Efficient Network Intrusion Detection using Grassmann Manifold

Keon Oh Kim, Dong Uk Kim, Ki Tae Kim, Sun-Moo Kang, Eui-Nam Huh and Choong Seon Hong
Department of Computer Science and Engineering, Kyung Hee University
Yongin, Republic of Korea
{keonoh, g9896, glideslope, etxkang, johnhuh, cshong}@khu.ac.kr

***Abstract*—Recently, the increasing complexity and volume of network traffic have posed significant challenges for conventional Network Intrusion Detection Systems (NIDS). Traditional NIDS often rely on predefined patterns and la making them inadequate for handling dynamic, high-dimensional network traffic. To address th studies have investigated the integration of m into NIDS. However, existing approaches are c on supervised learning models that are inefficie reliant on labeled datasets. Moreover, these mo process high-dimensional data and fail to gener attacks. To this extent, this paper explores an proach by introducing a framework that levera manifolds to overcome these limitations. Grassm which represent the set of all possible subspa dimension, provide a robust geometric foundatio network traffic data. Leveraging this structure method effectively captures abnormal patterns information, enabling precise identification of an iors. Comprehensive experiments using the NS demonstrate that the proposed GrassmannPCA only improves detection accuracy, precision, an also achieves lower false alarm rates and reduced cost. The proposed GrassmannPCA framework EuclideanPCA, achieving an accuracy of 82.97 of 88.32%, an F1-score of 84.37% , and a sig false alarm rate of 14.12%, while also reducing t 1.12%.**

***Index Terms*—anomaly detection, network intrusion detection systems(NIDS), grassman manifold, deep learning**

I. Introduction

Network intrusion detection systems (NIDS) play a critical role in safeguarding the integrity and security of modern networked environments. However, the increasing complexity and dynamism of network traffic, coupled with the rise of sophisticated cyber threats, present significant challenges for traditional detection methodologies. Conventional NIDS often rely on supervised learning models, which require extensive labeled datasets that are costly and time-consuming to generate [1]. Moreover, these systems tend to struggle when faced with high-dimensional data and fail to generalize effectively to novel or previously unseen attacks, making them less adaptable to rapidly evolving network environments. The dependence on predefined patterns limits their ability to detect emerging

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under (No.RS-2024-00398993) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation), and supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00207816), (No. RS-2024-00352423). *Dr. CS Hong is the corresponding author.

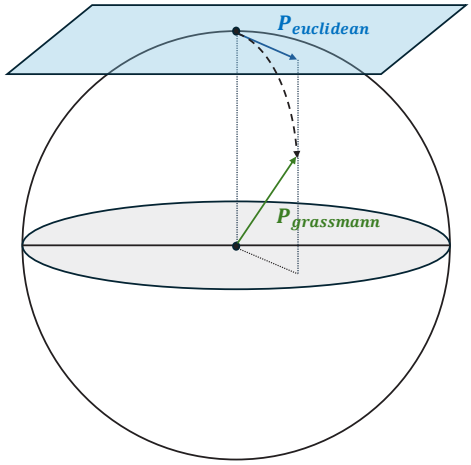


Fig. 1: Example of mapping Euclidean space vector to Grassmann vector. The figure depicts two components, $P_{euclidean}$ and $P_{grassmann}$, representing projections in high-dimensional and low-dimensional spaces respectively. $P_{euclidean}$ corresponds to the high-dimensional representation of the data, while $P_{grassmann}$ illustrates its lower-dimensional counterpart after being projected onto a Grassmann manifold. This mapping preserves critical structural information while reducing dimensionality, which aids in effective anomaly detection by isolating subtle deviations.

To overcome these limitations, researchers have explored alternative approaches that leverage advanced mathematical structures to model network traffic data more efficiently. One promising direction involves the use of Grassmann manifolds, which represent the set of all possible subspaces of a given dimension within a larger space. Figure 1 illustrates the mapping of data from Euclidean space to the Grassmann manifold, highlighting the transition from a high-dimensional representation $P_{\text{euclidean}}$ to a lower-dimensional subspace $P_{\text{grassmann}}$. This geometric framework is well-suited for capturing complex correlations and variances in high-dimensional network data, providing a means to project and analyze the data in a lower-dimensional space without losing critical structural

information. By mapping data onto Grassmann manifolds, the inherent properties of the manifold space can be used to isolate and detect subtle deviations that would otherwise go unnoticed using traditional techniques.

In this paper, we proposes a novel intrusion detection framework that utilizes Grassmann manifold techniques to enhance the detection and mitigation of network anomalies. Overall contribution can be summarized as below:

- The framework projects network traffic data onto these manifolds, enabling more efficient modeling and analysis while reducing the computational and memory overhead commonly associated with high-dimensional data processing.
- By leveraging the geometric properties of Grassmann manifolds, the proposed approach captures hidden patterns and complex data structures, which significantly improves the detection of a wide range of network intrusions.
- Through comprehensive experiments, the framework demonstrates its ability to identify anomalies with high accuracy and robustness, making it a promising solution for real-world network security applications.
- The results highlight its potential to address the limitations of existing NIDS, providing a path forward for more effective and scalable intrusion detection systems.

II. Related Works

Network Intrusion Detection Systems (NIDS) have seen significant advancements with the integration of machine learning and deep learning techniques. Principal Component Analysis (PCA) is one of the foundational methods employed for dimensionality reduction in NIDS. By projecting high-dimensional data into a lower-dimensional subspace, PCA preserves key variances to isolate anomalies based on reconstruction errors [3]. For example, Brauckhoff et al. [4] demonstrated the use of PCA for analyzing network traffic anomalies by projecting flow features and detecting deviations. Despite its simplicity, traditional PCA operates in Euclidean space, which limits its ability to model complex geometric structures inherent in high-dimensional network traffic. This limitation reduces its effectiveness in dynamic environments where attack patterns can be highly nonlinear and varied.

To address these challenges, recent years have seen the rise of deep learning-based methods in NIDS. Techniques such as Long Short-Term Memory (LSTM) networks and Transformer-based architectures excel in capturing temporal and contextual dependencies in sequential network data [5]. Additionally, Generative Adversarial Networks (GANs) have gained attention for their ability to generate synthetic attack data, which enhances the robustness and generalization of anomaly detection models [6]. However, while these methods achieve high accuracy, they are computationally intensive and heavily reliant on labeled datasets, which are often expensive and time-consuming to obtain. Moreover, their black-box nature makes interpretability and real-time deployment challenging.

Recent innovations have focused on mathematical frameworks like Grassmann manifolds to overcome the limitations of both traditional and deep learning-based methods. Grassmann manifolds represent the set of all possible subspaces of a given dimension within a larger space, providing a robust geometric foundation for analyzing high-dimensional data. Unlike PCA in Euclidean space, Grassmann manifold-based approaches effectively capture complex correlations and variances, enabling more precise detection of subtle deviations that are critical for anomaly detection. For instance, GrassmannPCA projects network traffic data onto these manifolds, allowing for efficient dimensionality reduction while preserving critical structural information.

A. Advantages of Grassmann Manifold

Unlike traditional NIDS that rely heavily on predefined patterns or supervised models constrained to Euclidean geometry, our proposed method leverages the Grassmann manifold to project high-dimensional network traffic data into a subspace that captures critical geometric structures. This approach provides several key advantages:

- **Robust Representation**: The Grassmann manifold enables the modeling of network traffic data with inherent geometric awareness, preserving complex patterns and correlations often overlooked by Euclidean-based techniques.
- **Improved Generalization**: By utilizing the manifold's structure, the proposed method demonstrates superior ability to generalize to unseen attacks, addressing a significant limitation of existing supervised learning approaches.
- **Reduced Computational Overhead**: The Grassmann manifold framework facilitates efficient dimensionality reduction while maintaining essential information, leading to faster convergence and lower computational costs compared to traditional methods.
- **Enhanced Detection Metrics**: Through comprehensive experiments, the proposed approach achieves higher detection accuracy, precision, and F1-scores while maintaining a lower false alarm rate, emphasizing its effectiveness in real-world scenarios.

By combining these advantages, the proposed GrassmannPCA framework in IIIrepresents a significant advancement in NIDS, leveraging manifold properties to model network traffic more efficiently and accurately. Compared to traditional EuclideanPCA, GrassmannPCA demonstrates superior performance in detecting anomalies, particularly in scenarios involving high-dimensional and dynamic network data. With its computational efficiency, robust geometric representation, and reduced reliance on labeled data, GrassmannPCA offers a scalable and adaptable solution for modern network environments, providing a promising alternative for next-generation NIDS while complementing the strengths of deep learning approaches.

III. Proposed Method

A. Mathematical Formulation

Given a set of high-dimensional data points $X \in \mathbb{R}^{n \times d}$, where d is the feature dimension and n is the number of samples, we aim to project X onto a Grassmann manifold $\mathcal{G}(p, d)$, where p is the subspace dimension. The Grassmann manifold $\mathcal{G}(p, d)$ represents the space of p-dimensional subspaces in $\mathbb{R}^d$.

To perform dimensional reduction on the Grassmann manifold, we first compute the covariance matrix $C = X^T X$. The projection matrix P is obtained through eigen decomposition, as shown in Eq. 1:

$$C = U \Lambda U^T, \tag{1}$$

where $U \in \mathbb{R}^{d \times p}$ are the eigenvectors corresponding to the top-p eigenvalues Λ. The projection $Y = XU$ then lies on $\mathcal{G}(p, d)$.

B. Algorithm

The two algorithms employed in the study are outlined below:

Algorithm 1 GrassmannPCA

Require: Dataset $X \in \mathbb{R}^{n \times d}$, Rank k, Learning rate η, Number of iterations T

1: Randomly initialize projection matrix $U \in \mathbb{R}^{d \times k}$ with orthonormal columns.
2: **for** $t = 1$ to T **do**
3: Compute the reconstruction error: $E = X - UU^T X$
4: Calculate the gradient: $\nabla_U = -2EX^T$
5: Project gradient onto tangent space: $\nabla_{Grassmann} = (I - UU^T)\nabla_U$
6: Update U using gradient descent: $U \leftarrow U - \eta \nabla_{Grassmann}$
7: Reorthogonalize U using QR decomposition: $U = QR(U)$
8: **end for**
9: **Return** the optimized projection matrix U

The **GrassmannPCA** algorithm is a specialized variant of PCA designed to optimize a low-dimensional subspace on the Grassmann manifold. Unlike traditional PCA, which optimizes in Euclidean space, GrassmannPCA enforces orthonormal constraints on the projection matrix and updates it using gradient descent projected onto the tangent space of the manifold. The algorithm starts by initializing a random orthonormal matrix and iteratively minimizes the reconstruction error by adjusting the projection matrix using a tangent space gradient. To maintain orthonormality, it reorthogonalizes the matrix at each step using QR decomposition. This process ensures that the optimized subspace remains a valid point on the Grassmann manifold. And the preserved subspace is used for accurate modeling. Hence, it allows to cluster subspaces which eventually leads to effective network intrusion detection.

The **EuclideanPCA** algorithm is a basic version of PCA optimized using Euclidean space properties. It aims to find a low-dimensional subspace by iteratively refining a projection matrix W to minimize the reconstruction error of the given dataset. The algorithm starts by randomly initializing W and uses gradient descent to update it over a specified number of iterations (T). At each iteration, the algorithm computes the reconstruction error $E = X - WW^T X$, which measures how well the current subspace defined by W approximates the original data. The gradient of the reconstruction error with respect to W is calculated as $\nabla_W = -2EX^T$, and W is updated using the learning rate η.

Algorithm 2 EuclideanPCA

Require: Dataset $X \in \mathbb{R}^{n \times d}$, Rank k, Learning rate η, Number of iterations T

1: Randomly initialize projection matrix $W \in \mathbb{R}^{d \times k}$
2: **for** $t = 1$ to T **do**
3: Compute the reconstruction error: $E = X - WW^T X$
4: Calculate the gradient: $\nabla_W = -2EX^T$
5: Update W using gradient descent: $W \leftarrow W - \eta \nabla_W$
6: **end for**
7: **Return** the optimized projection matrix W

GrassmannPCA and EuclideanPCA share a foundation in dimensionality reduction but differ in their geometric representations and optimization strategies:

- **GrassmannPCA**: Operates on the Grassmann manifold, enforcing orthonormal constraints and leveraging tangent space gradients for optimization. This enables the preservation of geometric structures in high-dimensional data.
- **EuclideanPCA**: Optimizes in Euclidean space, providing a simpler implementation but lacking the geometric awareness of GrassmannPCA, which can result in reduced detection capability for complex attack patterns.

IV. Experimental Setup

The proposed network intrusion detection framework was evaluated using the NSL-KDD dataset, a widely recognized benchmark in network security. Addressing issues of redundancy and class imbalance present in the original KDD Cup 1999 dataset, NSL-KDD offers a more reliable foundation for assessing intrusion detection systems. Comprising 41 features that reflect basic, content-based, and time-based characteristics of network traffic, the dataset includes labels categorizing traffic as normal or one of several attack types. GrassmannPCA and EuclideanPCA were independently applied to this dataset under a centralized training setup to distinguish normal from malicious traffic. To ensure consistency and unbiased evaluation, the data underwent z-score normalization for standardization and was split into balanced training and testing sets. Both algorithms leveraged PCA-based techniques to model the underlying structure of the data, with reconstruction errors used as the criterion for anomaly detection.

A. Dataset Description and Preprocessing

The NSL-KDD dataset [7] was chosen due to its widespread use in intrusion detection research and its improvements over

the original KDD Cup 1999 dataset, including the removal of redundant records and better class balance. While NSL-KDD provides a reliable benchmark.

The dataset contains 125,973 records and 41 features categorized into basic, content-based, and time-based characteristics, with labels indicating whether traffic is normal or an instance of one of 22 attack types. To ensure unbiased evaluation:

- The dataset was split into 70% for training and 30% for testing, maintaining a balanced distribution of normal and attack instances.
- Z-score normalization was applied to standardize feature values across all records.

B. Performance Metrics

The performance of the algorithms was measured using several key metrics:

- **Accuracy**: The ratio of correctly classified instances (both normal and attack) to the total number of instances.
- **Precision**: The proportion of true positive detections among all instances classified as anomalies, representing the model's ability to avoid false positives.
- **Recall**: The proportion of true positive anomalies detected out of all actual anomalies in the dataset, indicating the model's sensitivity.
- **F1-Score**: The harmonic mean of precision and recall, balancing detection sensitivity and specificity.
- **False Alarm Rate**: The proportion of normal instances incorrectly classified as attacks.
- **False Negative Rate**: The proportion of attack instances incorrectly classified as normal.

C. Rationale for Baseline Selection

Euclidean PCA was chosen as the baseline because it represents a fundamental and well-studied approach to dimensionality reduction in anomaly detection [8]. This choice allows for a clear demonstration of the benefits introduced by incorporating manifold geometry in the proposed GrassmannPCA method.

D. Results and Analysis

The results for the EuclideanPCA and GrassmannPCA algorithms are summarized below. GrassmannPCA demonstrated better overall performance compared to EuclideanPCA across multiple metrics, particularly in terms of precision, F1-score, and accuracy. This improvement is attributed to the Grassmann manifold's ability to preserve geometric structures in high-dimensional data, allowing for better differentiation between normal and malicious patterns. The results for EuclideanPCA and GrassmannPCA are summarized in Table I. GrassmannPCA demonstrated better overall performance across multiple metrics, particularly in terms of precision, F1-score, and accuracy. These improvements, though modest (e.g., a 2.17% increase in F1-score), highlight the effectiveness of Grassmann manifolds in capturing high-dimensional patterns while maintaining computational efficiency.

TABLE I: Performance Comparison on NSL-KDD Dataset

Metric	EuclideanPCA	GrassmannPCA
Training Time (s)	16,806.38	**16,618.74**
Accuracy	80.20%	**82.97%**
Precision	84.38%	**88.32%**
Recall	80.04%	**80.77%**
F1-Score	82.15%	**84.37%**
False Alarm Rate	19.59%	**14.12%**
False Negative Rate	19.96%	**19.23%**

Although the improvements in detection metrics are incremental, the reduced false alarm rate (-5.47%) underscores the robustness of GrassmannPCA in differentiating normal and malicious traffic. Additionally, its faster training time (-1.12%) indicates improved computational efficiency, which is critical for real-time NIDS deployment.

For the EuclideanPCA algorithm, the final test results indicated an accuracy of 80.20% with a precision of 84.38% and a recall of 80.04%. The F1-Score was 82.15%, while the false alarm rate and false negative rate were 19.59% and 19.96%, respectively. The total training time for the EuclideanPCA was approximately 16,806 seconds.

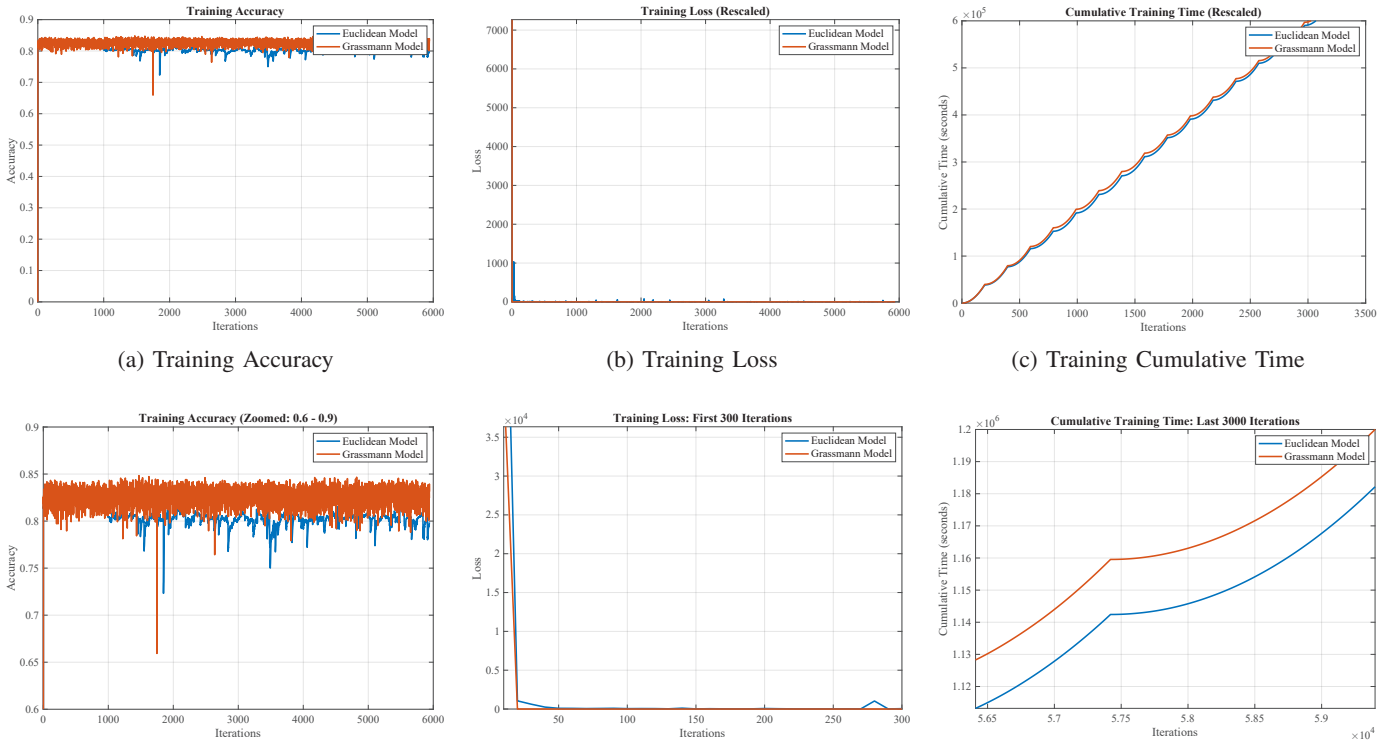
In contrast, the GrassmannPCA algorithm showed a higher overall performance with an accuracy of 82.97%, a precision of 88.32%, and a recall of 80.77%. The F1-Score for GrassmannPCA was 84.37%, and the false alarm rate was significantly lower at 14.12%, with a false negative rate of 19.23%. The GrassmannPCA model also required a shorter training time of 16,618 seconds, indicating a more efficient learning process.

The combined Fig.2 presents a detailed comparison of the training performance metrics for the Euclidean and Grassmann models. It consists of six subplots, providing insights into accuracy, loss, and cumulative time trends across the entire training process.

The training accuracy (2a) shows that the Grassmann model consistently achieves higher accuracy compared to the Euclidean model. The training loss (2b) reflects a faster convergence for the Grassmann model, evidenced by its steep decline in loss values. Meanwhile, the training cumulative time (2c) illustrates that the Grassmann model requires less computational time, showcasing its efficiency.

The zoomed training accuracy (2d) between 0.6 and 0.9 highlights the stability of the Grassmann model, as it maintains a smoother accuracy curve. The early-stage loss (2e) during the first 300 iterations shows a rapid decrease, confirming the Grassmann model's quicker adaptation to the training data. The cumulative time for the last 3000 iterations (2f) further emphasizes the Grassmann model's sustained efficiency even in the later stages of training.

Overall, the results show that the Grassmann model achieves better performance in terms of both accuracy and convergence speed while requiring less computational time compared to the Euclidean model. This suggests that the Grassmann manifold-based approach is more effective for handling high-dimensional network traffic data, making it a superior choice



(a) Training Accuracy (b) Training Loss (c) Training Cumulative Time

(d) Zoomed Training Accuracy (e) First 300 Iterations Loss (f) Last 3000 Iterations Cumulative Time

Fig. 2: Training Performance Metrics for Euclidean and Grassmann Models.

for network intrusion detection in complex environments.

The results demonstrate that the proposed GrassmannPCA method outperforms the Euclidean PCA baseline across multiple metrics. GrassmannPCA achieved a 2.17% improvement in F1-score and a 1.12% reduction in training time. While the improvements are modest, they highlight the effectiveness of Grassmann manifold geometry in capturing high-dimensional patterns.

The results underscore the following advantages of the GrassmannPCA approach:

- **Scalability and Efficiency**: GrassmannPCA achieves these improvements with reduced computational costs, making it well-suited for real-time applications.
- **Consistent Metric Improvements**: The approach demonstrates consistent gains across all metrics, including a significant 5.47% reduction in false alarm rate, which is critical for practical deployment [9].
- **Complementary Strengths**: While deep learning-based models like LSTMs and GNNs excel at sequential data modeling, GrassmannPCA complements these strengths by providing a geometrically informed representation of data.

V. Conclusion and Future Work

In this paper, we presented a comparative study of two network intrusion detection algorithms, GrassmannPCA and EuclideanPCA, using the NSL-KDD dataset. The results demonstrated that GrassmannPCA outperformed EuclideanPCA in terms of detection accuracy, precision, and F1-Score. The GrassmannPCA algorithm leveraged the unique geometric properties of the Grassmann manifold, which enabled it to capture complex patterns and relationships within high-dimensional network data more effectively. This led to a lower false alarm rate and better differentiation between normal and malicious traffic compared to the traditional Euclidean-based PCA approach. Additionally, GrassmannPCA exhibited a faster convergence rate, requiring less training time to achieve optimal performance, making it a computationally efficient solution for real-world network security applications. The findings suggest that projecting network data onto Grassmann manifolds is a promising approach for enhancing the detection capabilities of NIDS, particularly in environments with high-dimensional and dynamic traffic data. Future work will focus on extending GrassmannPCA evaluations to diverse datasets such as CICIDS2017 and UNSW-NB15 and comparing its performance with advanced techniques like Autoencoders, GANs, and Transformer-based models. Efforts will also explore hybrid approaches that integrate GrassmannPCA with other statistical or machine learning methods to detect complex attack patterns more effectively. Additionally, optimizing the framework for energy efficiency will ensure its suitability for resource-constrained environments, such as edge devices or IoT systems.

REFERENCES

[1] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *ACM SIGCOMM computer communication review*, vol. 34, no. 4, pp. 219–230, 2004.

[2] H. Huang, H. Al-Azzawi, and H. Brani, "Network traffic anomaly detection," *arXiv preprint arXiv:1402.0856*, 2014.

[3] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[4] D. Brauckhoff, K. Salamatian, and M. May, "Applying pca for traffic anomaly detection: Problems and solutions," *IEEE INFOCOM 2009*, pp. 2866–2870, 2009.

[5] N. T. Van, T. N. Thinh *et al.*, "An anomaly-based network intrusion detection system using deep learning," in *2017 international conference on system science and engineering (ICSSE)*. IEEE, 2017, pp. 210–214.

[6] A. Dunmore, J. Jang-Jaccard, F. Sabrina, and J. Kwak, "A comprehensive survey of generative adversarial networks (gans) in cybersecurity intrusion detection," *IEEE Access*, 2023.

[7] M. Tavallaee *et al.*, "A detailed analysis of the kdd cup 99 dataset," *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pp. 1–6, 2009.

[8] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[9] H. Huang *et al.*, "Network traffic anomaly detection," *arXiv preprint arXiv:1402.0856*, 2014.