadratic computational costs and potential inaccuracies. The Statistical methods build probabilistic models but struggle to scale well in high dimensions, hindering their effectiveness. Deep learning [9] is gaining popularity for fault and anomaly detection due to its ability to extract features from normal data. However, it's more suitable for supervised or semi-supervised scenarios and not in an unsupervised setting. Moreover, while deep learning approaches have shown promising results in many domains of anomaly detection, their reliance on large amounts of labeled data for optimal performance limits their utility in unsupervised settings [1].

Recent research has pointed towards manifold learning as a promising avenue to overcome these challenges. Manifold learning techniques, such as Latent Map Gaussian Processes (LMGP) and Autoen-

coders, aim to project high-dimensional data into a lower-dimensional space, maintaining the intrinsic geometric properties of the data [2]. This reduction not only alleviates the curse of dimensionality but also enhances the clustering of normal samples, thereby improving the detection of anomalies. The CICIDS2017 dataset, a comprehensive collection of real-world traffic data including labeled network flows, provides an ideal testbed for validating the efficacy of these methods [4].
This research builds on these foundational insights and proposes a novel approach employing manifold learning for the unsupervised detection of cyber threats. Our study leverages the latest advancements in manifold learning to develop models that are not only computationally efficient but also capable of handling the complexities of high-dimensional network data. By adapting these techniques to the specific challenges of network security, we aim to create a robust framework for IDS that can effectively identify and respond to emerging threats without the need for extensive labeled data.

3 Background

Despite significant advancements, current intrusion detection methods face challenges in unsupervised settings, where the absence of labeled attack samples hinders performance. Traditional approaches, such as neighbor-based and statistical methods, encounter scalability issues and may struggle to capture complex data distributions. Deep learning techniques, while promising, are primarily suited for supervised or semisupervised scenarios, limiting their applicability in unsupervised settings. To address these limitation of the existing techniques, this research advocates for the adoption of interpretable manifold learning techniques like Multi-dimensional Scaling, Isomap, LLE, Spectral Embedding, latent map Gaussian processes and Autoencoders which project high-dimensional data into a lower-dimensional space by preserving specific relationship between data points and can be used to capture different aspects of the data for attack detection.

3.1 Overview of Manifold Learning

Manifold Learning is a family of techniques used to reduce the dimensionality of large datasets while preserving their intrinsic properties. Unlike linear dimensionality reduction methods such as Principal Component Analysis (PCA), manifold learning is particularly adept at uncovering the underlying nonlinear structures within complex datasets. It achieves this by assuming that the high-dimensional data points are samples from a low-dimensional manifold embedded within the high-dimensional space. The key goal is to uncover this low-dimensional embedding which retains some meaningful properties of the original data, such as distances or local data densities.

3.2 Key Manifold Learning Techniques in Cybersecurity

- **Multi-dimensional Scaling (MDS):** aims to find a low-dimensional representation of data while preserving the pairwise distances between data points. One of the formulations of MDS is:

$$\mathbf{X} = \arg\min_{\mathbf{X}} \sum_{i,j} (d_{ij} - ||\mathbf{x}_i - \mathbf{x}_j||)^2$$

where $\mathbf{X}$ is the matrix of low-dimensional representations, d_{ij} is the pairwise distance between points i and j in the original high-dimensional space, $\mathbf{x}_i$ and $\mathbf{x}_j$ are the corresponding low-dimensional representations and $|| \cdot ||$ denotes the Euclidean norm.

- **Isomap:** aims to preserve the geodesic distances between all pairs of data points. One of the formulations of Isomap is:

$$\mathbf{D}_{geo} = (\mathbf{D}_{shortest})^2$$

where $\mathbf{D}_{geo}$ is the matrix of geodesic distances and $\mathbf{D}_{shortest}$ is the matrix of shortest path distances.

- **Locally Linear Embedding (LLE):** aims to preserve local relationships between neighboring data points. One of the formulations of LLE is:

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_i ||\mathbf{x}_i - \sum_j w_{ij}\mathbf{x}_j||^2$$

where $\hat{\mathbf{W}}$ is the weight matrix, $\mathbf{W}$ contains the weights w_{ij} indicating the contribution of point j to the reconstruction of point i and $|| \cdot ||$ denotes the Euclidean norm.

- **Spectral Embedding:** utilizes the spectral decomposition of a graph Laplacian matrix. One of the formulations of Spectral Embedding is:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}$$

where: $\mathbf{L}$ is the Laplacian matrix, $\mathbf{D}$ is the diagonal degree matrix and $\mathbf{A}$ is the adjacency matrix.

- **Latent Map Gaussian Processes** model relationships using Gaussian processes in a latent space. The formulations could be the Gaussian process regression equation:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

where $f(\mathbf{x})$ is the latent function, $\mathcal{GP}$ denotes a Gaussian process and $k(\mathbf{x}, \mathbf{x}')$ is a kernel function.

- **Autoencoders:** use neural network architectures to encode-decode the data. One of the formulations for an autoencoder is the reconstruction loss:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = ||\mathbf{x} - \hat{\mathbf{x}}||^2$$

where $\mathbf{x}$ is the input data, $\hat{\mathbf{x}}$ is the reconstructed output and $|| \cdot ||$ denotes the Euclidean norm.

Evaluation metrics are essential tools in assessing the performance of unsupervised machine learning algorithms, particularly clustering algorithms. The Calinski-Harabasz Index, Davies-Bouldin Index, and Silhouette Score [11] provide quantitative measures of how well the algorithm has grouped the data points into meaningful clusters without the need for labeled data.

- **Silhouette Score:** assesses the coherence of clusters by measuring the similarity of data points within clusters compared to those in other clusters. It produces a numerical representation ranging from -1 to 1, where a value of 1 indicates distinct clusters, -1 signifies incorrect clustering, and 0 suggests indistinct clusters. The Silhouette Score S for a single data point i is given by:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where: $a(i)$ is the average distance from i to other points in the same cluster, $b(i)$ is the smallest average distance from i to points in a different cluster.

- **Calinski-Harabasz Index:** evaluates cluster quality by quantifying the ratio of dispersion between clusters to dispersion within clusters. A higher index score indicates well-separated clusters, as it reflects greater differences in the squared distances between data points within clusters and those between clusters. The Calinski-Harabasz Index is computed as:

$$CH = \frac{B(k)}{W(k)} \times \frac{N - k}{k - 1}$$

where $B(k)$ is the between-cluster dispersion (sum of squared distances between cluster centroids), $W(k)$ is the within-cluster dispersion (sum of squared distances of points to their respective cluster centroids), N is the total number of data points and k is the number of clusters.

- **Davies-Bouldin Index:** evaluates clustering effectiveness by computing the average similarity between each cluster and its most similar one. It measures the ratio of within-cluster distances to between-cluster distances, aiming for clusters that are both distinct and well-separated. Unlike the other metrics, a lower Davies-Bouldin Index score signifies better clustering performance, indicating clearer boundaries between clusters and less overlap.
The Davies-Bouldin Index for a set of k clusters is given by:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{sim(C_i, C_j) + sim(C_j, C_i)}{d(C_i) + d(C_j)} \right)$$

where $sim(C_i, C_j)$ is a measure of similarity between clusters C_i and C_j (e.g., average distance between centroids or average pairwise distance), $d(C_i)$ is a measure of within-cluster dispersion for cluster C_i (e.g., average distance from each point in C_i to its centroid).

To test the quality of the manifold learning, Q-score has been used. It is a measure used to evaluate the effectiveness of manifold learning techniques in preserving the local and global structure of the data. Local Q score measures how well our manifold learning technique preserves the local relationships between neighboring data points. It is typically computed based on pairwise distances or similarities between points in the original high-dimensional space and the corresponding distances or similarities in the reduced-dimensional space. Global Q score evaluates how well a manifold learning technique preserves the global structure or overall relationships in the data. It considers the distribution of points in the reduced-dimensional space compared to the original space.

4 Proposed approach

To address the shortcomings of existing methods, the research proposes a novel approach grounded in interpretable manifold learning techniques to detect attacks in an unsupervised setting more efficiently. By leveraging manifold learning, low-dimensional data representations are constructed which facilitate the automatic clustering of normal samples, allowing effective identification of attacks. Subsequently, k-means clustering is applied to group the encoded latent points based on their positions in the manifold, facilitating the differentiation between normal and attacked samples, while identifying outliers exhibiting significant deviations from expected behavior. The research focuses on the following Manifold Learning algorithms -
1. **Multi-dimensional Scaling (MDS)**: preserves the pairwise distances between data points, capturing the underlying structure and seeks to place each object in a low-dimensional space such that the between-object distances are preserved as well as possible. It starts with a matrix of item-item similarities, then assigns a location to each item in N-dimensional space. For IDS, MDS help us to understand the similarity relationships between various network traffic samples, which aids in identifying attack patterns.
2. **Isomap**: extends MDS by incorporating geodesic distances measured along the manifold and is particularl useful for capturing the true geometric distances in datasets that lie on curved manifolds, rather than just 'as the crow flies' distances. In this work, Isomap unravels complex attack patterns that traditional methods might miss due to their non-linear nature.
3. **Locally Linear Embedding (LLE)**: works on the premise that each data point can be linearly reconstructed from its neighbors, preserving the local properties of the data. It's advantageous for attack detection as it helps in retaining the local data relationships even in the reduced space, which is crucial

for spotting subtle deviations indicative of cyber threats.
4. **Spectral Embedding**: utilizes the spectral decomposition of a graph Laplacian matrix to embed data points into a lower-dimensional space. It operates by representing the data as a graph, where nodes represent data points and edges encode relationships between them. It is particularly effective in exploring the structure of the data that forms clusters, which can correspond to different types of network behavior, including potential intrusions.
5. **Latent Map Gaussian Processes (LMGP)**: is a probabilistic approach that models the relationship between input and output data using Gaussian processes in a latent space. In this work, LMGP help in constructing a low-dimensional manifold that captures the variability in the data, aiding in the identification of normal and malicious behaviour.
6. **Autoencoders**: are neural networks designed to learn an efficient encoding of the input data in an unsupervised manner. By training to minimize the difference between the input and its reconstruction, autoencoders learn to capture the most salient features of the data. In IDS, autoencoders detect the malicious attacks by evaluating the reconstruction error.

The research work has been using the 'CICIDS2017 Dataset'[4] which contains benign and the most up-to-date common attacks, which resembles the true real-world data. The dataset includes the most common attacks based on the 2016 McAfee report, such as Web based, Brute force, DoS, DDoS, Infiltration, Heartbleed, Bot, and Scan covered in this dataset. This research uses the data which has benign and PortScan attacks. Manifold learning techniques like autoencoder and latent map Gaussian processes model requires higher computational resources. Thus, considering the time and computational limitation, 5000 data points were randomly sampled for experimenting with different techniques.

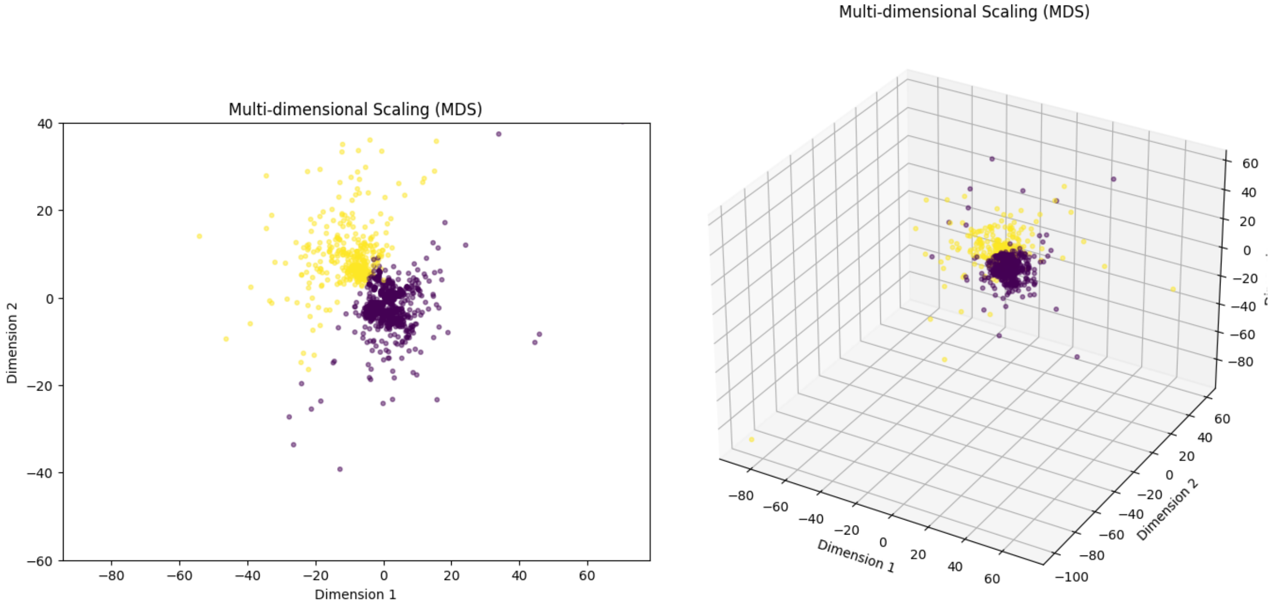
5 Results

In this research work, all the experiments are performed on data which has benign and PortScan attacks with 5000 randomly selected data points. The MDS, Isomap, LLE, Spectral Embedding, Autoencoder, and latent map Gaussian processes model are implemented for reducing the high dimensional data into latent space in a interpretable way. The Kmeans and Hierarchical Agglomerative Cluster(HAC) are subsequentively applied. Figures 2,3 show the latent space representation of data using different manifold learning techniques.

With just the randomly selected 5000 data points, it required around 5-6 hours to train the GPLVM model using GPU. However increasing the GPU capacity would significantly reduce the training time. Tables 1,2 show the results demonstrating the performance of the Kmeans and HAC (Hierarchical Agglomerative Clustering) using different Manifold Learning techniques. The performance from both Kmeans and HAC are pretty consistent. However, it is observed that Kmeans performs slightly better than HAC on the latent space data points.

Table 1: Performance of the Kmeans with Manifold Learning

Method	Silhouette	Davies-Bouldin	Calinski-Harabasz
Without Manifold	0.719	1.354	1303.728
Spectral Embedding	0.373	1.427	2174.072
MDS	0.682	1.003	1683.037
ISOMAP	0.641	0.782	2545.307
LLE	0.673	0.536	1721.240
Autoencoder	0.769	0.571	4271.125
GPLVM	0.999	0.139	32971.644



(a) MDS 2D

(b) MDS 3D

Figure 1: Multi-dimensional Scaling Latent Space Representation

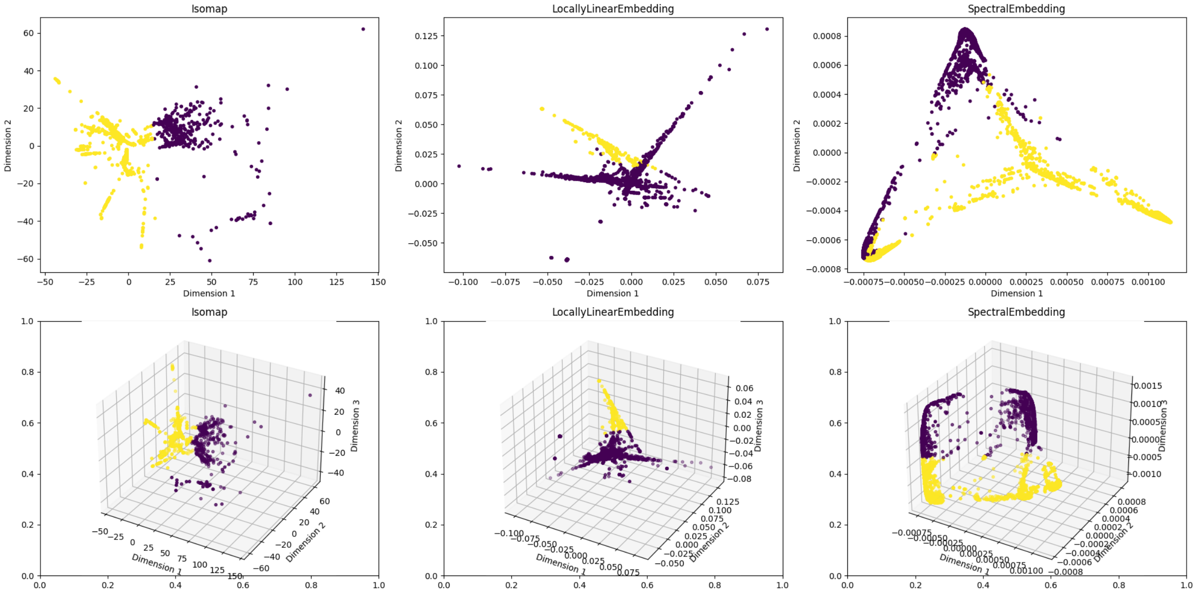
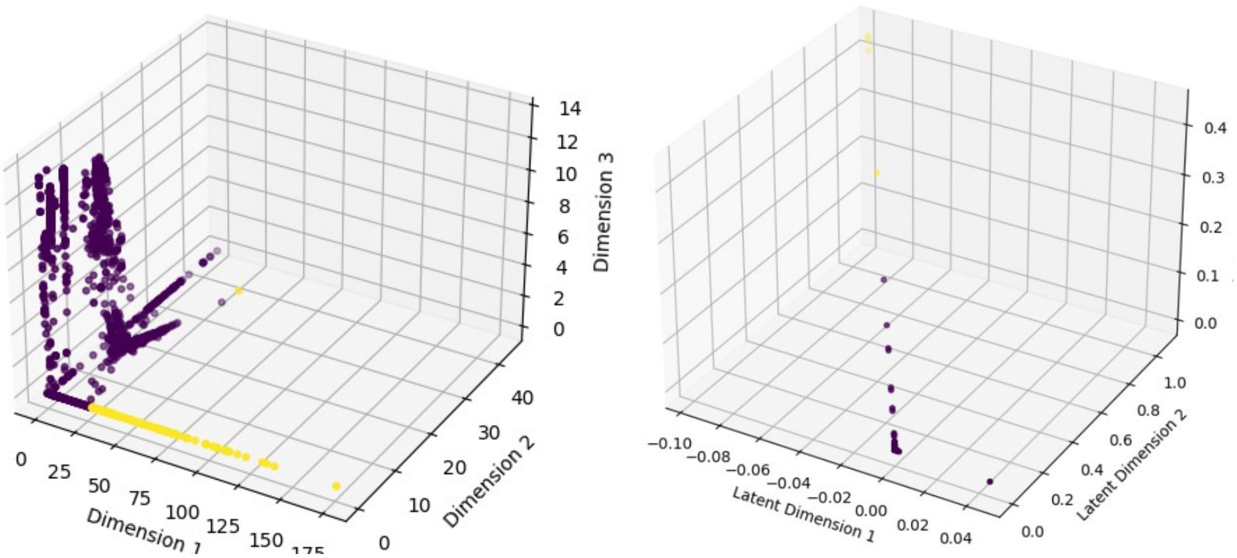


Figure 2: Latent Representation of Isomap, Locally Linear Embedding (LLE), Spectral Embedding



(a) Auto-encoder

(b) GPLVM

Figure 3: Latent Representation of Autoencoder and Gaussian Process Latent Variable Model (GPLVM)

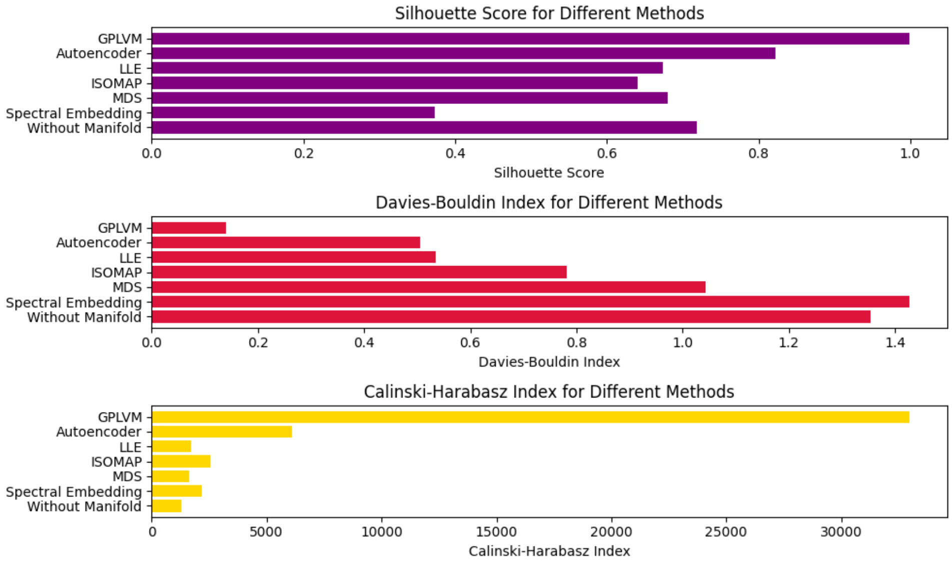


Figure 4: Comparision Plot of Kmeans Clustering using different metric

Table 2: Performance of the HAC with Manifold Learning

Method	Silhouette	Davies-Bouldin	Calinski-Harabasz
Without Manifold	0.676	1.413	1276.093
Spectral Embedding	0.367	1.312	2099.666
MDS	0.649	1.060	1533.639
ISOMAP	0.642	0.779	2543.644
LLE	0.691	0.536	1733.573
Autoencoder	0.728	0.677	3857.542
GPLVM	0.999	0.139	32971.644

Figure 4 show the performance comparison plot of the Kmeans clustering indicating significant improvement in the performance when we use GPLVM and Autoencoder compared to the case where we do not use manifold learning. Th reason could be because while Multi-dimensional Scaling, Isomap, LLE, Spectral Embedding preserve certain specific characteristics/features, GPLVM and Autoencoder consider the overall features into consideration. GPLVM model relationships using Gaussian processes in a latent space, capturing the data variability and Autoencoder is trained to minimize the reconstruction error, thus effectively learning a compressed representation of the data. Therefore these two techniques are observed to be performing the best amongst all. Table 3 shows the Q-scores of all the manifold learning techniques implemented, indicating the quality of the manifold learning. It has the (local, global) Q scores respectively. Both the global and local features are best described by MDS followed by Autoencoders.

Table 3: Quality of the Manifold Learning

Method	Q-Score
Spectral Embedding	(0.558, 0.711)
MDS	(0.733, 0.911)
ISOMAP	(0.649, 0.808)
LLE	(0.319, 0.610)
Autoencoder	(0.673, 0.859)
GPLVM	(0.395, 0.897)

6 Discussion and Conclusion

The Kmeans and HAC clustering applied after using various manifold learning algorithms was compared with the clustering without manifold learning. We can observe a significant improvement in the clustering performance using manifold learning especially Latent map Gaussian processes, and Autoencoders. We can observe that Silhouette Score and Calinski-Harabasz Index is much higher for clustering with LMGP and Autoencoder compared to without manifold learning. Davies-Bouldin Index is much lower with manifold learning thus suggesting improved performance. The overall result indicate that manifold learning techniques are effective in improving the performance of clustering of cyber attack data and thus aids in identifying cyber-attacks.

The research work can be refined by using more data points. Additionally, this work can be extended for identifying different types of web/cyber attacks like Brute Force, XSS, SQL injection, DoS, DDoS, Infiltration, Heart-bleed, Bot besides distinguishing them from benign ones.

7 Reference

[1] Markou, M., Singh, S., ”Novelty detection: A review, part 2: Neural network based approaches,” Signal Processing, vol. 83, no. 12, pp. 2499–2521, 2003.

[2] Olson, C.C., Judd, K.P., Nichols, J.M., ”Manifold learning techniques for unsupervised anomaly detection,” Expert Systems with Applications, 2018.

[3] Schölkopf, B., Smola, A.J., ”Learning with Kernels,” MIT Press, Cambridge, MA, 2002.

[4] Sharafaldin, I., Lashkari, H., Ghorbani, A., ”Toward generating a new intrusion detection dataset and intrusion traffic characterization,” International Conference on Information Systems Security and Privacy, pp. 108–116, 2018.

[5] Fujimaki, R., Yairi, T., Machida, K., ”An approach to spacecraft anomaly detection problem using kernel feature space,” Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, IL, USA: ACM Press, 2005, pp. 401–410.

[6] Knorr, E.M., Ng, R.T., Tucakov, V., ”Distance-based outliers: algorithms and applications,” The VLDB Journal, vol. 8, pp. 237–253, 2000.

[7] Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., ”Lof: Identifying density-based local outliers,” Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA: ACM Press, 2000.

[8] Bezdek, J., Ehrlich, R., Full, W., ”Fcm: The fuzzy c-means clustering algorithm,” Computers and Geosciences, vol. 10, no. 2, pp. 191–203, 1984.

[9] Liu, X., Liu, J. Malicious traffic detection combined deep neural network with hierarchical attention mechanism. Sci Rep 11, 12363 (2021).

[10] Amin Yousefpour, Mehdi Shishehbor.Unsupervised Anomaly Detection via Nonlinear Manifold Learning. In: arXiv:2306.09441 (2023).
[11] https://www.kdnuggets.com/2023/04/exploring-unsupervised-learning-metrics.html