Exploring the Impact of PCA Variants on Intrusion Detection System Performance

CHENTOUFI Oumaima[1], CHOUKHAIRI Mouad[2], CHOUGDALI Khalid[3], ALLOUG Ilyas[4]
Engineering Science Laboratory, ENSA Kenitra, Ibn Tofail University, Kenitra, Morocco[1, 3, 4]
LARI, Department of Computer Science Ibn Tofail University,Kenitra, Morocco[2]

***Abstract*—Intrusion detection systems (IDS) play a critical role in safeguarding network security by identifying malicious activities within network traffic. However, the effectiveness of an IDS hinges on its ability to extract relevant features from the vast amount of data it collects. This study investigates the impact of different feature extraction methods on the performance of IDS. We compare the performance of various feature extraction techniques on two widely used intrusion detection datasets: KDD Cup 99 and NSL-KDD. By evaluating these techniques on both datasets, we aim to gain insights into the generalizability and robustness of each method across different dataset characteristics. The study compares the performance of these methods using standard metrics like detection rate, F-measure and FPR for intrusion detection.**

***Keywords*—Intrusion detection; dimensionality reduction; feature extraction; KDDCup'99; NSL-KDD**

I. INTRODUCTION

Machine learning (ML), a subfield of Artificial Intelligence (AI), has seen explosive growth in recent years. ML algorithms learn from data to make predictions or classifications, making them ideal for various applications requiring intelligent behaviour [1]. However, incorporating ever-growing amounts of data can be challenging across various fields, including data analysis, text mining, and even machine learning itself [2]. ML excels at building models for specific tasks like classification (categorizing data), clustering (grouping similar data points), and prediction (forecasting future outcomes) [2]. There are two main Machine Learning approaches: supervised learning and unsupervised learning. On one hand, supervised learning is where the model learns a mapping between input data and desired output based on labelled examples (data with known outcomes) [3]. In other words, we are giving the computer the input data and know what the output should be. Common supervised tasks include classification (e.g., spam vs. non-spam email) and regression (e.g., predicting house prices). In contrast, unsupervised learning analyses unlabelled data (data without predefined categories) to identify patterns or structures [3]. This is like giving a computer data and letting it discover patterns on its own. A common unsupervised technique is clustering, which groups similar data points together. Machine learning algorithms are revolutionizing cybersecurity by enhancing the effectiveness of Intrusion Detection Systems (IDS). IDS is a security tool or software application designed to monitor constantly the traffic, system logs, and events for suspicious patterns that might indicate cyberattacks.

The Host Intrusion Detection System (HIDS) and the Network Intrusion Detection System (NIDS) are the two types of IDS that may be invoked. The first is a system that operates on hosts, analysing logs and system calls on a specific device. Although this sort of IDS may identify intrusions on a single host, its primary downside is that it consumes a lot of resources, which hurts the host's performance. The second, NIDS, operates on the network, analysing packets while remaining undetected and detecting abnormalities and suspicious activities, it analyses packets send and received from different nodes of a network. When establishing a NIDS, we may encounter blind spots, the NIDS's location may have a detrimental impact on the system, and encrypted data may elude detection.

Signature-based IDS and Anomaly based IDS are two methods of intrusion detection systems. On one hand the signature-based IDS is based on comparing the signatures of known attacks with the collected data. In other word, the collected and observed data is compared with a database containing different signatures of known attacks, once there is a match, alerts are triggered. The downfall of this approach is this system cannot detect new attacks, meaning the database should always be updated with the newly found attacks. On the other hand, the anomaly-based IDS focuses on detecting deviation or anomalies from established baselines of normal behaviour. In other words, the "normal" behaviour of the user is determined first, then any type of action is analysed and is considered as an attack if it deviates from normality. This method allows us to detect unknown and new attacks, but it generates a huge number of false positive. We can also mention the hybrid intrusion detection systems that are both Signature-based IDS and Anomaly based IDS. Each type of IDS has benefits as well as drawbacks, and most organizations use a combination of them to offer a degree of protection that detects threats and intrusions throughout the network.

This research explores the potential of machine learning for intrusion detection in computer networks [4] [5] [19]. We begin by examining and analysing existing approaches in the field. Section III delves into the specific anomaly detection methods used in our study. Here, we'll provide a detailed flowchart illustrating the entire process, from data pre-processing to anomaly identification. Following this, Section IV presents the results obtained from implementing these methods. Finally, Section V offers concluding remarks, summarizing the key findings, and demonstrating their value in identifying network threats.

II. Related Work

This research in [8] explores various techniques for feature selection in network intrusion detection systems (NIDS). The authors propose using several algorithms, including Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), and Firefly Algorithm (FFA), either individually or in combinations. Before feature selection, the authors performed essential data preprocessing steps: Label Removal and features removal, then Label Encoding where Categorical features are converted into numerical values and finally Data Binarization where Features are transformed into binary values (0 or 1).Following feature selection, two classifiers are employed: J48 decision tree and Support Vector Machine (SVM).

Zhou & Al [9] presented a novel intrusion detection framework that combines feature selection and ensemble learning techniques to enhance the efficiency of intrusion detection systems. The proposed methodology includes a Correlation-based Feature Selection with Bat Algorithm (CFS-BA) for selecting optimal feature subsets based on feature correlations. An ensemble classifier, comprising C4.5, Random Forest (RF), and Forest Parallel Algorithm (ForestPA) with an Average of Probabilities (AOP) combination rule, is utilized to construct the classification model. The proposed CFS-BA-Ensemble method outperforms other feature selection methods in terms of accuracy, F-Measure, Attack Detection Rate (ADR), and False Alarm Rate (FAR) across different datasets (NSL-KDD, AWID, and CIC-IDS2017). The study highlights the importance of feature selection in reducing computational complexity and improving the performance of intrusion detection systems.

The research in [10] introduces a novel approach called AE-IDS for enhancing classification accuracy and reducing training time in network security. The system utilizes deep learning techniques, specifically auto-encoders, for unsupervised clustering in network intrusion detection. By incorporating random forest feature selection, the method aims to improve the overall performance of intrusion detection systems. The workflow includes setting up decision trees, constructing sub-decision trees, determining output results, calculating classification errors, and assessing feature importance. The study evaluates the method using the DDoS: HIOC dataset and highlights the significance of feature selection in improving system performance. The authors acknowledge the support received for the research and declare no competing interests. Overall, the paper presents a promising approach that combines deep learning and random forest feature selection for effective network intrusion detection. Improved Classification Accuracy: The proposed AE-IDS method demonstrated superior performance in terms of classification accuracy compared to traditional machine learning-based intrusion detection methods. This improvement is attributed to the effective deep learning approach combined with the random forest algorithm.

Venkatesan & al [11] investigated the effectiveness of a new approach for intrusion detection using the NSL-KDD dataset. The authors primarily focus on accuracy, a crucial metric for intrusion detection systems (IDS).The authors leverage the ANOVA F-Test to identify the most relevant features from the NSL-KDD dataset. This helps focus on the information that best distinguishes normal network traffic from intrusions. After feature extraction, the Recursive Feature Elimination (RFE) technique is employed. RFE eliminates features deemed less important based on a ranking system, ultimately reducing the number of features from its original size to a set of 13 most relevant features. To assess the effectiveness of the selected features and the overall approach, the authors employ three different machine learning algorithms: Decision Tree, Random Forest, and Support Vector Machine (SVM). The performance of each algorithm is then evaluated based on accuracy, comparing their ability to correctly identify intrusions within the network traffic data.

In study [12], they introduced a novel and potentially impactful hybrid feature selection method (HFS) designed for intrusion detection systems (IDS). This HFS method combines three techniques: Genetic Search Technique, Rule-Based Engine and CfsSubsetEval. The selected features are then fed into a classifier called KODE for attack classification. The authors demonstrate that their HFS method not only achieves promising results in terms of standard performance metrics (accuracy, precision, recall, etc.), but it also offers benefits in terms of model building and testing time. This suggests that the HFS method can be both effective and efficient for intrusion detection.

The research conducted by Zahid Halim & al [13], and their team focuses on utilizing machine learning and data mining techniques to enhance cybersecurity measures. The study introduces a novel fitness function for genetic algorithms to rank features and develop a feature selection technique, GbFS, for intrusion detection systems. The researchers train machine learning classifiers using the selected optimum features and evaluate performance on benchmark datasets. The proposed method demonstrates effectiveness through comparisons with existing intrusion detection methods and standard feature selection techniques. The paper provides insights on improving detection accuracy, optimizing feature selection, and enhancing cybersecurity measures using genetic algorithms and machine learning approaches.

Pranto & al [14] explores various approaches to using machine learning for effective intrusion detection in network traffic data. The study compares the performance of different algorithms. And to improve computational efficiency, a basic feature selection strategy was employed. The research conducted by Talukder & al [15] propose a novel hybrid machine learning model designed to improve network intrusion detection. The model prioritizes both dependability and effectiveness, offering a reliable solution for identifying malicious activity within network traffic. The model addresses the challenge of imbalanced datasets, often encountered in intrusion detection, by incorporating SMOTE and highlights on the importance of using efficient dimensionality reduction methods to improve computational efficiency without compromising the model's accuracy. The proposed approach is evaluated on two benchmark datasets: KDDCUP'99 and CIC-MalMem-2022. This evaluation ensures the model's

generalizability and adaptability to different types of network traffic data.

III. Proposed Approach

A. System Model and Problem Formulation

One of the issues encountered is the use of enormous datasets to work on new approaches to improve signature-based IDS, therefore the usage of data mining. Data mining is a pre-processing technique before using machine learning models. It is used to explore and extract useful information from data, as well as minimize its dimensions, before using machine learning algorithms.

In this paper, we describe the suggested method in detail. The main idea of the approach is to improve intrusion detection and detect each type of attack by applying different machine learning methods.

Fig. 1 represents the flowchart of the proposed signature-based IDS. In this study, we offered to construct a robust IDS using each time a different method of feature extraction and dimensionality reduction, aiming to improve the accuracy and decrease the false positive rate.

B. Data Preparation

Our work will be applied on two different and well-known datasets the KDDCup'99 and the NSL KDD. These datasets have been widely used for so many approaches and by different researchers all over the globe for evaluating intrusion detection systems and asses the performance of these approaches in cybersecurity domain [16] [17] [18] [24].

The KDD Cup 99 is a well-known dataset used in the field of cybersecurity and network intrusion detection. It was organized as part of the KDD (Knowledge Discovery and Data Mining) conference in 1999 and aimed to help researchers to propose and try new approaches in detecting network intrusions or attacks within computer systems. The KDD-CUP 99 dataset features 41 attributes describing network traffic and categorizes them into five classes: normal, Denial of Service, User to Root, Remote to Local, and Probe attacks.

The presence of redundant and irrelevant information in the KDD Cup 99 dataset can negatively affect the performance of analysis and machine learning models. Thus, the use of the NSL-KDD where multiple challenges have been resolved. The NSL-KDD is the modified version of the KDD CUP 99, where they worked on reducing the redundant and irrelevant data, and solve the problem of imbalanced data, where it made the machine learning models be bias towards the majority class and reduce the effectiveness of detecting the attacks that were underrepresented.

Overall, NSL-KDD aimed to provide a more suitable and realistic dataset for evaluating intrusion detection systems. These improvements have contributed to more robust and accurate intrusion detection models that are better suited for real-world applications. This Dataset contains the same five classes of patterns, but with different representations.

Through providing appropriate data set to reduce the data afterwards, both the KDD Cup'99 and the NSL-KDD dataset passe through multiple steps of pre-processing.

- Step1: Collecting and splitting the data.

Building an effective intrusion detection model relies on having some well-prepared data, and collecting this data and splitting it is the first important step. When separating the used dataset into training and testing sets, it's crucial to acknowledge that these sets should not be from the same underlying probability distribution. This implies that certain attack types present in the testing data might be absent in the training data, enhancing the realism of the evaluation but also posing challenges for accurate detection.

- Step2: Vectorizing the data using one hot encoding:

Vectorizing data is essential for machine learning as most of machine learning algorithms require numerical input. This involves transforming data into numerical vectors. Since our datasets contain both numerical and categorical features, we'll leverage one-hot encoding for the categorical ones. This technique essentially creates separate binary vectors for each category, effectively expanding the feature space and enhancing the model performance.

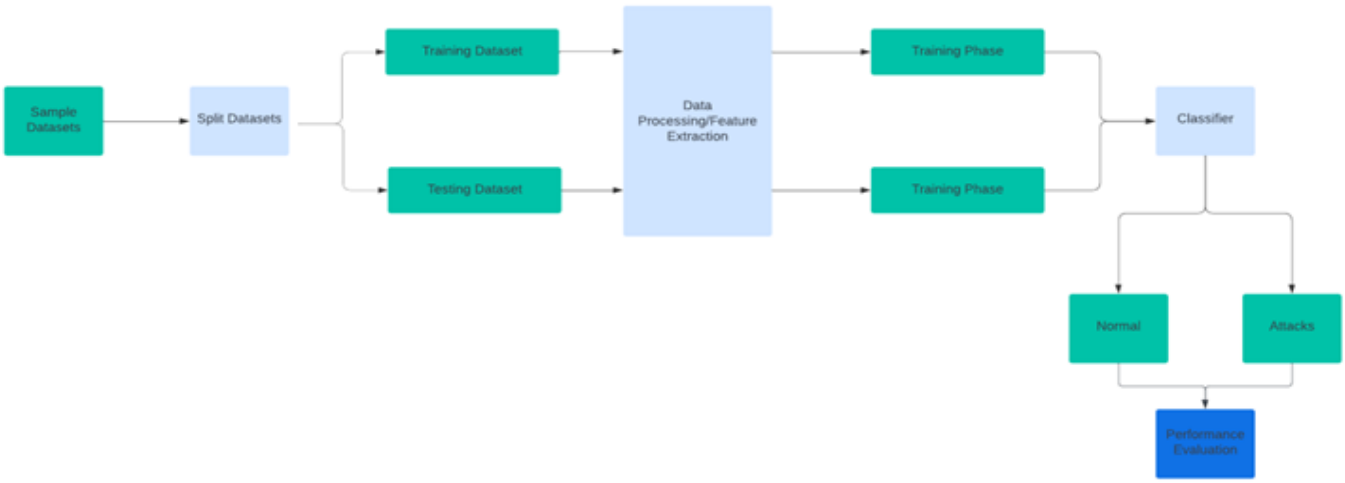

Fig. 1. Flowchart of the proposed Approach of NIDS.

- Step 3: Feature Scaling:

Data scaling is the act of transforming the values of features of a dataset into a specific range.

C. Dimensionality Reduction

Reducing dimensions can lead to simpler models, faster computation, improved generalization by reducing noise and redundancy, and easier visualization of data.

Principal Component Analysis (PCA) is a technique for dimensionality reduction, aiming to transform a high dimensional dataset into a lower dimensional subspace while preserving the most significant information [7]. Transforming several correlated variables into a set of mutually orthogonal variables called Principal components (PCs) where the initial PCs encapsulate the highest information density. Let's assume we have a training data matrix described as follow:

$$X = \begin{pmatrix} x11 & \cdots & xpn \\ \vdots & \ddots & \vdots \\ xn1 & \cdots & xpn \end{pmatrix} = \begin{pmatrix} x1 \\ \vdots \\ xn \end{pmatrix} \quad (1)$$

Where, p is the columns vectors and n is the data size. To get the PCs of the training set, we'll first compute the average of this set:

$$\bar{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \quad (2)$$

The covariance matrix C(xj) will be calculated to identify the scatter degree of the feature vectors to identify the key features, it can be determined as follow:

$$\mathrm{C(xj)} = \frac{1}{n} \sum_{i=0}^{n} \left(x_{ij} - \bar{x_j}\right)\left(x_{ij} - \bar{x_j}\right)^t \quad (3)$$

Following the computation of the covariance matrix, the next step involves computing the eigenvectors and their corresponding eigenvalues. These eigenvectors, also known as principal components (PCs) should be sorted in a descending order where the first PCs encapsulate most of the data variance. The selection of the principal components should strike a balance between retaining critical information and achieving the desired level of dimensionality reduction. This ratio is defined by the following formula:

$$\beta = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n'}} \lambda_{\mathrm{k}}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \lambda_{\mathrm{k}}} \quad (4)$$

Once these PCs are chosen and validated, the original data is projected into the PCs, creating a new lower dimensional projection.

L2-p norm based PCA is a variant of principal component analysis that incorporates the L2-p norm as a measure of distance or similarity between data points, allowing for more robust and flexible dimensionality reduction. The L2-p norm based PCA offers a valuable approach for dimensionality reduction, as it allows for adjusting the importance of different dimensions based on the chosen value of p. Wang & al. [20] proposed this approach where:

$$\min_{W} \sum_{i=1}^{n} \|x_i - WW^T x_i\|_2^p \quad (5)$$

Subject to : $W^T W = I$

where, $0 < p \leq 2$.

$$\sum_{i=1}^{N} \|x_i - WW^T x_i\|_2^2 \|x_i - WW^T x_i\|_2^{p-2}$$

$$= \sum_{i=1}^{N} tr\left\{(x_i - WW^T x_i)^T * (x_i - WW^T x_i)\right\} d_i \quad (6)$$

$$= \sum_{i=1}^{N} tr\left\{x_i^T x_i - x_i WW^T x_i - x_i WW^T x_i + x_i WW^T WW^T x_i\right\} d_i \quad (7)$$

$$= \sum_{i=1}^{N} tr\left\{x_i^T x_i - x_i WW^T x_i\right\} d_i \quad (8)$$

where: $\mathrm{d}_i = \|x_i - WW^T x_i\|_2^{p-2}$

By substituting Eq. (8) into Eq. (5), we'll obtain the following objective function:

$$\min_{W} \sum_{i=1}^{N} tr\left\{x_i^T x_i\right\} d_i - \sum_{i=1}^{N} tr\left\{W^T x_i x_i^T W\right\} d_i \quad (9)$$

The primary focus at this juncture is on devising a method to determine the optimal projection matrix W for the objective function (8). The goal is to find a projection matrix W that reduces the objective function value to the minimum. This objective function (8) involves the unknown variables W and di, which are interlinked with W. Given that the objective function (8) lacks a straightforward, closed-form solution, directly addressing it poses a significant challenge. An approach that can be developed involves iteratively updating W (holding di constant) and di (holding W constant).

$$W^* = \text{argmax } tr(W^T XDX^T W) \quad (10)$$

Subject to : $W^T W = I$

In this context, D represents a diagonal matrix with its diagonal elements being di, and the column vectors of W in the objective function (10) consist of eigenvectors from XDXT, which correspond to the k highest eigenvalues. Following this, the diagonal element di within the matrix D is updated. This iterative process is carried out repeatedly until the algorithm reaches convergence.

The Double L2, p-norm based Principal Component Analysis (DLPCA), introduced by Huang & al [6], presents an innovative technique for feature extraction. It is designed to reduce reconstruction error while increasing data variance within a cohesive structure. DLPCA incorporates the L2,p-norm distance metric into its objective function, improving its ability to manage outliers with greater robustness and efficiency. Through the identification of two transformation matrices, the method optimizes both data variance and reconstruction error, providing an effective approach to feature extraction challenges.

To maximize the data variance and achieve robust results to outliers, we'll use the following formulation:

$$\max_{W} \sum_{i=1}^{n} \|W^T x_i\|_2^p \quad (11)$$

Subject to: $W^T W = I$

They propose a robust model for minimizing reconstruction error. This model incorporates utilization of different transformation matrices for each role involved in the feature extraction process.

$$\min_{W,U} \sum_{i=1}^{n} \|x_i - UW^T x_i\|_2^p \quad (12)$$

Subject to : $W^T W = I \text{ and } U^T U = I$

Eq. (11) defines a two-step process for data transformation. First, matrix W projects the data into a lower-dimensional space for efficient processing. Then, matrix U recovers the data from this compressed form. To fulfil the goal of using both the minimization of reconstructed error and the maximization of data variance into account, we combine (10) and (11) to get the objective function of the double L2-p norm PCA formulated as follow:

$$\min_{W,U} \frac{\sum_{i=1}^{n} \|x_i - UW^T x_i\|_2^p}{\sum_{i=1}^{n} \|W^T x_i\|_2^p} \quad (13)$$

Subject to : $W^T W = I \text{ and } U^T U = I$

Unlike existing robust PCA methods that focus solely on either minimizing reconstruction error or maximizing data variance, this approach takes a unified perspective, by combining these aspects into a single framework, allowing them to contribute more effectively to the projection learning process.

IV. Results and Discussion

A. Performance Metrics

Accuracy (AC): is the ability of identifying accurately both known and novel malicious activities. Can be determined by the following equation:

$$AC = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} * 100 \quad (14)$$

Precision (PR): Configurable hyper-parameter used for accurate classification of attacks within the intrusion detection system. PR is defined by the following formula:

$$\text{PR} = \frac{\text{Tp}}{\text{Tp+Fp}} * 100 \quad (15)$$

Recall (RC): Also known as the detection rate (DR), it refers to the ability to identify and flag malicious activities and can be calculated as follow:

$$\text{RC} = \text{DR} = \frac{\text{Tp}}{\text{Tp+Fn}} * 100 \quad (16)$$

False Positive Rate (FPR): Is the proportion of falsely identified normal behaviour detected as abnormal action which is expressed by the following formula:

$$\text{FPR} = \frac{\text{Fp}}{\text{Tn+Fp}} * 100 \quad (17)$$

F-measure (FM): It offers a balance view of the performance of individual metrics precision and recall. Is computed by the following formula:

$$\text{FM} = 2 * \frac{\text{PR*RC}}{\text{PR+RC}} * 100 \quad (18)$$

B. Performance Evaluation

In cybersecurity, a high F-measure implies that the system accurately identifies most attacks while minimizing false alarms that can overburden security personnel. This helps prioritize genuine threats and optimize security response measures.

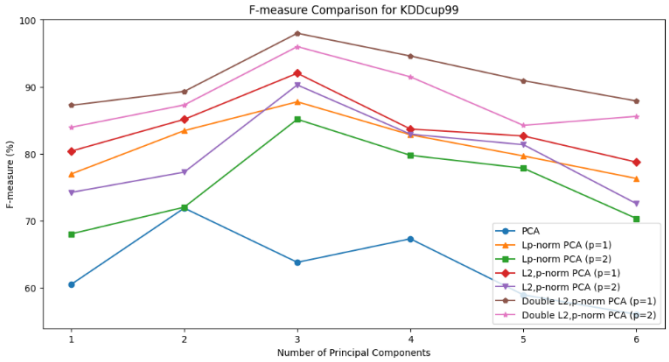

Fig. 2. Principal Components vs. F-Measure for KDDcup99.

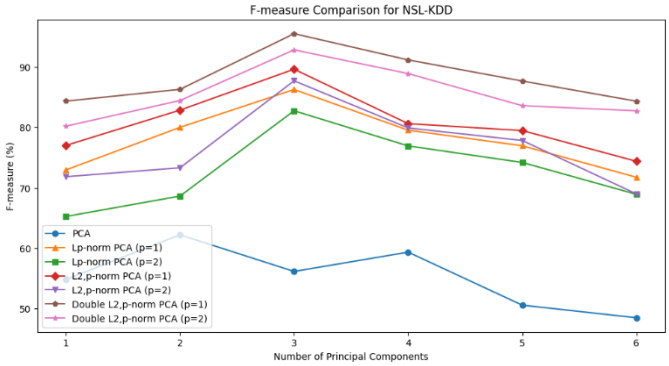

Fig. 3. Principal Components vs. F-Measure for NSL-KDD.

Fig. 2 and 3 provide complimentary perspective on the relationship between the number of principal components chosen for feature extraction and F-measure, a metric that balances precision and recall. Techniques utilizing Lp-norms (p=1 and 2) [22] [23] or double L2, p-norms generally achieve superior F-measures compared to standard PCA across most numbers of principal components. This suggests these methods extract more relevant features, leading to better overall performance in intrusion detection. As the number of principal components increases, F-measures tend to improve for most techniques. This indicates that higher-dimensional feature representations capture more information, potentially leading to better precision (correctly identifying intrusions) and recall (minimizing missed attacks). Notably, Double L2,p-norm PCA consistently demonstrates the highest F-measures regardless of the number of principal components chosen. This finding highlights its effectiveness in feature extraction for the KDD Cup dataset and the NSL-KDD dataset. It suggests that Double L2,p-norm PCA excels at selecting

informative features across different dimensionalities, leading to a good balance between precision and recall in intrusion detection.

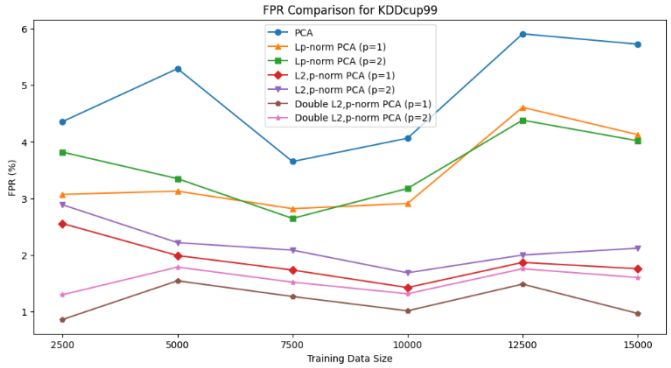

Fig. 4. Training data vs. F-measure for KDDCup99.

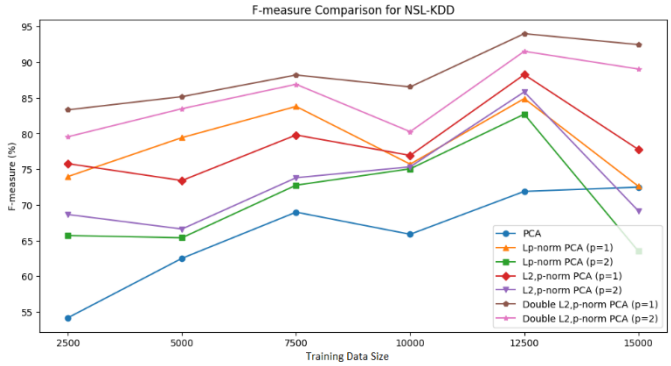

Fig. 5. Training data vs. F-measure for NSL-KDD.

Fig. 4 and Fig. 5 offers distinct visual representation of how the feature extraction techniques that incorporate Lp-norms (p=1 and 2) or double L2,p-norms generally achieve higher F-measures compared to standard PCA across most training sizes. This suggests they extract more informative features, leading to better overall performance in intrusion detection. As the amount of training data increases, F-measures tend to improve for most techniques. This highlights the importance of larger datasets for achieving better precision and recall in intrusion detection. Notably, Double L2,p-norm PCA consistently demonstrates the highest F-measures across all training sizes. This finding underscores its effectiveness in feature extraction for the KDD Cup dataset, as it leads to a better balance between correctly identifying intrusions (high precision) and minimizing missed attacks (high recall).

In both Fig. 6 and Fig. 7, traditional PCA generally performed worse than the other techniques, especially as the number of features analysed increased. Double L2,p-norm PCA consistently achieved the highest detection rates in all scenarios. This suggests it's the most effective method for extracting relevant information from network traffic data for intrusion detection on the KDD Cup dataset and the NSL-KDD Dataset. Within the Double L2,p-norm PCA technique, using p=1 typically led to better results than using p=2 in most cases.

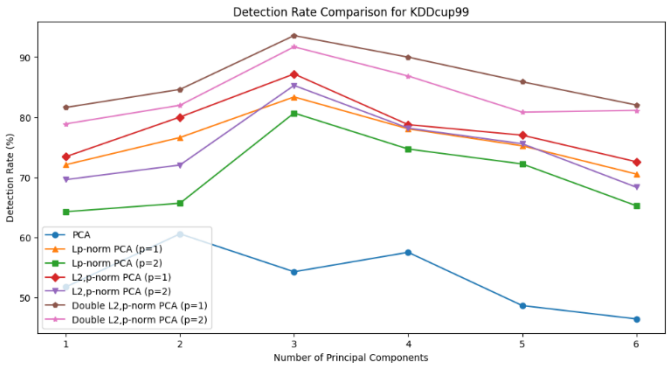

Fig. 6. Principal Components vs. DR for KDDcup99.

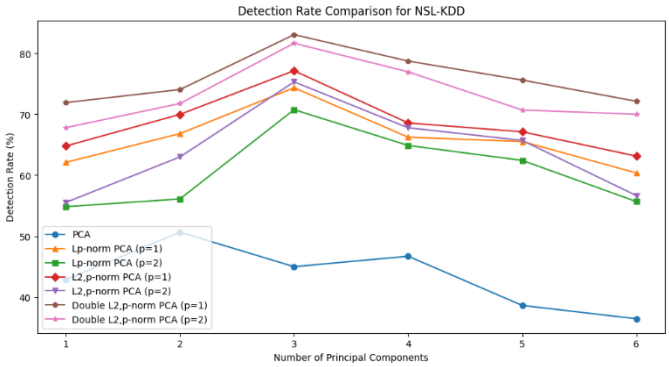

Fig. 7. Principal Components vs. DR for NSL-KDD.

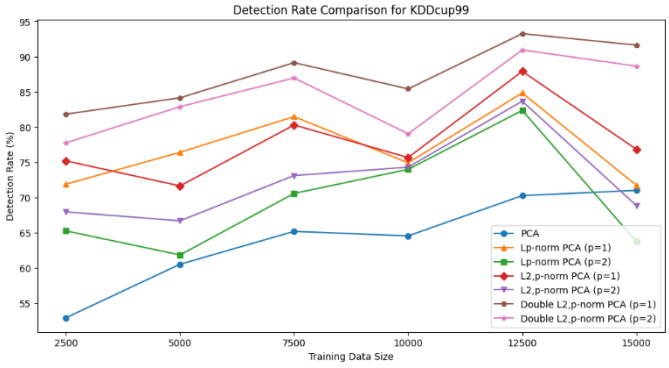

Fig. 8. Training data vs. DR for KDDCup99.

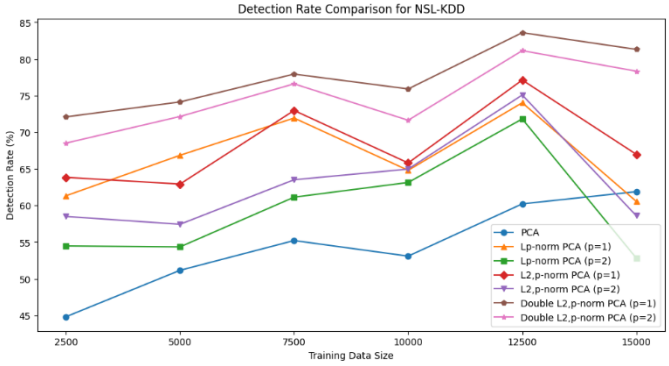

Fig. 9. Training data vs. DR for NSL-KDD.

As anticipated, larger training sets consistently elevate detection rates across all feature extraction techniques examined. This reinforces the notion that ample data is crucial for optimal intrusion detection system (IDS) performance. The relationship between the training data size and detection rate is further elucidated by Fig. 8 and Fig. 9. Techniques incorporating Lp-norms (p=1 and 2) or double L2, p-norms generally surpass standard PCA, particularly as the training size increases. This suggests that these methods capture more relevant information from the data, leading to more effective intrusion detection. Notably, double L2, p-norm PCA (p=1) consistently achieves the highest detection rates, demonstrating its efficacy in feature extraction for the KDD Cup dataset.

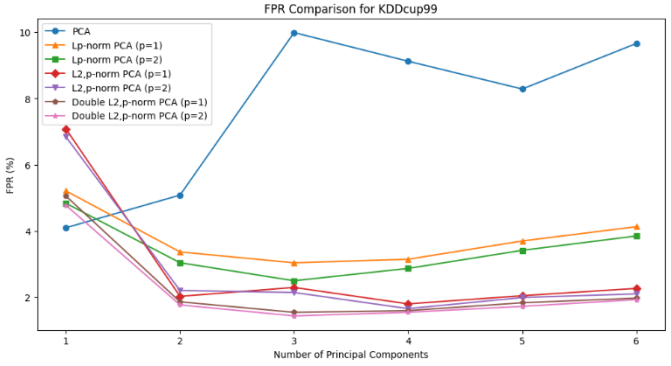

Fig. 10. Principal Components vs. FPR for KDDcup99.

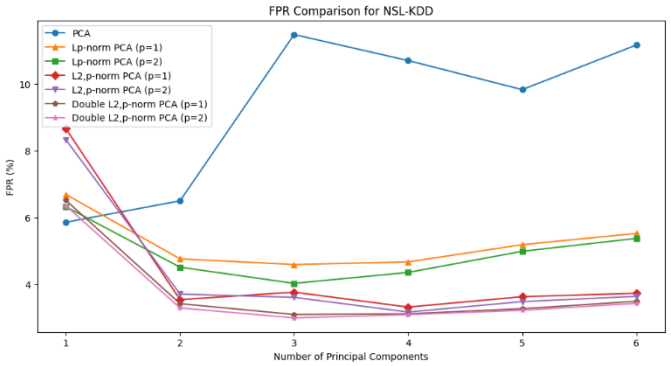

Fig. 11. Principal Components vs. FPR for NSL-KDD

Fig. 10 and Fig. 11 examines how feature extraction techniques impact false positive rates, a crucial metric in intrusion detection systems (IDS). Feature extraction methods incorporating Lp-norms (p=1 and 2) or double L2,p-norms generally achieve lower false positive rates compared to standard PCA across most settings. As the number of principal components increases (higher dimensionality), false positive rates tend to decrease for most techniques. This indicates that higher-dimensional feature spaces allow for better separation between normal and abnormal network activities, reducing the chances of misidentification. Notably, Double L2,p-norm PCA consistently demonstrates the lowest false positive rates. This finding highlights its effectiveness in feature extraction for the KDD Cup dataset where it excels at creating features that effectively distinguish between normal and attack traffic, minimizing the number of false alarms generated by the IDS.

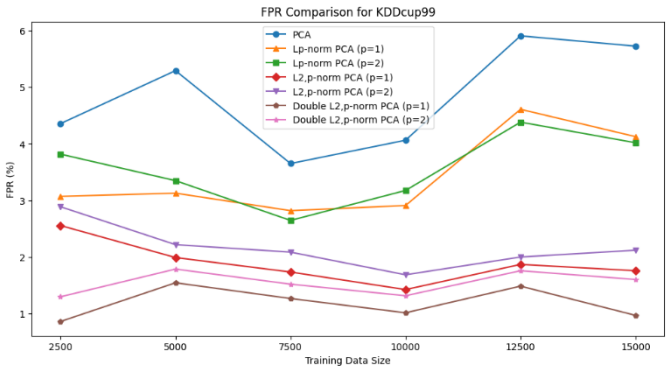

Fig. 12. Training data vs. FPR for KDDCup99.

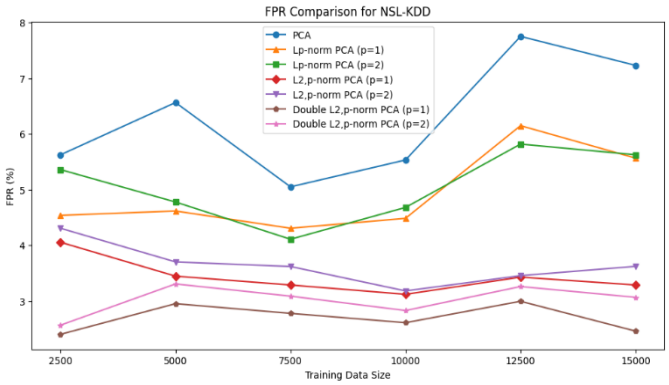

Fig. 13. Training data vs. FPR for NSL-KDD.

False positives occur when an IDS mistakenly identifies normal traffic as an attack. Here's what the findings reveal based on both Fig. 12 and Fig. 13. Techniques that incorporate Lp-norms (p=1 and 2) or double L2, p-norms generally achieve lower false positive rates compared to standard PCA across most training sizes. This suggests these methods create more robust feature representations, leading to fewer instances of misclassifying normal data as intrusions. As the amount of training data increases, false positive rates tend to decrease for most techniques. This highlights the importance of larger datasets for an IDS to learn the subtle differences between normal and abnormal network activities, ultimately reducing false alarms. Notably, Double L2,p-norm PCA consistently demonstrates the lowest false positive rates across all training sizes.

TABLE I. OBTAINED RESULTS FOR THE KDDCUP99

Used Method	Performance Metrics (%)		
	DR	***FPR***	***F-measure***
PCA	70,26	5,91	78,71
Lp norm PCA(p=1)	84,83	4,61	88
Lp norm PCA(p=2)	82,35	4,38	84,60
L2-p norm PCA(p=1)	87,90	1,86	90,63
L2-p norm PCA(p=2)	83,66	1,99	87,40
Double L2-p norm PCA(p=1)	93,24	1,48	96,37
Double L2-p norm PCA(p=2)	90,93	1,75	94,44

TABLE II. OBTAINED RESULTS FOR THE NSL-KDD

Used Method	Performance Metrics (%)		
	DR	*FPR*	*F-measure*
PCA	60,23	7,75	71,89
Lp norm PCA(p=1)	74,03	6,15	84,91
Lp norm PCA(p=2)	71,82	5,82	82,72
L2-p norm PCA(p=1)	77,13	3,43	88,27
L2-p norm PCA(p=2)	75,08	3,46	85,84
Double L2-p norm PCA(p=1)	83,59	3,00	94,01
Double L2-p norm PCA(p=2)	81,14	3,26	91,54

The tables demonstrate how various dimensionality reduction techniques, including PCA, Lp-norm PCA, L2-p norm PCA, and double L2-p norm PCA, impact intrusion detection performance. These results highlight the importance of dimensionality reduction and feature extraction in building efficient and robust intrusion detection systems (IDS). Notably, double L2-p norm PCA appears to be a promising method for reducing dimensionality in network security. The analysis reveals that applying double L2-p norm PCA with p=1 consistently achieved the highest detection rate (DR), F-measure (a balanced metric for accuracy), and lowest false positive rate (FPR) across both KDD Cup'99 and NSL-KDD datasets. This suggests that double L2-p norm PCA is the preferred approach for enhancing IDS due to its ability to preserve crucial data features. It achieves this by simultaneously maximizing data variance and minimizing reconstruction error.

V. CONCLUSION

This paper investigates the application of Principal Component Analysis (PCA) for network intrusion detection. We propose several PCA-based models and evaluate their effectiveness on the KDD Cup 99 and NSL-KDD datasets. Our goal is to assess their ability to detect a wide range of attacks. The experiments highlight the importance of feature extraction techniques like PCA in improving intrusion detection. Among the models tested, Double L2,p-norm PCA emerged as the most with promising method among those tested. These observations offer valuable insights into the interplay between training size and feature extraction techniques in IDS performance. The research compared several dimensionality reduction techniques for their impact on noise reduction and overall effectiveness in cybersecurity intrusion detection. Analysis of the KDDCup99 and NSL-KDD datasets revealed a clear trend techniques achieved a wider range of detection rates. Principal Component Analysis (PCA) resulted in the lowest detection rate (70.26%) for KDD Cup'99, while Double L2-p norm PCA with p=1 achieved the highest (93.24%). Similar variations were observed for NSL-KDD (60.23% to 83.59%). Future work should explore the computational demands and scalability of these methods. This could involve testing them on a broader range of network scenarios and considering data imbalances to assess their real-world applicability [21] [25]. Striking a balance between detection accuracy and computational cost will be crucial for deploying these techniques in practical Intrusion Detection Systems (IDS).

REFERENCES

[1] Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-driven cybersecurity: An overview, security intelligence modeling and research directions. SN Computer Science, 2(3), 160. [DOI: 10.1007/s42979-021-00592-x].

[2] Jadidoleslamy, H. (2011). A high-level architecture for intrusion detection on heterogeneous wireless sensor networks: Hierarchical, scalable and dynamic reconfigurable. Wireless Sensor Network, 3(07), 241.

[3] Han, J., Pei, J., & Kamber, M. (2011). Data mining: Concepts and techniques. Amsterdam: Elsevier.

[4] Agrawal, S., Sarkar, S., Aouedi, O., Yenduri, G., Piamrat, K., Alazab, M., Bhattacharya, S., Reddy Maddikunta, P. K., Gadekallu, T. R., & Thippanna Reddy (2022). Federated Learning for Intrusion Detection System: Concepts, Challenges and Future Directions. Computer Communications, 195(November), 346–61. [DOI: 10.1016/j.comcom.2022.09.012].

[5] Ahmad, Z., Khan, A. S., Shiang, C. W., Abdullah, J., & Ahmad, F. (2021). Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches. Transactions on Emerging Telecommunications Technologies, 32(1), e4150. [DOI: 10.1002/ett.4150].

[6] Huang, P., Ye, Q., Zhang, F., Yang, G., Zhu, W., & Yang, Z. (2021). Double L2,p-norm based PCA for feature extraction. Information Sciences, 573, 345–359.

[7] Almaiah, M. A., Almomani, O., Alsaaidah, A., Al-Otaibi, S., Bani-Hani, N., Al Hwaitat, A. K., Al-Zahrani, A., Lutfi, A., Bani Awad, A., & Theyazn H. H. Aldhyani (2022). Performance Investigation of Principal Component Analysis for Intrusion Detection System Using Different Support Vector Machine Kernels. Electronics, 11(21), 3571. [DOI: 10.3390/electronics11213571].

[8] Almomani, O. (2020). A Feature Selection Model for Network Intrusion Detection System Based on PSO, GWO, FFA and GA Algorithms. Symmetry, 12(6), 1046. [DOI: 10.3390/sym12061046].

[9] Zhou, Y., Cheng, G., Jiang, S., & Dai, M. (2020). Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier. Computer Networks, 174(June), 107247. [DOI: 10.1016/j.comnet.2020.107247].

[10] Li, X., Chen, W., Zhang, Q., & Wu, L. (2020). Building Auto-Encoder Intrusion Detection System Based on Random Forest Feature Selection. Computers & Security, 95(August), 101851. [DOI: 10.1016/j.cose.2020.101851].

[11] Venkatesan, S. (2023). Design an Intrusion Detection System Based on Feature Selection Using ML Algorithms. 72(1).

[12] Jaw, E., & Wang, X. (2021). Feature Selection and Ensemble-Based Intrusion Detection System: An Efficient and Comprehensive Approach. Symmetry, 13(10), 1764. [DOI: 10.3390/sym13101764].

[13] Halim, Zahid, Muhammad Nadeem Yousaf, Muhammad Waqas, Muhammad Sulaiman, Ghulam Abbas, Masroor Hussain, Iftekhar Ahmad, et Muhammad Hanif. « An Effective Genetic Algorithm-Based Feature Selection Method for Intrusion Detection Systems ». Computers & Security 110 (novembre 2021): 102448. https://doi.org/10.1016/j.cose.2021.102448.

[14] Pranto, M. B., Ratul, M. H. A., Rahman, M. M., Diya, I. J., & Zahir, Z.-B. (2022). Performance of Machine Learning Techniques in Anomaly Detection with Basic Feature Selection Strategy - A Network Intrusion Detection System. Journal of Advances in Information Technology, 13(1), 36-44. [DOI: 10.12720/jait.13.1.36-44].

[15] Talukder, M. A., Hasan, K. F., Islam, M. M., Uddin, M. A., Akhter, A., Yousuf, M. A., Alharbi, F., & Moni, M. A. (2023). A Dependable Hybrid Machine Learning Model for Network Intrusion Detection. Journal of Information Security and Applications, 72(February), 103405. [DOI: 10.1016/j.jisa.2022.103405].

[16] Choudhary, S., & Kesswani, N. (2020). Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets Using Deep Learning in IoT. Procedia Computer Science, 167, 1561–73. [DOI: 10.1016/j.procs.2020.03.367].

[17] El-Sappagh, S., Mohammed, A. S., & AlSheshtawy, T. A. (2019). CLASSIFICATION PROCEDURES FOR INTRUSION DETECTION

BASED ON KDD CUP 99 DATA SET. International Journal of Network Security & Its Applications, 11(03), 21–29. [DOI: 10.5121/ijnsa.2019.11302].

[18] Gumusbas, D., Yldrm, T., Genovese, A., & Scotti, F. (2021). A Comprehensive Survey of Databases and Deep Learning Methods for Cybersecurity and Intrusion Detection Systems. IEEE Systems Journal, 15(2), 1717–31. [DOI: 10.1109/JSYST.2020.2992966].

[19] Jaradat, Ameera S., Malek M. Barhoush, et Rawan S. Bani Easa. « Network Intrusion Detection System: Machine Learning Approach ». Indonesian Journal of Electrical Engineering and Computer Science 25, no 2 (1 février 2022): 1151. https://doi.org/10.11591/ijeecs.v25.i2.pp1151-1158..

[20] Wang, Q., Gao, Q., Gao, X., & Nie, F. (2018). *L2-p Norm Based PCA for Image Recognition. IEEE Transactions on Image Processing, 27(3), 1336–1346*. doi:10.1109/tip.2017.2777184.

[21] Kilincer, I. F., Ertam, F., & Sengur, A. (2021). Machine Learning Methods for Cyber Security Intrusion Detection: Datasets and Comparative Study. Computer Networks, 188(April), 107840. [DOI: 10.1016/j.comnet.2021.107840].

[22] Kwak, N. (2014). Principal Component Analysis by Lp-Norm Maximization. IEEE Transactions on Cybernetics, 44(5), 594–609. [DOI: 10.1109/TCYB.2013.2262936].

[23] Liang, Z., Xia, S., Zhou, Y., Zhang, L., & Li, Y. (2013). Feature Extraction Based on Lp-Norm Generalized Principal Component Analysis. Pattern Recognition Letters, 34(9), 1037–45. [DOI: 10.1016/j.patrec.2013.01.030].

[24] Ngueajio, M. K., Washington, G., Rawat, D. B., & Ngueabou, Y. (2023). Intrusion Detection Systems Using Support Vector Machines on the KDDCUP'99 and NSL-KDD Datasets: A Comprehensive Survey. In Intelligent Systems and Applications (pp. 609–29). Cham: Springer International Publishing. [DOI: 10.1007/978-3-031-16078-3_42].

[25] Thakkar, Ankit, et Ritika Lohiya. « A Review of the Advancement in Intrusion Detection Datasets ». Procedia Computer Science 167 (2020): 636-45. https://doi.org/10.1016/j.procs.2020.03.330.