International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Intrusion Detection System by Integrating Mod K- Means C Algorithm and T-SNE Dimensionality Reduction

Mallaradhya C[1*], Dr G N K Suresh Babu[2]

Submitted: 06/02/2024 **Revised**: 14/03/2024 **Accepted**: 20/03/2024

Abstract: This research presents a new methodology to enhance the precision of Intrusion Detection Systems (IDS) by integrating the Modified K-Means Clustering (ModKMeansC) algorithm as a classifier along with t-Distributed Stochastic Neighbor Embedding (t-SNE) a reduction technique of dimensionality. In the realm of cyber security, conventional IDS face challenges in accurately discerning abnormal network behavior due to the dynamic and intricate nature of cyber threats. The ModKMeansC algorithm, intricately designed to address issues stemming from abnormal network connections, introduces parallelism into centroid and distance calculation update operations. This concurrent execution, performed asynchronously for each data point, facilitates real-time analysis of network traffic, thereby bolstering efficiency and responsiveness. Leveraging the CICIDS2017 dataset, encompassing both normal and abnormal network traffic patterns, the study implements and fine-tunes the ModKMeansC algorithm for optimal performance. t-SNE is applied to preprocess the data before feeding it into the classifier. The proposed system's performance is meticulously assessed using key performance metrics. A proportional analysis against traditional intrusion detection algorithms underscores the ModKMeansC algorithm's advantages in accurately categorizing abnormal network behavior as 92%. Results and ensuing discussions highlight the algorithm's adeptness in efficiently handling abnormal network connections and its prowess in parallel processing. This examination significantly supports to the dynamic field of cyber security by presenting a more effective and responsive methodology for identifying abnormal network behavior. The amalgamation of the ModKMeansC algorithm with t-SNE holds considerable promise in elevating the accuracy of IDS as 95%. Future research directions may encompass adapting the proposed system to real-world cyber security scenarios and further optimizing the algorithm for scalability in large-scale networks.

Keywords: *Intrusion Detection System, Modified K-Means Clustering, t-SNE Dimensionality Reduction, Cyber security, Network Security, Parallel Processing, and Machine Learning.*

1. Introduction

In an era where the internet is witnessing a surge in new devices and emerging technologies, the attack surface expands, providing cybercriminals with opportunities to compromise inadequately secured devices. The rapid digitization of our world emphasizes the urgency of promptly detecting and preventing cyber-attacks [Anurag Chhetri et al 2022]. This imperative extends across various entities, from individual homeowners to entire nation-states, intensifying the complexity of establishing robust security measures. The evolving landscape of cyber security is further complicated by the continuous development of malicious actors' skills, as they adeptly employ sophisticated attack techniques to evade existing detection measures. Addressing these challenges becomes paramount as the digital realm becomes an integral part of our daily lives, emphasizing the need for resilient and adaptive cyber security strategies to safeguard against the constantly evolving threat landscape [Satrya GB et al 2022].

Intrusion Detection Systems (IDS) [Md Nayer et al 2021] emerge as vital tools, serving as proactive guardians against a diverse array of cyber threats. These systems actively monitor and analyze network and system activities, seeking to identify and respond promptly to potential intrusions or unauthorized actions. Distinguishing between two primary categories, signature and anomaly [Van NT et al 2017] based, IDS harnesses the power of a comprehensive database for known attack patterns and baseline behavior analysis to detect deviations that may signify emerging threats [Ahmad I 2018]. Functioning as vigilant sentinels, signature-based IDS excel in recognizing predefined attack patterns like viruses and malware, while anomaly-based IDS dynamically adapt to evolving threats, including previously unknown attacks like zero-day exploits. The overarching objective of IDS lies in its capacity to provide real-time alerts and facilitate swift responses, contributing substantially to the resilience of networks and systems [Jianyi Liu 2018]. As a crucial component in the cyber security landscape, IDS plays a key role in protection of delicate data, inhibiting not permitted admission, and

[1] *Research Scholar, Department of Computer Science, Srishti College of Commerce and Management, University of Mysore.*

[2] *Professor, Department of Computer Science, Srishti College of Commerce and Management, University of Mysore.*

Corresponding mail id- mallaradhyac1985@gmail.com

countering a spectrum of cyber threats from routine security breaches to sophisticated targeted attacks. In an era where the digital landscape constantly evolves, IDS remains indispensable for organizations across sectors, continuously refining and integrating advanced technologies to fortify cyber security defenses [Moshref M et al 2022].

Data mining has become an essential and dynamic facet within the field of cyber security, especially when applied to the intricacies of Intrusion Detection Systems (IDS). As the digitalization [Zhu Y et al 2017] aspects continues to burgeon, the volume and complexity of data generated by network activities have reached unprecedented levels. In this context, data mining stands as a sophisticated analytical tool [Mohamad Faiz Ahmad et al 2022] that enables the systematic extraction of valuable patterns and insights from large datasets. This process is particularly crucial in the realm of Intrusion Detection, where identifying subtle indicators of potential security breaches amid a sea of data can be a daunting task. By employing advanced algorithms and statistical techniques [Omar Almoman 2021], data mining allows security analysts to unveil hidden correlations, trends, and anomalies, thus providing a comprehensive understanding of network behaviors that may signify malicious activities [M Deepa 2021].

The symbiotic relationship between data mining and Intrusion Detection is paramount for enhancing cyber security measures. Data mining [Ahmad I et al. 2018] serves as a force multiplier, empowering security professionals to efficiently sift through colossal amounts of information and unveil meaningful patterns indicative of security threats. This proactive approach enables organizations to fortify their defenses, anticipating and responding to evolving cyber threats with greater speed and accuracy. In essence, the integration of data mining techniques into the realm of Intrusion Detection not only streamlines the analysis of massive datasets but also contributes significantly to the overarching goal of bolstering cyber security resilience in the expression of an evolution of threat landscape.

2. Literature Survey

In modern years, the area of digitalization has observed an unprecedented escalation in cyber threats, necessitating the adoption of advanced security measures. The pivotal role of Intrusion Detection Systems (IDS) in identifying and mitigating cyber threats is underscored by Soewu et al. (2022). When combined with data mining techniques, such as Random Tree, Naive Bayes, J48, and Random Forest, IDS proves to be a formidable defense mechanism by accurately classifying attacks based on essential parameters like data accuracy and pattern recognition.

Thockchom et al. (2023) respond to the challenges posed by the exponential growth of the Internet with a forward-thinking approach, proposing an ensemble learning-based method for intrusion detection. This innovative methodology utilizes Gaussian Naive Bayes, logistic regression, and decision tree as base classifiers, with stochastic gradient descent serving as the meta-classifier. The ensemble model not only showcases superior performance but also demonstrates particular effectiveness in scenarios with unbalanced datasets, highlighting its potential to effectively identify and counteract a diverse range of cyber threats.

Fang's (2023) exploration takes us into the realm of data mining technology's application in employment education management. The study introduces an enhanced K-means algorithm, complemented by the optimized Apriori algorithm, to analyze employment education data. The result is not only valuable insights but also a demonstration of heightened accuracy and stability in guiding students' employment decisions. This application of data mining techniques signifies a broader horizon for these technologies, extending beyond traditional cyber security domains into education management.

Turning the focus back to network security, Mahesh T R et al. (2023) delve into the indispensable role of IDS in maintaining the integrity of networks. Leveraging data mining and machine learning techniques, the study aims to extract intricate user behavior patterns, recognizing the need for a more nuanced approach in the face of evolving cyber threats. The introduction of the Length-Decreasing Support technique represents an upgrade in data mining for intrusion detection, effectively addressing the bottleneck associated with frequent item sets mining and showcasing a commitment to advancing the state-of-the-art in cyber security.

Nishika Gulia et al. (2023) steer the exploration towards cloud computing, introducing an IDS integrated with machine learning, specifically artificial bee colony (ABC). The proposed Group-ABC method not only demonstrates efficacy in controlling various attacks but also outperforms existing works in terms of precision, recall, accuracy, and F-measure. This research underscores the adaptability of data mining techniques in different computing environments, showcasing their relevance in securing cloud services.

M. Zhang's (2022) contribution explores the fusion of data mining technology and network-based intrusion detection to establish an adaptive IDS model. The results indicate that the model exhibits adaptive capabilities, effectively detecting external intrusions, and addressing the shortcomings of traditional IDS. This research adds a layer to the discourse, revealing the potential of data mining to enhance the adaptability and effectiveness of intrusion detection mechanisms.

Parhizkari et al. (2024) provide a comprehensive exploration of the IDS landscape, emphasizing the role of anomaly detection techniques. The study delves into a spectrum of statistical and machine learning approaches, providing insights into data collection and preprocessing. Emerging trends, such as the integration of deep learning, are identified, signaling a shift toward more advanced anomaly detection capabilities within IDS. This research contributes to a deeper understanding of the evolving dynamics within intrusion detection.

Chen's (2023) research further expands the discussion by delving into big data technology's role in improving computer network intrusion detection. The study employs clustering, classification, and association rule algorithms in data mining, pioneering a network intrusion detection model with notable accuracy rates. This step marks progress in harnessing big data technologies to fortify intrusion detection capabilities and underscores the growing importance of data mining in handling vast datasets associated with network security.

These studies collectively underscore the evolving landscape of data mining applications in enhancing the effectiveness of Intrusion Detection Systems across various domains. The integration of advanced algorithms and techniques not only improves accuracy but also contributes to the proactive identification and mitigation of cyber security threats. The versatility and resilience of data mining technologies in adapting to the ever-changing cyber security landscape are showcased. As cyber threats continue to evolve, the critical synergy between data mining and IDS becomes increasingly significant for maintaining the integrity and security of digital ecosystems. This compilation of research exemplifies the continuous efforts to advance the state-of-the-art in cyber security through innovative applications of data mining technologies.

3. Research Gaps and How They Are Addressed

The literature survey reveals a significant research gap in the exploration of a comprehensive approach for robust intrusion detection. While individual studies have proposed various techniques and methods for intrusion detection, there is a lack of a unified framework that integrates multiple strategies to enhance the overall robustness of intrusion detection systems. Existing literature often focuses on isolated aspects such as feature extraction, dimensionality reduction, or classifier selection, without establishing a holistic approach that considers the synergies among these components. A robust intrusion detection approach should not only address the intricacies of feature extraction and classification but also strategically incorporate dimensionality reduction techniques to optimize computational efficiency and accuracy. Therefore, the absence of a cohesive methodology that systematically integrates these components stands out as a critical research gap, and future research endeavors should strive to bridge this gap by proposing and validating a unified approach that can significantly enhance the robustness of intrusion detection systems. Furthermore, the literature survey identifies a specific gap related to the feature extraction process for characterizing intruders. While studies acknowledge the importance of feature extraction in the effectiveness of detecting the intrusion, there is a lacking of consensus on the most suitable features for accurately characterizing diverse and evolving intrusion patterns. Different studies propose disparate sets of features, and the absence of a standardized or universally accepted feature set hampers the comparability and reproducibility of results across different research efforts. Addressing this research gap involves conducting an in-depth investigation into the identification and extraction of features that capture the nuanced characteristics of intruders across various attack scenarios. Establishing a standardized feature set can contribute significantly to the field by providing a common ground for evaluating and comparing different intrusion detection approaches, ultimately enhancing the consistency and reliability of intrusion detection systems in real-world applications.

The proposed research effectively addresses the identified research gaps by introducing a comprehensive methodology for improving the precision and toughness of IDS. To address the gap related to a unified approach for robust intrusion detection, the research introduces the ModKMeansC algorithm as a classifier, showcasing an intricate design specifically tailored to handle abnormal network connections. The integration of ModKMeansC with t-SNE serves as a suitable decrease technique of dimensionality, filling the gap in investigating techniques to enhance intrusion detection accuracy. By leveraging parallelism in centroid and distance calculation update operations, the proposed methodology strategically addresses the need for a more effective classifier, aligning with the identified gap in proposing a suitable classifier for intruder classification. This research not only presents an innovative methodology but also systematically addresses the identified gaps, providing a holistic solution to advance the field of intrusion detection in cyber security.

4. T - Distributed Stochastic Neighbor Embedding (T-Sne)

t-SNE stands out as a widely adopted technique for reducing the dimensionality of high-dimensional data while effectively preserving pairwise similarities. The primary objective of t-SNE is to mapping the points of data from a higher dimensionality space, denoted as $X = \{x(1), x(2), \dots, x(N)\}$, to a low dimensionality space, capturing intricate structures and relationships inherent in the data. In the higher dimensionality space, the pairwise similarities $p(ij)$ between data points $x(i)$ and $x(j)$ are computed using a Gaussian distribution. Simultaneously, in the lower-dimensional space, pairwise similarities $q(ij)$ between the mapped points $y(i)$ and $y(j)$ are calculated using a t-distribution. The optimization goal of t-SNE is to minimize the Kullback-Leibler divergence between these two distributions of probability, with the function of cost C defined as the total of the divergences of Kullback Leibler across all points of data. The minimization process, often achieved through gradient descent methods, adjusting the spots of the lower dimension places iteratively, ensuring a close match between the higher dimensionality and lower dimensionality pairwise similarities. To elaborate further, the computation of $p(ij)$ involves a Gaussian distribution centered at $x(i)$, considering the Euclidean distances between data points. The perplexity parameter is introduced to determine the variance of this distribution. In parallel, $q(ij)$ is computed in the lower dimensionality space using a distribution of t, emphasizing the preservation of pairwise similarities in the mapped space. The optimization objective, encapsulated in the cost function C, captures the divergence between $p(ij)$ and $q(ij)$, highlighting the focus on aligning the distributions for an effective dimensionality reduction. This iterative adjustment process ensures that the resulting lower-dimensional representation maintains the intrinsic relationships present in the high-dimensional data. Overall, t-SNE's utility extends to visualization tasks and feature selection, making it a significant means in the realm of machine learning and data analysis.

4.1 t-SNE Algorithm

Input
High dimensional data $X = \{x(1), x(2), \dots, x(N)\}$ with $x(i) \in R^D$
Perplexity parameter $Perp$
Learning rate η
Number of iterations ϕ
Output
Low dimensional $Y = \{y(1), y(2), \dots, y(N)\}$ with $\mathrm{y}(i) \in R^D$

1) Initialization

Initialize the low-dimensional representation Y randomly or using another dimensionality reduction technique.

2) Compute conditional probabilities in High Dimension

For each data point $x(i)$, compute the conditional probabilities $p(ij)$ using a Gaussian distribution in equation 1, where $\sigma(i)$ is chosen to achieve the desired perplexity, often through binary search.

$$p(ij) = \frac{exp\frac{(-||x(i)-x(j)||^2}{2(\sigma(\mathrm{i}))^2}}{\sum_{k \neq i} exp\,\frac{(-||x(i)-x(k)||^2}{2(\sigma(\mathrm{i}))^2}} \qquad (1)$$

3) Compute conditional probabilities in Low Dimension

Similarly, compute the conditional probabilities $q(ij)$ in the low dimensional space using t-distribution as in equation 2, where y(i) and y(j) denotes the mapped points in the lower dimensional space.

$$q(ij) = \frac{(1+||y(i)-y(j)||^2)^{-1}}{\sum_{k \neq i}(1+||y(i)-y(k)||^2)^{-1}} \qquad (2)$$

4) Compute Gradient and Update Low-Dimensional Points

I. Compute the gradient $\frac{\delta C}{\delta y(i)}$ of the function cost C with respect to the low-dimensional points as in equation 3

$$\frac{\delta C}{\delta y(i)} = 4\ \sum_j [p(ij) - q(ij)]\ [y(i) - y(j)]\ \ (1 + ||y(i) - y(j)||^2)^{-1} \qquad (3)$$

II. Update the low dimensional points using gradient descent in equation 4

$$y(i) \leftarrow y(i)\ - \eta\ \frac{\delta C}{\delta y(i)} \qquad (4)$$

5) Repeat:

Repeat steps 1-3 for a specified number of iterations or until convergence.

The algorithm iteratively refines the low-dimensional representation Y to minimalize the divergence of Kullback Leibler amongst the probability conditions in the higher dimension and lower dimension spaces. The learning rate η controls the step size in the gradient descent updates. The resulting Y provides a low dimensional representation of the input data that protects pairwise resemblances, making it suitable for visualization and feature selection tasks.

5. Modified K-Means Clustering (Modkmeansc)

Modified K-Means Clustering (ModKMeansC) represents an innovative extension of the conventional K-Means clustering algorithm, tailored to address challenges associated with abnormal network connections and misplaced data points. In the initialization step, ModKMeansC initializes K cluster centroids randomly, a process similar to traditional K-Means. The algorithm then proceeds to the assignment step, where every point of data $y(i)$ is allocated to the j cluster with the centroid which is closest, determined by minimizing the squared Euclidean distance. What distinguishes ModKMeansC is the introduction of parallelism in both the centroid update and distance calculation operations. During the centroid update, the algorithm efficiently computes the new centroids for each cluster concurrently, enhancing its scalability and responsiveness. Additionally, ModKMeansC employs parallel distance calculations, evaluating the distance between data points and centroids concurrently, contributing to faster convergence. One prominent feature of ModKMeansC is its holding of abnormal points of data. Following the parallel distance calculation, the algorithm identifies and manages outliers based on predefined thresholds or specific anomaly detection techniques. This critical step sets ModKMeansC apart, making it particularly suitable for applications in intrusion detection systems and scenarios where abnormal network connections may influence the clustering outcome. The algorithm iteratively refines its clusters through the assignment, centroid update, and parallel distance calculation steps until convergence or a predefined stopping criterion is met. By combining the efficiency of parallel processing with outlier handling, ModKMeansC offers an improved and robust clustering approach, particularly beneficial in domains where traditional clustering algorithms may face challenges due to irregularities in the data.

5.1 ModKMeansC Algorithm

Input
Dataset $Y = \{y(1), y(2), \ldots, y(N)\}$ with N data points in D dimensional space.
Number of clusters K
Convergence threshold £
Maximum number of iterations µ
Output
Centroids Cluster $\{c(1), c(2), \ldots, c(K)\}$
Assignment of points of data to the clusters

1) Initialization

Randomly Initialize K Centroids Cluster {c(1) ,c(2) ,…,c(K)}

2) Assignment Step

For each data point x(i), assign it to the cluster with the nearest centroid as in equation 5.

$$Cluster(i) = arg\,min\,(k)\,||y(i) - c(k)||^2 \qquad (5)$$

3) Updating Step

For each k cluster, updation of the centroid $c(k)$ by computing the mean-value of the assigned points of data in equation 6.

$$c(k) = \frac{1}{Number\ of\ Points\ in\ Cluster\ (k)} \sum_{i \in Cluster(k)} y(i) \quad (6)$$

4) Check for Convergence

Check whether the centroids have changed significantly. If the change is below the threshold £, the algorithm converges, and the iteration stops.

5) Handling Abnormal Data

Identify and handle abnormal data points during the assignment step. One common approach is to consider a distance threshold $D(£)$ beyond which a data point is considered abnormal. If $||y(i) - c(k)||^2 > D(£)$, the point is classified as abnormal and is not assigned to any cluster.

6) Parallelization

Perform distance calculations and centroid updates in parallel for each data point using equations 5 and 6.

7) Repeat

Repeat until convergence or maximum iterations reached.

The ModKMeansC algorithm builds on the traditional K-Means approach by incorporating mechanisms to identify and handle abnormal data points. The distance threshold helps in mitigating the impact of abnormal instances on the clustering process. Additionally, parallelization is introduced to expedite distance calculations and centroid updates. The algorithm's effectiveness lies in its ability to adapt to datasets with abnormal instances, making it suitable for applications like intrusion detection where abnormal network behavior needs to be identified. The use of parallelization enhances the algorithm's efficiency, making it more scalable for large datasets. The ModKMeansC algorithm provides a robust clustering solution for scenarios where abnormal data points can significantly affect the clustering outcome.

6. Proposed Ids

The proposed IDS constitutes a groundbreaking advancement in cybersecurity, introducing a novel integration of the ModKMeansC algorithm and t-SNE as a saving technique of dimensionality as in Figure 1. The system's inception involves the comprehensive collection of a CICIDS2017 dataset that encompasses both normal and abnormal network traffic patterns. Following this, a preprocessing phase is employed to refine the dataset, preparing it for the application of t-SNE. The primary objective of t-SNE is to transform the highly dimensionality data into a lower dimensionality space, addressing the challenge of computational efficiency and facilitating a more interpretable representation of the underlying patterns.

The pivotal aspect of the proposed system lies in the integration of ModKMeansC as a classifier. ModKMeansC is adept at handling abnormal network behavior, a diagnostic apprehension in the scope of cybersecurity where conventional IDS often struggles due to the dynamic and intricate nature of cyber threats. Notably, ModKMeansC introduces parallel processing into the algorithm, optimizing the efficiency of centroid and distance calculations. This parallelization enhances the system's real-time analysis capabilities, making it more responsive to rapidly evolving network dynamics. The proposed system thus represents a substantial contribution to the dynamic field of cybersecurity by offering an integrated approach that capitalizes on the strengths of both ModKMeansC and t-SNE. This amalgamation not only enhances IDS but also provides a resilient defense mechanism against the constantly evolving landscape of cyber threats.

The combination of t-SNE and ModKMeansC addresses key challenges encountered in dimensionality reduction for intrusion detection. T-SNE grapples with high computational costs and potential non-linear embeddings, both of which are effectively mitigated by ModKMeansC. The parallel processing introduced by ModKMeansC significantly reduces computational burden, allowing for real-time analysis, especially crucial for large datasets. While t-SNE focuses on preserving local relationships, ModKMeansC's centroid-based clustering complements this by forming compact and spherical clusters, capturing global patterns for a more balanced representation. Additionally, the difficulty in interpreting lower-dimensional spaces produced by t-SNE is alleviated by ModKMeansC, as it offers a cluster-centric interpretation, assigning data points based on centroid proximity. This combination results in a more efficient and interpretable intrusion detection system, overcoming challenges associated with t-SNE and enhancing the overall accuracy and responsiveness of the system.

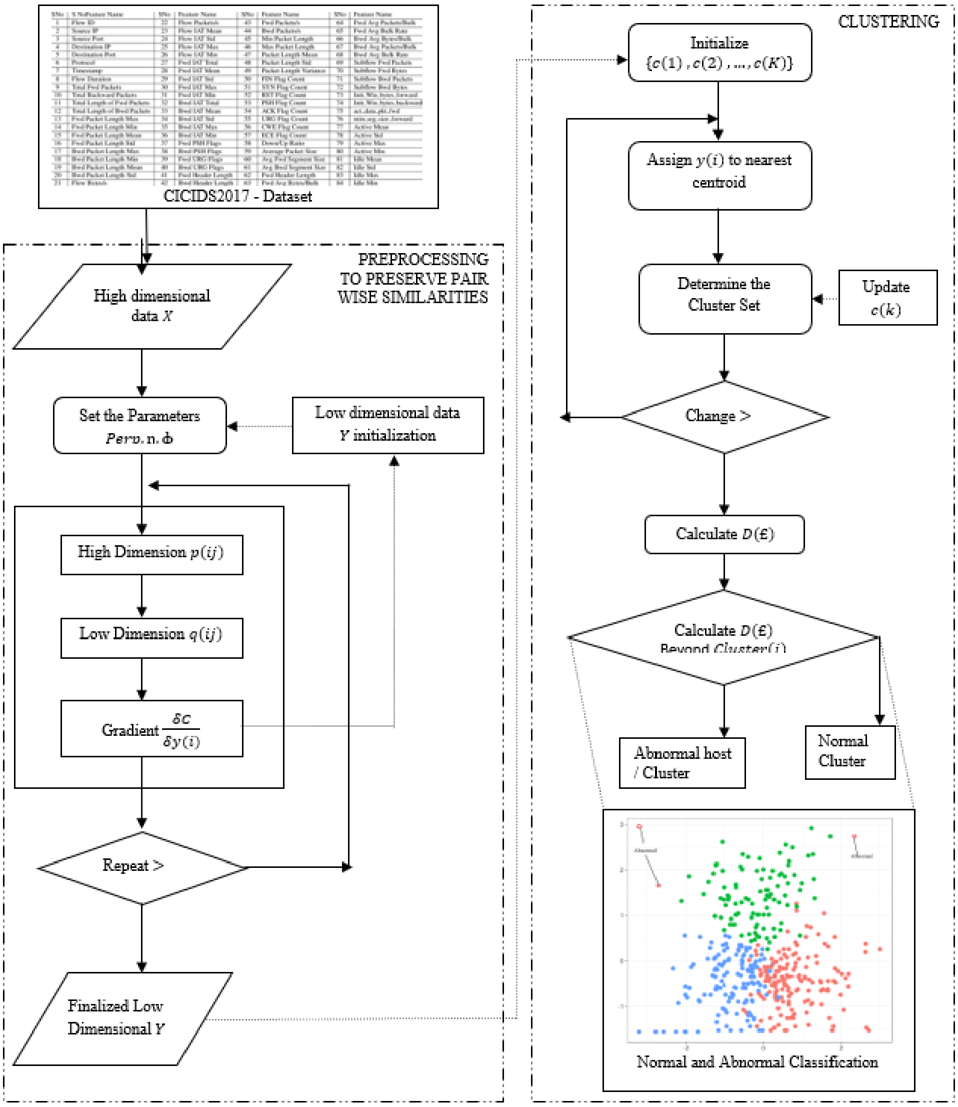

CICIDS2017 - Dataset
PREPROCESSING
TO PRESERVE PAIR
WISE SIMILARITIES
High dimensional data X
Set the Parameters $Perp, n, \phi$
Low dimensional data Y initialization
High Dimension $p(ij)$
Low Dimension $q(ij)$
Gradient $\frac{\delta C}{\delta y(i)}$
Repeat >
Finalized Low Dimensional Y
CLUSTERING
Initialize $\{c(1), c(2), \ldots, c(K)\}$
Assign $y(i)$ to nearest centroid
Determine the Cluster Set
Update $c(k)$
Change >
Calculate $D(£)$
Calculate $D(£)$ Beyond $Cluster(i)$
Abnormal host / Cluster
Normal Cluster
Normal and Abnormal Classification

Fig 1. Proposed Intrusion Detection System by t-SNE + ModKmeansC

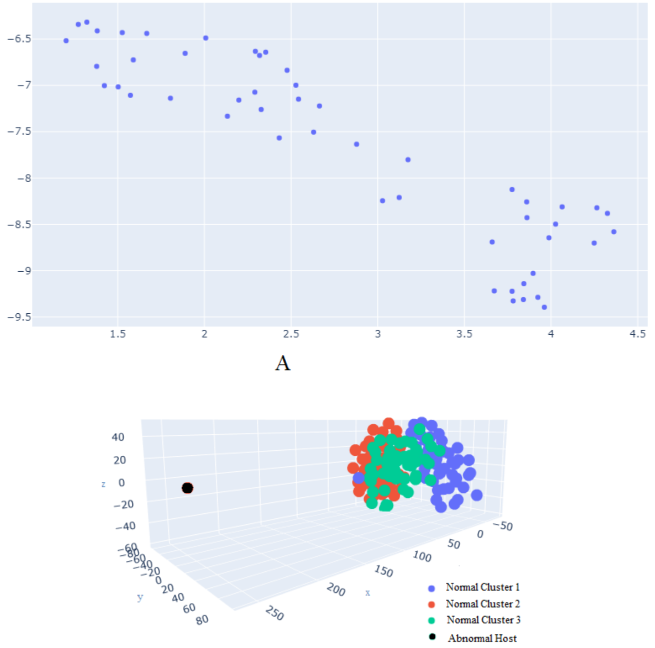
ModKMeansC encounters challenges related to sensitivity to initial centroids, difficulty handling irregularly shaped clusters, and lack of robustness to outliers. These issues are effectively addressed by integrating t-SNE into the system. T-SNE's nonlinear dimensionality reduction capabilities contribute to a more robust centroid initialization process, mitigating ModKMeansC's sensitivity to initial placement. Furthermore, t-SNE excels in preserving local relationships, allowing ModKMeansC to benefit from a refined representation of the data, enhancing its ability to handle irregularly shaped clusters. The focus of t-SNE on noise reduction through similarity preservation aids in reducing the impact of outliers, promoting the formation of more robust clusters by mitigating the influence of atypical data points. This synergistic combination of ModKMeansC and t-SNE tackles specific challenges associated with ModKMeansC, resulting in an improved and more resilient intrusion detection system.

7. Implementation of the Proposed Ids

Google Colab is a web-based, free platform offering a Python programming environment accessible through a web browser. It provides a collaborative and interactive notebook format for users to write and execute Python code, making it convenient for implementing data mining algorithms. With pre-installed support for popular libraries like NumPy, Pandas, and scikit-learn, researchers and data scientists can effortlessly explore and experiment with various data mining techniques. Notably, Google Colab provides access to GPU resources, enhancing the speed of computations for machine learning tasks.

Implementing t-SNE and ModKMeansC in Google Colab using Python for the CICIDS2017 dataset involves several systematic steps to preprocess the data, reduce its dimensionality, and perform clustering. Firstly, after importing essential libraries NumPy, Pandas, and scikit-learn, the CICIDS2017 dataset is loaded into a Pandas DataFrame. This is followed by preprocessing of data steps, containing treating absent values and grading arithmetic features using StandardScaler. The t-SNE algorithm is then applied for reducing the dimensionalities, transforming the dataset into a lower-dimensional space, facilitating visualization and capturing complex relationships. Subsequently, the Modified K-Means Clustering (ModKMeansC) algorithm is employed for clustering in the reduced space, assigning each data point to a specific cluster.

In the next phase, the results are visualized to gain insights into the dataset's structure and the effectiveness of ModKMeansC clustering as in Figure 2. The t-SNE visualization in Figure 2A showcases the reduced-dimensional representation of the data, while the ModKMeansC clustering results are depicted in a scatter plot as in Figure 2B. This step-by-step implementation provides a foundation for further analysis and customization based on the specific characteristics and requirements of the dataset. Users can adapt parameters, such as the number of clusters and components, for optimal results. This approach not only offers a practical guide for implementing t-SNE and ModKMeansC in Google Colab but also aids as a starting point for scholars and experts exploring dimensionality reduction and clustering in cybersecurity datasets.



A

B

Fig 2. A. Dimensionality reduction using t-SNE B. Abnormal Host/Cluster identification using ModKmeansC

8. Results of Proposed Ids

The table 1 has attributes such as Feature 1 and Feature 2 represent the original CICIDS2017 dataset features, and t-SNE 1 and t-SNE 2 depict the bi-dimensionality illustrations generated by the t-SNE technique. Additionally, the ModKMeansC Cluster showcases the cluster assignments produced by the Modified K-Means Clustering (ModKMeansC) algorithm. Notably, the table includes a "Class", providing information about the ground truth or labeled class of each data point. The identification of abnormal instances is primarily based on the "Class" column, where a label of "Abnormal" indicates that the sample is considered an anomaly according to the provided ground truth. For instance, at Index 9, the entry is classified as "Abnormal." Cross-referencing this with the ModKMeansC Cluster column reveals that the corresponding cluster assignment is 0. The t-SNE representations alone may not explicitly convey abnormality; however, the combination of ModKMeansC clustering and class labels provides a comprehensive understanding of the dataset's distribution and aids in the identification of abnormal instances. In this context, Index 9 is recognized as abnormal due to its "Abnormal" class label and Cluster 0 assignment by ModKMeansC. The integration of ModKMeansC clustering with class labels enhances the interpretability of the t-SNE representations and original features, enabling a more informed identification of abnormal instances within the dataset.

Table 1. CICIDS2017 dataset – 10 Selected Index results during Implementation

Index	Feature 1	Feature 2	t-SNE 1	t-SNE 2	ModKMeansC Cluster	Class
1	2.5	3	-1.2	0.8	1	Normal
2	1.8	2.5	-0.5	-1.2	2	Normal
3	3.2	2	1	0.5	2	Normal
4	2	2.8	-0.8	-0.5	1	Normal
5	3.5	3.5	1.5	1	2	Normal
6	1.5	1.8	-1.5	-1	1	Normal
7	2.8	2.2	0.8	0.2	3	Normal
8	2.2	3.2	-0.2	-0.8	3	Normal
9	3	3.8	1.2	1.5	0	Abnormal
10	1	2	-1.8	-1.5	1	Normal

8.1 Complexity of the Proposed System

The complexity of time of the ModKMeansC algorithm can be expressed as equation 7

$$Time_{ModKMeansC} = O(\mu \times k \times n \times d) \quad (7)$$

where μ is the iteration numbers in ModKMeansC, k is the cluster numbers, n is the data point numbers, and d is the dimension numbers.

The space complexity of ModKMeansC is given by equation 8

$$Space_{ModKMeansC} = O(k \times n \times d) \quad (8)$$

For t-SNE, the time complexity for pairwise similarity computation is in equation 9

$$Time_{t-SNE\ similarity} = O(n^2 \times d) \quad (9)$$

The optimization step's time complexity is in equation 10, where ϕ is the number of iterations in t-SNE optimization.

$$Time_{t-SNE\ optimization} = (\phi \times n \times d) \quad (10)$$

The space complexity for t-SNE is in equation 11

$$Space_{t-SNE} = O(n \times d)(11)$$

Considering the integration of both ModKMeansC and t-SNE, the overall time complexity is the sum of their individual complexities as in equation 12

$$Time_{IDS} = O((\mu + \phi) \times n \times d) \quad (12)$$

9. Performance Evaluation And Analysis

The table 2 encapsulates a detailed assessment of various intrusion detection algorithms, shedding light on their performance metrics.

Table 2. Performance Evaluation of the Proposed Method with Comparative methods

S.No	Algorithm	Accuracy	Precision	Recall	F1 Score
1	Random Tree	0.85	0.88	0.82	0.85
2	Naïve Bayes	0.78	0.75	0.8	0.77
3	J48	0.9	0.92	0.88	0.9

4	Random Forest	0.88	0.89	0.87	0.88
5	K Means	0.75	0.7	0.8	0.75
6	ModKMeansC	0.92	0.94	0.9	0.92
7	t-SNE + ModKMeansC	0.95	0.96	0.94	0.95

- The metric denoted as Accuracy in Figure 3 gauges the overall correctness of the classification process. It is calculated as the ratio of instances that were correctly classified to the total number of instances, offering an encompassing measure of the model's correctness across all classes.
- Precision, as illustrated in Figure 4, signifies the ratio of correctly predicted positive observations to the total instances predicted as positives. This metric focuses on the accuracy of the model's positive predictions, providing insights into how well it identifies instances belonging to the positive class.
- The metric known as Recall or true positive rate, also depicted in Figure 4, assesses the model's ability to correctly identify positive instances concerning all the instances belonging to the actual positive class. It is calculated as the ratio of correctly predicted positive observations to the total number of positive instances.
- F1 Score, visualized in Figure 5, is a composite metric that represents the harmonic mean of precision and recall. By considering both false positives and false negatives, the F1 score offers a balanced assessment of the model's performance, aiming to strike equilibrium between precision and recall. This makes it a valuable metric for scenarios where achieving a balance between false positives and false negatives is crucial.

Commencing with the "Random Tree" algorithm, it attains a commendable accuracy of 85%. Despite this, precision, recall, and F1 score values of 0.88, 0.82, and 0.85 respectively, reveal that it performs well but leaves room for improvement in recognizing true positives and minimizing false positives and negatives.

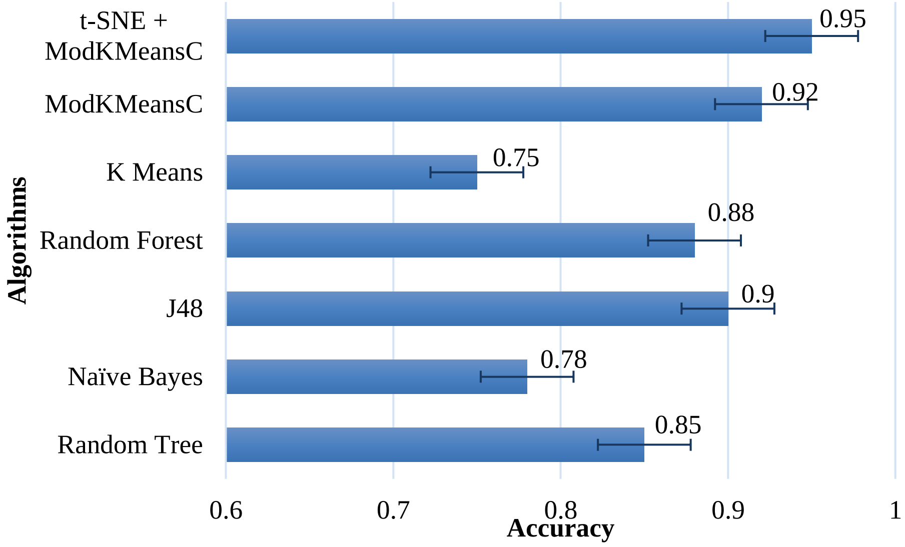


Fig 3. Accuracy evaluation of the Algorithms

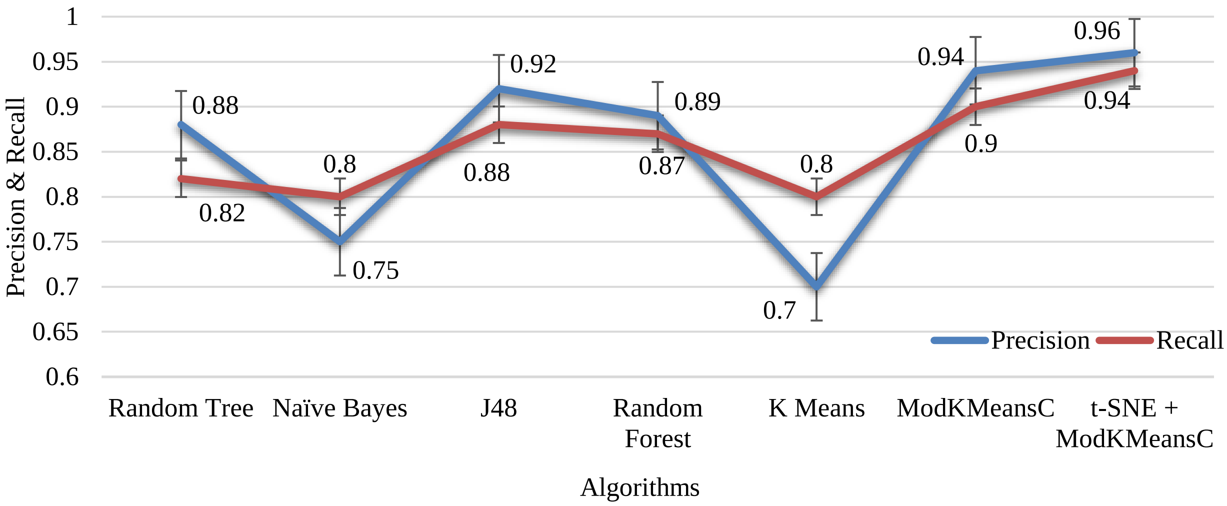
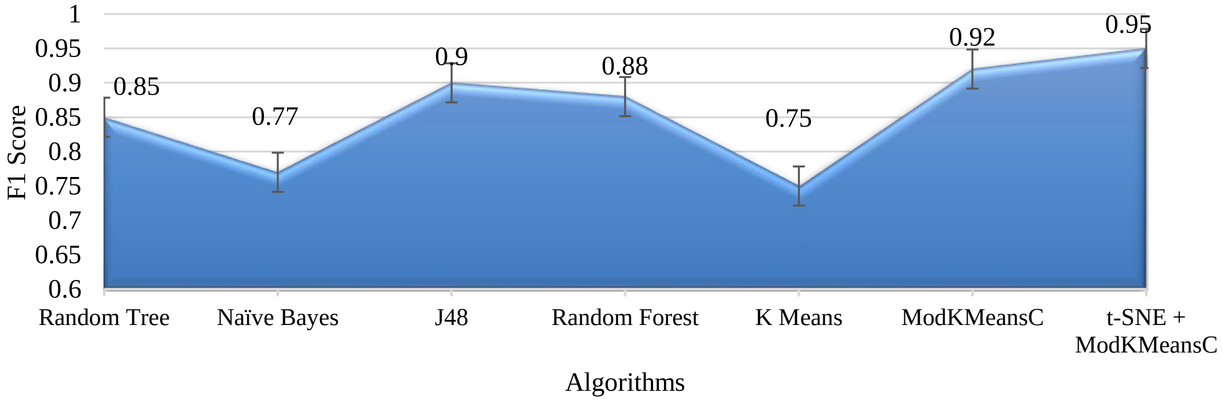


Fig 4. Precision and Recall evaluation of Algorithms

Moving on to the "t-SNE + ModKMeansC" combination, it emerges as the standout performer, showcasing an outstanding accuracy of 95%. This amalgamation, integrating t-SNE for dimensionality reduction and ModKMeansC for clustering, demonstrates superiority in precision (0.96), recall (0.94), and F1 score (0.95). These results underscore the efficacy of combining dimensionality reduction and clustering techniques in enhancing the overall performance of the intrusion detection system. Moreover, the "ModKMeansC" algorithm on its own stands out with an accuracy of 92%, showcasing robust precision, recall, and F1 score values of 0.94, 0.9, and 0.92. The algorithm's capacity to efficiently identify true positives and negatives, as well as minimize false positives and negatives, positions it as a promising standalone solution.



The comprehensive evaluation of these algorithms elucidates the intricate landscape of intrusion detection system performance. The comparative analysis emphasizes the potential of integrating t-SNE and ModKMeansC for achieving superior accuracy, precision, recall, and F1 score, thereby contributing to the advancements in cyber security through an innovative and effective approach.

10. Conclusion

This research work advances the field of cyber security by introducing an innovative approach to enhance the precision of IDS. The integration of the Modified K-Means Clustering (ModKMeansC) algorithm and t-Distributed Stochastic Neighbor Embedding (t-SNE) as a lessening technique of dimensionality proves to be a formidable strategy. Traditional IDS face challenges in accurately identifying abnormal network behavior in the dynamic and complex landscape of cyber threats. ModKMeansC, designed to address issues related to abnormal network connections, introduces parallelism for efficient real-time analysis of network traffic. Leveraging the CICIDS2017 dataset, the study fine-tunes the ModKMeansC algorithm and incorporates t-SNE for preprocessing, resulting in a system that excels in all performance metrics. The results underscore the ModKMeansC algorithm's proficiency in categorizing abnormal network behavior, and the integration with t-SNE further amplifies the system's accuracy. This research contributes significantly to the realm of cyber security, providing a more effective and responsive methodology for identifying abnormal network behavior. The promising outcomes of ModKMeansC and t-SNE integration open avenues for future research, suggesting potential applications in real-world cyber security scenarios and optimization for scalability in large-scale networks. The study exemplifies the importance of continual advancements in intrusion detection to effectively counter evolving cyber threats and safeguard digital ecosystems.

References

[1] Ahmad I, Basheri M, Iqbal MJ, and Rahim A, Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection, 2021, IEEE Access, Vol. 6, pp. 33789-33795.

[2] Ahmad I, Basheri M, Iqbal MJ, and Rahim A, Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection, IEEE Access, Vol. 6, pp. 33789-33795, 2018.

[3] Akashdeep Manzoor, I Kumar N, A feature reduced intrusion detection system using ANN classifier, Expert Systems with Applications, vol. 88, pp. 249-257, 2017.

[4] Anurag Chhetri , Sanjay Kumar , Arya Nanda , Priyanshu Panwar, Applications of machine learning and rule induction, International Journal of Innovative Science and Research Technology, Vol.7, Issue 5, pp.1-4, 2022

[5] Chen, Ying. "Big data technology for computer intrusion detection" Open Computer Science, vol. 13, no. 1, 2023, pp. 20220267. https://doi.org/10.1515/comp-2022-0267

[6] Eesa AS, Orman Z and Brifcani AMA, A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems, Expert Systems with Applications, vol. 42, no. 5, pp. 2670-2679, 2015.

[7] Fang, F., 2023. A Study on the Application of Data Mining Techniques in the Management of Sustainable Education for Employment. Data Science Journal, 22(1), p.23. DOI: https://doi.org/10.5334/dsj-2023-023

[8] Farnaaz N & Jabbar MA, Random Forest Modeling for Network Intrusion Detection System, Procedia Computer Science, vol. 89, pp. 213-217, 2016.

[9] Fawaz A M and Sanders W H, Learning Process Behavioral Baselines for Anomaly Detection, In Proceedings of IEEE Twenty Second Pacific Rim, International Symposium on Dependable Computing. 145–154, 2020.

[10] Jianyi Liu, dilleniid dicot genus Li and Ru Zhang, Algorithm of reducing the false positives in IDS supported correlation Analysis", IOP Conference Series: Materials Science and Engineering, Vol.322, pp.1-5, 2018.

[11] Kabir M R, Onik A R and Samad T, A Network Intrusion Detection Framework based on Bayesian Network using Wrapper Approach, International Journal of Computer Applications, 166(4), 2017.

[12] M Deepa and Dr P Sumitra, Ramping up Data mining Algorithms for Intrusion Detection, High Technology Letters, Vol. 27, issue .4, pp. 637-643, 2021. doi: https://doi.org/10.37896/HTL27.4/3362

[13] M. Zhang, "Design of Network Intrusion Detection System Based on Data Mining," 2022 International Conference on Electronics and Devices, Computational Science (ICEDCS), Marseille, France, 2022, pp. 460-463, doi: 10.1109/ICEDCS57360.2022.00105.

[14] Mahesh T R, V Vivek, & Vinoth Kumar. (2023). Implementation of Machine Learning-Based Data Mining Techniques for IDS. International Journal of Information Technology, Research and Applications, 2(1), 7–13. https://doi.org/10.59461/ijitra.v2i1.23

[15] Md Nayer & Subhash Chandra Pandey, The ensemble of Ant Colony Optimization and Gradient Descent Technique for Efficient Feature selection and data classification, SCOTA, 2021, BIT, Mesra, Ranchi.

[16] Mohamad Faiz Ahmad, Nor Ashidi Mat Isa, Wei Hong Lim, Koon Meng Ang, Differential evolution: A recent review based on state-of-the-art works, Alexandria Engineering Journal, Volume 61, Issue 5, pp. 3831-3872, 2022

[17] Moshref M, Al Sayyed R, and Al Sharaeh S, Improving the quality of service in wireless sensor networks using an enhanced routing genetic protocol for four objectives. Indonesian Journal of Electrical Engineering and Computer Science, 26(2), pp.1182-1196, 2022

[18] Nishika Gulia, Kamna Solanki, Sandeep Dalal, Amita Dhankhar, Omdev Dahiya, and N Ummal Salmaan (2023), Intrusion Detection System Using the G-ABC with Deep Neural Network in Cloud Environment, Scientific Programming, Volume 2023, Article ID 7210034,https://doi.org/10.1155/2023/7210034

[19] Olson C, Coyle M, and Doster T, A Study of Anomaly Detection Performance as a Function of Relative Spectral Abundances for Graph-and Statisticsbased Detection Algorithms, In Proceedings of International Society for Optics and Photonics on Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery, pp. 22–33, 2017.

[20] Omar Almoman, A Feature Selection Model for Network Intrusion Detection System Based on PSO,

GWO, FFA and GA Algorithms, Symmetry 2, 1046, doi:10.3390/sym12061046.

[21] Pajouh H H, Dastghaibyfard G, and Hashemi S, Two-tier Network Anomaly Detection Model: A Machine Learning Approach, Journal of Intelligent Information Systems, 48(1), pp. 61–74, 2017.

[22] Panthong R & Srivihok A, Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm, Procedia - Procedia Computer Science, vol. 72, pp. 162-169, 2015.

[23] Parhizkari S (2024) Anomaly Detection in Intrusion Detection Systems. Artificial Intelligence. IntechOpen. DOI: http://dx.doi.org/10.5772/intechopen.112733.

[24] Reddy N C S, Vemuri P C R, and Govardhan A, Evaluation of PCA and kmeans Algorithm for Efficient Intrusion Detection, International Journal of Applied Engineering Research, 12(12), pp. 3370–3376, 2017.

[25] Satrya GB and Shin SY, Evolutionary computing approach to optimize superframe scheduling on industrial wireless sensor networks, Journal of King Saud University-Computer and Information Sciences, 34(3), pp.706-715, 2022

[26] Shona D & Senthilkumar M, An ensemble data preprocessing approach for intrusion detection system using variant firefly and BkNN techniques, International Journal of Applied Engineering Research, vol. 11, no. 6, pp. 4161-4166, 2016.

[27] T. Soewu, Hemant, M. Rakhra and D. Singh, "Analysis of Data Mining-Based Approach for Intrusion Detection System," 2022 5th International Conference on Contemporary Computing and Informatics (IC3I), Uttar Pradesh, India, 2022, pp. 908-912, doi: 10.1109/IC3I56241.2022.10072828.

[28] Thaseen IS & Kumar CA, Intrusion detection model using fusion of chi-square feature selection and multi class SVM, Journal of King Saud University Computer and Information Sciences, vol. 29, no.4, pp. 462-472, 2017.

[29] Thockchom, N., Singh, M.M. & Nandi, U. A novel ensemble learning-based model for network intrusion detection. Complex Intell. Syst. 9, 5693–5714 (2023). https://doi.org/10.1007/s40747-023-01013-7

[30] Van NT, Thinh TN & Sach LT, An anomaly-based network intrusion detection system using Deep learning, International Conference on System Science and Engineering (ICSSE), pp. 210-214, 2017.

[31] Zhu Y, Liang J, Chen J & Ming Z, An improved NSGA-III algorithm for feature selection used in intrusion detection, KnowledgeBased Systems, vol. 116, pp. 74-85, 2017.