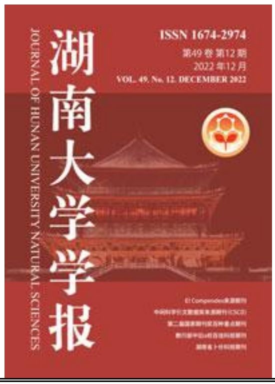


湖南大学学报（自然科学版）
Journal of Hunan University (Natural Sciences)

第 49 卷 第 12 期
2022 年 12 月

Available online at http://jonuns.com/index.php/journal/index

Vol. 49 No. 12
December 2022

Open Access Article https://doi.org/10.55463/issn.1674-2974.49.12.7

An Improved Approach Based on Density-Based Spatial Clustering of Applications with a Noise Algorithm for Intrusion Detection

Shahneela Pitafi*, Toni Anwar, Zubair Sharif

Department of Computer and Information Sciences, Universiti Teknologi Petronas, Bandar Seri Iskandar, Perak, 32610, Malaysia

* Corresponding author: shahneela_22000124@utp.edu.my

Received: August 5, 2022 / Revised: October 7, 2022 / Accepted: November 2, 2022 / Published: December 30, 2022

Abstract: Network Intrusion detection systems (NIDS) are extremely important for make the network secure from unauthorized access. Numerous studies have already been conducted to detect the unauthorized access to achieve security. As the NIDS are still lacking in terms of accuracy, true positive rate (TPR) and the false positive rate (FPR) of the invasive events. The main cause of high FPR in intrusion detection systems is run with a default set of signatures. Issues in the detection rate are caused by feature similarities between man-made events and environmental events. Considering this fact, in this paper, we introduced a new intrusion detection algorithm named as I-DBSCAN by focusing on the above-mentioned issues to get the better results from the previously done experiments. We used clustering and classification techniques. The proposed algorithm is an enhanced version of the existing DBSCAN algorithm. However, this research can spot attacks on data from IDS. It is found that the novel algorithm achieved more accuracy when it is applied to four classification methods on KDD Cup 99 and NSL-KDD Cup99 data. The results of our proposed methodology are more efficient with the achievement of better accuracy level and false positive rate (FPR).

Keywords: density-based spatial clustering of applications with noise, false positive rate, intrusion detection system, network intrusion detection system.

一种改进的基于密度的空间聚类应用噪声算法的入侵检测方法莎妮拉·皮特菲、托尼·安瓦尔、祖拜尔·谢里夫

摘要：网络入侵检测系统(NIDS)对于保护网络免受未经授权的访问非常重要。已经进行了大量研究来检测未经授权的访问以实现安全性。由于 NIDS 在入侵事件的准确性、真阳性率(热塑性弹性体)和假阳性率(FPR)方面仍然存在不足。入侵检测系统中高 FPR 的主要原因是使用默认签名集运行。检测率的问题是由人为事件和环境事件之间的特征相似性引起的。考虑到这一事实，在本文中，我们针对上述问题引入了一种名为数据库扫描仪的新入侵检测算法，以从先前所做的实验中获得更好的结果。我们使用了聚类和分类技术。所提出的算法是现有数据库扫描算法的增强版本。然而，这项研究可以发现对来自入侵检测系统的数据的攻击。结果发现，当将新算法应用于 KDD 杯 99 和 NSL-KDD 99 杯数据的四种分类方法时，

其准确性更高。我们提出的方法的结果更有效，实现了更好的准确度水平和误报率(FPR)。
关键词：具有噪声、误报率、入侵检测系统、网络入侵检测系统的应用程序的基于密度的空间聚类。

1. Introduction

A secure computer or network system should provide the services of data confidentiality, data and communications integrity, and assurance against denial-of-service to achieve these services network may combine several strategies to provide a comprehensive security system. Furthermore, current systems typically include an intelligence that comes naturally which allows for use in real-time sensor surveillance through a control center [1-3]. Security has grown to be a key worry as technology and automation progress. A security system, which immediately notifies the owners of any intrusion, is always the first line of defense for any property or network. Numerous security systems available today use various motion sensors to detect any movement and alert the owner about an entry. A network intrusion detection system (NIDS) is a tool or sensor that recognizes the presence of an intruder trying to access the data or tries to damage the confidentiality of the data [4]. In both the detection and prevention perspectives of attacker's information is critical to lowering the frequency of untrue alarms and improve the security systems efficiency.

For improving security numerous studies were conducted and yet many are ongoing on the intrusion detection, current systems still must differentiate between an intrusion and a nuisance. As such, the existing network intrusion detection systems (NIDS) have yet to establish a balance between the accuracy of detection (AOD) and false positive rate (FAR) [5]. Four forms of attacks (sequential, over-soliciting, temporal, and direct) were explored by a method proposed for spotting fraudulent commands in separate systems. To detect malicious commands that pass to the physical system from the control system, the Security Approach based on Filter Execution (SAFE) method was used [6]-[7]. The application of the intrusion detection system to the CPS was discussed. A CPS integrated with an intrusion detection system has been developed by the authors. The inspection of the CPS's unique qualities and requirements for dependability and security resulted in the development of a design platform [8]-[9]. For the first time, a fiber laser cavity was used in a fiber-optic multi-zone perimeter intrusion detection system. Experiments were conducted in four distinct weather situations, with a zero FAR as a consequence [10]-[11].

Applications for the Internet of Things can be anything from a fundamental device for a smart home to a specialized device for a smart grid, as shown in Figure 1. The IoT offers society worldwide a massive opportunity. Contrasting IoT apps share several traits while having diverse goals [12]-[14].

Traditional Intrusion Detection systems (IDS) lack in accuracy, false positive rates, and true positive rates of invasive events that remains a contentious issue in the field of detection and identification [15]. Therefore, to overcome those issues in IDS, we introduced a novel algorithm that focuses on the above-mentioned issues for improving the recognition and accuracy. We applied the improved DBSCAN algorithm on the KDD Cup99, NSL-KDD Cup99 datasets to achieve better accuracy and elimination of false positive rate (FPR). Furthermore, we applied K-NN, SVM, Random Forest, and Naïve Bayes as classifier, it is found that the novel algorithm achieved more accuracy when it was applied to K-NN method. This evaluation was performed by measuring the accuracy of the attack classification.

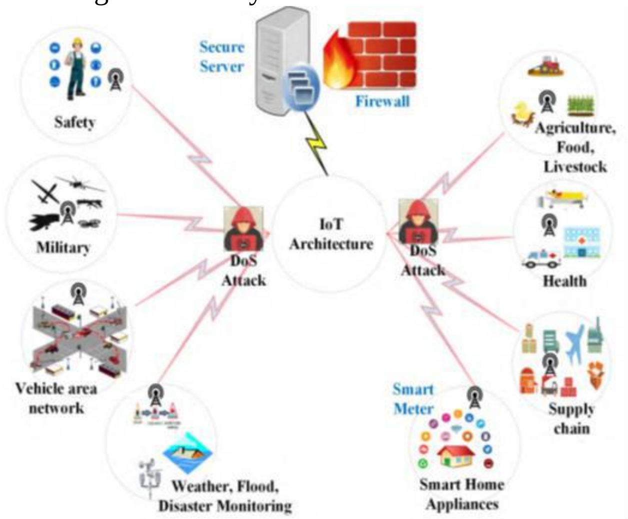


Fig. 1 An example of IOT applications

The rest of this paper is organized as follows: Section 2 discusses the literature review, and Section 3 describes the proposed methodology with the necessary explanations. Section 4 contains the results and discussion along with comparisons to the other related techniques, and Section 5 presents the conclusions and developments that can be continued in future work.

2. Related Work

Authors in [16] proposes an intrusion detection system (ML-IDS) based on ML for detecting IoT network threats. The prime goal of this study was to use ML-supervised algorithm-based IDS for IoT applications. The first part of the process they used was feature scaling, in which they applied the minimum-maximum (min-max) normalization idea or concept on the UNSW-NB15 dataset to reduce leakage information on the test data. The given data set consists of a mix of recent attacks and typical network traffic activities, which are then classified into nine different attack categories. In the second step by using principal component analysis (PCA) the dimensionality reduction was performed. In the end 6 Machine learning models were used for the analysis. The results of this study have been evaluated in terms of data validation.

The study [9] proposed a model of intrusion detection, which uses a classification module along with two tiers and two-dimension reduction. Furthermore, U2R and R2L attacks are detected by this model. The dimensions are reduced by employing the PCA and LDA.by using the NSL-KDD dataset, the whole experiment was conducted. In the two-tier classification module, NB and the Certainty Factor version of K-NN were employed to detect suspicious activity and exhibited a solution based on the classification for cloud-based threat detection [17]. An ELM scaled in the Apache Spark cloud architecture is used to analyze the data in this article. Net flow structured data simulated [18]-[19]. The framework was proposed in [20] grounded on the IoT to determine and track COVID-19 existence. Machine learning algorithms and other techniques such as NN and K-NN are used. It was found from the experimental results that algorithm of classification provided more than 90% accuracy. In [21] using the Internet of Things and artificial intelligence, author developed a system for medical specialists in the COVID-19 pandemic. The usage of IoT was discovered to decrease the difficulties experienced by medical personnel.

A method for detecting intrusions [22] combines oversampling, outlier identification, and metric learning. In three aspects, the proposed approach improves intrusion detection. by integrating outlier detection with distance metric learning: 1) it uses a novel technique to oversample minority classes, 2) it adds a new feature based on the imbalance ratio, and 3) To make the decision border clearer, it actively minimizes outliers and rescales original samples. Furthermore, the best collection of features is extracted using a genetic algorithm. On the UNSW-NB15 dataset, the experimental findings suggest that the recommended technique can achieve 98.51 percent accuracy while maintaining a 0.82 percent false alarm rate.

An IoT attack detection solution was developed in [23] based on distributed deep learning that achieves 96 percent accuracy as a final result. Intrusion Detection Systems were proposed in [24] for IoT applications with low capacity devices. It was seen that the 99.4% for the denial of services was achieved by their final experimental outcome. In this paper authors not provided information about the dataset that will be used in the study. The investigation of [25] worked on the cybersecurity with deep learning using the NSL-KDD dataset to perform unsupervised learning of features on the trained data by self-taught deep learning approach where sparse-auto encoders were used. To sort the labeled test data into abnormal and normal categories, the learned features were used. The performance was evaluated by the methodology of n-fold cross-validation and the results are sensible.

SVM, and ELMS with K-means techniques were used in [26] to focus on denial of services outcomes of this study are 96.02% precision, 76.19% TP rate, and 5.92 untrue level and the main drawback of this study is truncated TP level and maximum untrue alarm level. The ELM technique was used in [27] and found 83% of accuracy but the main drawback of this study is that it takes a high training time. Similarly the drawback of [28] is that the proposed model training takes a long period of time although it provides 99.98% precision and 97.39% recall. Moreover, 97.7% recall, 97.7% precision, 97.7% F-measure and 83% accuracy were obtained in [20] using the Naïve Bayes but still this study has the limitations that it requires long periods of training and the dataset's feature does not represent network activity in various environments. The ANN method was applied in [29] to reveal 99.4% of accuracy but in this study the author did not provide the information about the dataset they used. Self-taught DL sparse auto encoder was studied in [30], as a result they found STL: F-measure 98.84% and SMR: F-measure 96.76%, but they used the dataset obtained in a traditional network, which is not suitable for IoT protocols.

The First, privacy-enhancing edge intelligence model was provided in [31] using a federated machine learning mechanism is defined in this research. Differential privacy and Paillier homomorphic encryption go beyond 5G networks. Second, an Intrusion Detection System for Artificial Immune has been developed to monitor and identify nodes in the edge network that are causing an abnormality, allowing the network to form a result, a seamless and secure data transmission is provided as required. Security concerns, irregularity, and service failure are all significant challenges for this system. As a result, there is a need for an effective system that can address these problems [31]. This article investigates these issues and proposes a paradigm for improved communication, specifically the Energy Aware Smart Home (EASH) architecture. EASH analyzes the problem of communication failures and types of network attacks with this effort. The anomaly causes of the communication paradigm are distinguished using the machine learning technique. To assess the performance, we examine the suggested work for accuracy, efficiency, and performance. As a result, we get superior results, particularly the 85 percent accuracy rate. In the future, we will strive to improve our high accuracy rate [32].

Table 1 Existing methods for IOT attacks classification using different ML strategies

Reference	Method/Technique name	Outcomes	Drawbacks
[26]	SVMs and ELMS with K-means	96.02% precision, 76.19% TP rate, and 5.92 Untrue level	Truncated TP level and maximum untrue alarm level
[33]	DNN and shallow NN models	For probe attack Shallow NN = 96.75% Precision, DNN = 98.27% Precision	The NSLKDD A dataset was used that did not reflect the current attacks.
[27]	ELM	83% Accuracy	High training times
[28]	Decision tree	99.98, Precision 97.39 Recall	Model training takes a long period.
[20]	NB	97.7% Recall 97.7% Precision 97.7% F-measure 83% accuracy	Long periods of training The dataset's features do not represent network activity in various environments. s
[15]	Self-organized ant colony networks	DoS attack and accuracy = 98.55 Accuracy = 99.79	This dataset does not reflect present day attack
[34]	LDA for dimensionality reduction with NB and CF- KNN for classification of network traffic	Accuracy = 84.82% and false alarm rate = 5.56	Low detection rate and high FP rate
[29]	ANN	Accuracy = 99.4%	No information on the dataset used
[30]	Self-taught DL sparse auto encoder	STL: F-measure = 98.84% SMR: F-measure = 96.76%	The dataset obtained in traditional network and not suitable for IoT protocols

The usage of distributed FBG for invasion monitoring was expanded upon to establish the location of an intruder. They employed empirical wavelet packet and characteristic entropy techniques for mode decomposition to deconstruct the signals from several FBGs, by detecting in-ground and fence detection. The method was equitably extensive, and it worked well for interpreting vibrational signals from various FBGs and estimating the location of an intruder. LabVIEW was used to create a simple graphical user interface (GUI), allowing for real-time monitoring of the perimeter It could not, however, determine false alarms and needed to be improved [35]. A fiber brag grating sensor (FBG) perimeter intrusion detection sensor based on an armored cable was presented by [36]-[37]. The above-mentioned techniques and algorithms for IoT attack classification using different ML strategies are presented Table 1.

3. Proposed Methodology

This section describes the research framework from the process of clustering by using the proposed algorithm and classification. As for the classification, we used K-NN, SVM, Random Forest, and naïve Bayes. To do the evaluation, the results were compared with relatable studies.

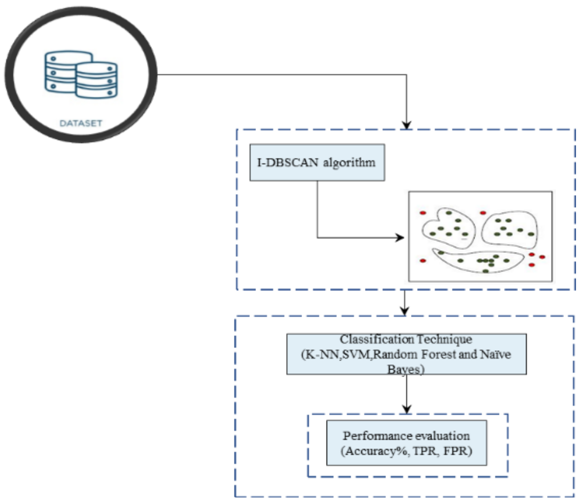
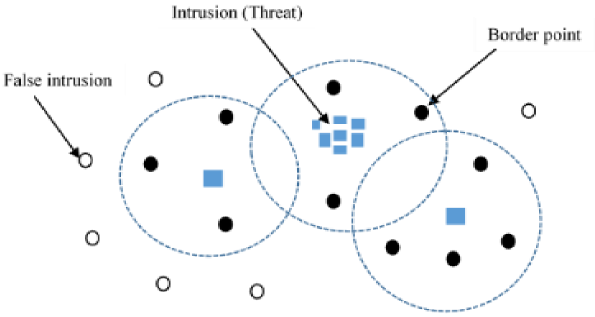


Fig. 2 Architecture of the proposed method

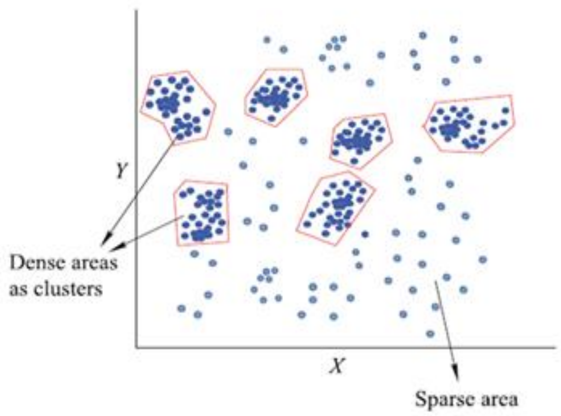
In Figure 2, the architecture of our method has been described, initially we should load the dataset and after that we applied I-DBSCAN to generate the clusters more efficiently once the clusters are generated, we then applied the classifier techniques K-NN, SVM, RF, and Naïve Bayes to get accuracy %, TPR, FPR accordingly.

3.1. Proposed DBSCAN Algorithm for Clustering

The proposed density-based spatial clustering of applications with noise (DBSCAN) method is used to form clusters, that are dense and similar types of intrusion. DBSCAN is a density-based clustering technique. It can find clusters of various forms and sizes in a vast amount of data that is noisy and contains outliers. Figure 3(a). shows an improved DBSCAN (I-DBSCAN) that can cluster similar data points in neighboring means if the data points are closer that are considered the same type of intrusion, and it will recognize as of core points including the few more data points, which form a single cluster, and false intrusion is shown, which is not forming any cluster because it's false positive where we don't want to have an alarm border point is considered an intrusion if it is coming under the region of the cluster. I-DBSCAN algorithm computation is further explained in Algorithm 1.



(a)



(b)

Fig. 3 Clustering of dense areas by density-based clustering algorithm (a) clustering the dense areas in circular form (b) clustering the dense areas regardless of the shape of clusters

A density-based clustering algorithm presents the dense areas as clusters, as shown in Figure 3(b). The density-based clustering algorithm also enables us to detect clusters in areas of uniform density regardless of the size of the region.

Algorithm 1. I-DBSCAN

1 Variables *minpts, eps and p*
2 Initialize *minpts, eps*
3 Initialize *p* at random
4 Calculate *eps* against *p* using equation (1)

$$E(p,q) = \sqrt{\sum_{i=0}^{n} (p_i - q_i)^2}$$

5 *If (p > minpts)*
Then P is core point
And cluster is generated
End if
6 *If (! (p==visited))*
Then, go to step 3
End if
7 *End*

Step-by-step explanation of the algorithm 1 is given below:

Step 1: Declare the two variables which are required for the I-DBSCAN

Step 2: In this step we are initializing both variables declared above

Step 3: In this step we are initializing the p variable with random value

Step 4: In this step we are calculating the epsilon (eps) against p by using equation (1)

Step 5: At this stage, we are checking that if the p point is greater than the minpts then cluster is generated

Step 6: At this step we are finding that if all points are not visited, then go to step 3

The flow of I-DBSCAN algorithm is further explained in Figure 4. where the process starts from initializing the variables minimum points (minpts), epsilon (eps) and point p at a random value, then it goes with the calculation of eps against point p by using equation (1) finally if the point p is greater than the minpts then a cluster is generated and after that, we checked that if the points are visited, it will go to the end, else it will be redirect to step 3.

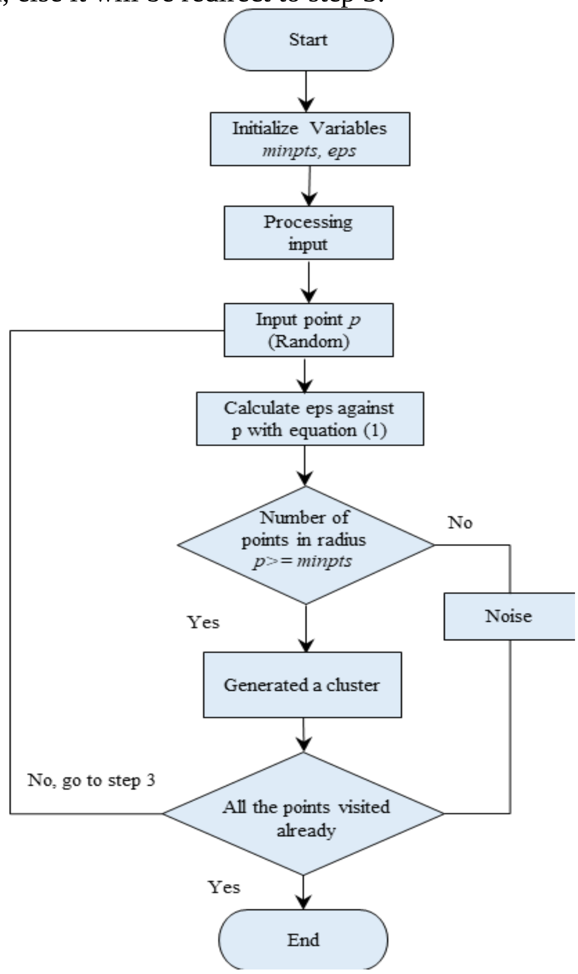


Fig. 4 I-DBSCAN algorithm flow

3.2. Classification

Classification is one part of the data mining process.

If the cluster algorithm process has no label or target class. But the classification can be considered supervised learning. In this research, we use the Naive Bayes Random Forest, SVM and K-Nearest Neighbor classification algorithm

Naïve Bayes is a straightforward probabilistic classification approach that computes a set of probabilities based on the total of a dataset's frequencies and value combinations. The categorization procedure with this method only requires a small quantity of data, yet it frequently produces unexpected findings that don't match the facts.

Random Forest is a classification technique that generates the most decision tree-generated classes, using several decision trees as classifiers and boosting accuracy through voting on the available decision trees,

The Support Vector Machine (SVM) is a technique that can be applied to both regression and classification. SVM performs best with data that have several dimensions. However, SVM training time is often slow, SVM is particularly accurate at handling complex nonlinear models. Unlike other approaches, SVM's shortcoming makes it susceptible to overfitting.

K-NN classifies objects by using raster learning data that is closest to the object. This technique seeks to categorize new objects based on characteristics and training data. This method is incredibly straightforward and simple to use like the clustering method, grouping a new set of data is dependent on how far away its neighbors are.

4. Experimental Results and Discussions

We discovered 362 clusters in the training phase with 22 different epsilon values and 23 different minpts values. 3 large and 359 minor size clusters were discovered. The results of the detection phase for the final three data sets include the detection rate, false positive rate, number of clusters generated, and number of updated cluster sizes. I-DBCSAN detected 1,772 attacks in the second section, added 3 new normal clusters to 87 new clusters, adjusted the size of 70 clusters, and left 74 uncertain spots. I-DBSCAM detected 2,736 attacks in the third section, generated 80 new clusters with 3 more normal clusters, adjusted the size of 72 clusters, and left 161 uncertain points. I-DBSCAN detected 2,135 assaults in the fourth, generated 86 new clusters with 3 more normal clusters, revised the size of 78 clusters and left 75 uncertain points. We compared the results of I-DBSCAN with those of the original DBSCAN by setting different epsilon values, which we found from the training phase from all clusters. The outcome shows that the highest detection rate of the original DBSCAN is lower and false positive rate is higher than the I-DBSCAN as depicted in Table 2. Furthermore, we can differentiate the performance of both the algorithm in Figures 5 and 6.

Table 2 Comparison of DBSCAN with I-DBSCAN

DBSCAN applied to KDD CUP 99			**I-DBSCAN applied to KDD CUP 99**		
Epsilon	**Detection Rate**	**False Rate**	**Epsilon**	**Detection Rate**	**False Rate**
0.4	0.833	0.558	0.4	0.955	0.458
0.8	0.933	0.623	0.8	0.945	0.523
1.2	0.946	0.429	1.2	0.948	0.399
1.6	0.961	0.362	1.6	0.964	0.262
2	0.646	0.316	2	0.8	0.216
2.4	0.759	0.289	2.4	0.857	0.189
2.8	0.588	0.278	2.8	0.688	0.178
3.2	0.547	0.265	3.2	0.648	0.165
3.6	0.663	0.239	3.6	0.732	0.139
4	0.641	0.223	4	0.721	0.123
4.4	0.616	0.207	4.4	0.716	0.107
5.6	0.54	0.197	5.6	0.645	0.097
6	0.534	0.177	6	0.638	0.077
6.8	0.384	0.132	6.8	0.449	0.032
7.2	0.372	0.123	7.2	0.447	0.023
7.6	0.358	0.128	7.6	0.442	0.028
8.4	0.338	0.112	8.4	0.429	0.012
8.8	0.319	0.104	8.8	0.407	0.004
9.6	0.331	0.093	9.6	0.436	0.0093
10.4	0.306	0.082	10.4	0.419	0.0082
10.2	0.299	0.073	10.2	0.398	0.0073
12	0.294	0.068	12	0.395	0.0068

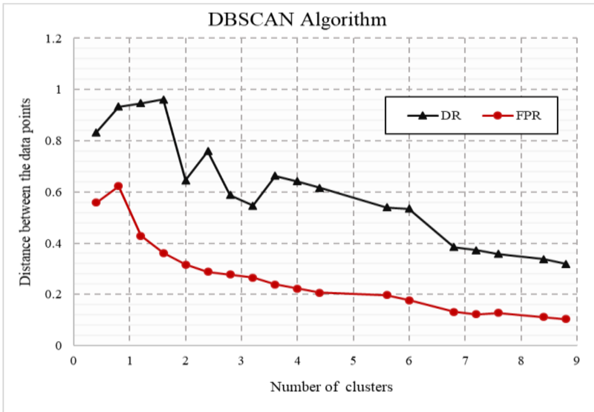


Fig. 5 Performance of DBSCAN on different epsilon values and clusters

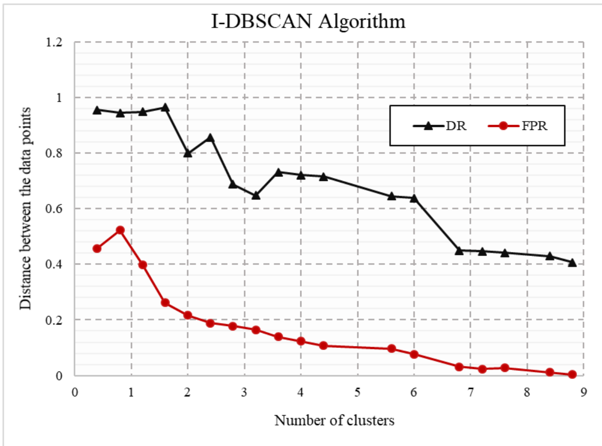


Fig. 6 Performance of I-DBSCAN on different epsilon values and clusters

In this research, using the WEKA tool, the feature selection procedure was carried out [38]. Its accuracy, True Positive Rate (TPR), and False Positive Rate (FPR) are used to assess performance (see 2, 3, 4, respectively). The overall number of attacks that go undetected is known as the false negative (FN). The total number of normal conditions that were identified as normal is known as True Negative (TN). False positives (FP) are any normal condition mistakenly identified as attack conditions. The number of attacks that were identified as an attack condition is known as True Positive (TP) [39]. The ratio of precision to trueness comes next. TPR is the ratio of attacks that were detected in all attacks combined. FPR is the ratio of false attacks or normal activity incorrectly detected in all data.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (2)$$

$$\text{Detection Rate} = (TP)/(TP+FP) \quad (3)$$

$$\text{False Alarm} = (FP)/(FP+TN) \quad (4)$$

The KDD CUP 99 dataset, with 6297 rows of data and cross validation 10, is used in this study. The KDD CUP 99 dataset, with 6297 rows of data and cross validation 10 is used in this study. Table 3 shows the performance evaluation of the initial experiment on the KDD Cup 99 dataset. The best Random Forest classification was determined to have a 99.954% accuracy and a TPR of 1 with an FPR of 0.

Table 3 Results of KDD Cup 99 classification

Classifier	Accuracy %	TPR	FPR
Random Forest	99.954	1	0
SVM	99.87	0.91	0
K-NN	99.92	0.999	0
Naïve Bayes	92.0223	0.920	0

Table 4 Results of KDD CUP 99 classification with the proposed method

Classifier	Accuracy %	TPR	FPR
Random Forest	99.9853	1	0
SVM	99.9136	0.995	0
K-NN	99.9862	1	0
Naïve Bayes	98.5334	0.981	0

In the second KDD Cup 99 experiment, by selecting the attributes sequentially, a new dataset is created. An evaluation of the experiment's performance is shown in Table 4. Table 4 demonstrates that the suggested approach can improve the accuracy for all the classifiers used in this research SVM, K-NN, Naïve Bayes and Random Forest. In the second experiment, the K-NN had better performance compared to the other three classifications with an accuracy of 99.9862% and TPR of 1 and FPR of 0. The drastic performance increase occurred in the SVM classification, which originally has the accuracy and TPR respectively 99.87% and 0.91 rise to 99.9136% and 0.995 respectively in the second experiment.

The NSL-KDD data set is then used. It provides a solution for the issues with the KDD Cup 1999 dataset (KDD- 99). KDD-99 Cup has been around for more than 17 years. However, it still often used in IDS research because there aren't many readily available, publicly accessible datasets. The 39 different attack types and other normal classes are available in this data collection.

The KDD CUP 99 dataset containing 74094 rows of data was used in this experiment, and cross validation was set at 10. The experiment was repeated twice, as in the first instance. The accuracy level for the first trial, which generated the performance evaluation in Table 5, showed that the Random Forest classification performed best, with a TPR of 0.996 and an accuracy level of 99.603 percent.

Table 5 Results of NSL-KDD CUP 99 classification

Classifier	Accuracy %	TPR	FPR
Random Forest	99.603	0.996	0.004
SVM	98.2849	0.973	0.029
K-NN	98.9042	0.989	0.00
Naïve Bayes	90.694	0.907	0.092

Table 6 Results of classification of NSL-KDD CUP 99 with the proposed method

Classifier	Accuracy %	TPR	FPR
Random Forest	99.8933	0.899	0.00
SVM	97.9405	0.970	0.002
K-NN	99.8982	0.999	0.00
Naïve Bayes	96.8833	0.998	0.00

In the second experiment, a fresh data set was produced using the NSL-KDD Cup 99 data with the suggested methodology. The performance assessment for this experiment is shown in Table 6. Due to the volume of data and choice of characteristics, the classification procedure in this experiment took a little longer to complete.

Table 6 demonstrates that, except SVM, the accuracy of the Nave Bayes, Random Forest, and k-NN models has increased in the second trial using the NSLKDD Cup 99 data set.

K-NN progressed better than the other three categories in the second experiment, with an accuracy of 99.8982 percent and TPR of 0.999. The accuracy and TPR of the Naive Bayes classification significantly increased accuracy and TPR from 90.694 percent and 0.907 in the first experiment to 96.883 percent and 0.998 respectively in the second experiment.

Table 7 provides a comparison of the performance while employing the suggested strategy. KDD Cup 99 shows the best accuracy then NSL-KDD cup, as shown in Table 7.

Table 7 Accuracy performance comparison of the proposed method

Classifier	KDD Cup 99(%)	NSL (%)
Random Forest	99.9853	99.8933
SVM	99.9136	97.9405
K-NN	99.9862	99.8982
Naïve Bayes	98.5334	96.8833

At the end, we present the accuracy results from our proposed method and compare them to the results obtained by [40]. It is clear from Table 8 that our strategy performs well on both Datasets. By using our proposed method on the KDD CUP 99 the SVM performs more efficient with an accuracy of 99.9136 whereas K-NN classifier provide the accuracy of 99.9862 similarly Naïve Bayes produced an accuracy of 98.5334 and the Random Forest provided an accuracy of 99.9853.

Similarly to the above discussion, when we applied our proposed method to NSL-KDD Cup 99 the SVM produced an accuracy of 97.9405, K-NN provided an accuracy of 99.8982, likewise Naïve Bayes is at the accuracy of 96.8833 and random forest provide the accuracy of 99.8933.

Table 8 Performance evaluation with proposed method compression

Classifier	KDD Cup 99 (%)	NSL (%)	Classifier	KDD Cup 99(%)	NSL (%)
SVM	99.9136	97.9405	SVM	99.5218	97.0405
K-NN	99.9862	99.8982	K-NN	99.9851	99.7982
Naïve Bayes	98.5334	96.8833	Naïve Bayes	98.1334	96.7883
Random Forest	99.9853	99.8933	Random Forest	99.981	99.8823

5. Conclusion

In this study, we developed an improved DBSCAN algorithm called I-DBSCAN that can be used for more effective clustering is the main strength of this study. For this purpose, we used the clustering and classification techniques. Additionally, this research can spot attacks in data from intrusion detection systems. The public will identify attack patterns and signatures with high accuracy and learn how to defend against them. The usage of various datasets and the idea of deep learning can both be tested further researcher can use the I-DBSCAN on different other datasets in the future. From the overall results obtained, the combination of I-DBSCAN with classification methods of Random Forest, SVM, K-NN and Naïve Bayes in the KDD Cup 99 and NSL-KDD Cup 99 datasets improves the accuracy. The effectiveness of this work is to determine the accuracy, TPR, and FPR of classification based on intrusion detection system (IDS) data will rise because of the usage of I-DBSCAN in data preparation. Additionally, this study contrasts four classification techniques. Results of comparing the four classifications show that K-NN tends to perform better in both the experiments, whereas the dominating Random Forest (RF) approach performs worse when using the suggested method. The SVM method with the proposed strategy has 99.9136% percent accuracy, where the improvement is found. Moreover, it is not performing well with NSL-KDD Cup 99 and proposed work is not focused on cost effectiveness. Similarly, Naïve Bayes accuracy found 98.5334 with the proposed method on KDD cup 99 whereas with NSL-KDD Cup 99, the result of its accuracy comes to 98.133%. Results of our study proved to be better in terms of accuracy when they are compared to the already available work of Khadija et al. previous work.

Acknowledgments

This research was conducted at Universiti Teknologi PETRONAS (UTP) under the YUTP grant scheme with the cost center of 015LC0-350.

References

[1] WANG B, YAO X, JIANG Y, SUN C, and SHABAZ M. Design of a real-time monitoring system for smoke and dust in thermal power plants based on improved genetic algorithm. [J] *Journal of Healthcare Engineering*, 2021, 2021, article id 7212567: 1-10, https://doi.org/10.1155/2021/7212567

[2] SEVILLA F R S, *et al.* State-of-the-art of data collection, analytics, and future needs of transmission utilities worldwide to account for the continuous growth of sensing data. [J] *International Journal of Electrical Power Energy Systems*, 2022, 137, article id 107772.

[3] SHARIF Z, JUNG L T, and AYAZ M. Priority-based Resource Allocation Scheme for Mobile Edge Computing. [C] *Proceeding* of the *2nd International Conference on Computing and Information Technology (ICCIT)*, 2022: 138-143.

[4] LULLA G, KUMAR A, POLE G, and DESHMUKH G. IoT based Smart Security and Surveillance System. [C] *Proceeding* of the *International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2021: 385-390.

[5] SICARD F, ZAMAÏ É, FLAUS J-M, and SAFETY S. An approach based on behavioral models and critical states distance notion for improving cybersecurity of industrial control systems. [J] *Reliability Engineering*, 2019, 188: 584-603.

[6] JAHROMI A N, KARIMIPOUR H, DEHGHANTANHA A, and CHOO K-K R. Toward Detection and Attribution of Cyber-Attacks in IoT-Enabled Cyber–Physical Systems. [J] *IEEE Internet of Things Journal*, 2021, 8(17): 13712-13722.

[7] SHARIF Z, JUNG L T, AYAZ M, YAHYA M, and KHAN D. Smart Home Automation by Internet-of-Things Edge Computing Platform. *International Journal of Advanced Computer Science Applications*, 2022, 13(4): 474-484.

[8] LIN K-S, YEH K-H, CHIANG Y-J, and WANG L. Fiber-optic perimeter intrusion detection by employing a fiber laser cavity in each defensed zone. [J] *IEEE Sensors Journal*, 2018, 18(20): 8352-8360.

[9] AJOUH H H, JAVIDAN R, KHAYAMI R, *et al.* A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks. [J] *IEEE Transactions on Emerging Topics in Computing*, 2016, 7(2): 314-323.

[10] SHARMA J, GIRI C, GRANMO O-C, and GOODWIN M. Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and

softmax aggregation. [J] *EURASIP Journal on Information Security*, 2019, (1): 1-16.

[11] SHARIF Z, *et al.* Internet-of-Things based Home Automation System using Smart Phone. [J] *Sir Syed University Research Journal of Engineering Technology*, 2021, 11(2): 70-76.

[12] JAN S U, AHMED S, SHAKHOV V, and KOO I. Toward a lightweight intrusion detection system for the internet of things. [J] *IEEE Access*, 2019, 7: 42450-42471.

[13] AYAZ M, AMMAD-UDDIN M, SHARIF Z, *et al.* Internet-of-Things (IoT)-based smart agriculture: Toward making the fields talk. [J] *IEEE Access*, 2019, 7: 129551-129583.

[14] SHARIF Z, JUNG L T, RAZZAK I, and ALAZAB M. Adaptive and Priority-based Resource Allocation for Efficient Resources Utilization in Mobile Edge Computing. [J] *IEEE Internet of Things Journal*, 2021: 1-15, https://doi.org/10.1109/JIOT.2021.3111838.

[15] CHEN P, YOU C, and DING P. Event classification using improved salp swarm algorithm based probabilistic neural network in fiber-optic perimeter intrusion detection system. [J] *Optical Fiber Technology*, 2020, 56, article id 102182, https://doi.org/10.1016/j.yofte.2020.102182

[16] SAHEED Y K, ABIODUN A I, MISRA S, *et al.* A machine learning-based intrusion detection for detecting internet of things network attacks. [J] *Alexandria Engineering Journal*, 2022, 61(12): 9395-9409.

[17] KOZIK R, CHORAŚ M, FICCO M, *et al.* A scalable distributed machine learning approach for attack detection in edge computing environments. *Journal of Parallel*, 2018, 119: 18-26.

[18] TSIKALA VAFEA M, *et al.* Emerging technologies for use in the study, diagnosis, and treatment of patients with COVID-19. *Cellular*, 2020, 13(4): 249-257.

[19] UDDIN M, AYAZ M, MANSOUR A, *et al.* Cloud-connected flying edge computing for smart agriculture. [J] *Peer-to-Peer Networking Applications*, 2021, 14(6): 3405-3415.

[20] OTOOM M, OTOUM N, ALZUBAIDI M A, *et al.* An IoT-based framework for early identification and monitoring of COVID-19 cases. [J] *Biomedical Signal Processing*, 2020, 62, article id 102149.

[21] KUMAR S, RAUT R D, and NARKHEDE B. A proposed collaborative framework by using artificial intelligence-internet of things (AI-IoT) in COVID-19 pandemic situation for healthcare workers. [J] *International Journal of Healthcare Management*, 2020, 13(4): 337-345.

[22] JIN F, CHEN M, ZHANG W, *et al.* Intrusion detection on internet of vehicles via combining log-ratio oversampling, outlier detection and metric learning. [J] *Information Sciences*, 2021, 579: 814-831.

[23] DIRO A A, and CHILAMKURTI N. Distributed attack detection scheme using deep learning approach for Internet of Things. [J] *Future Generation Computer Systems*, 2018, 82: 761-768.

[24] E. HODO *et al.* Threat analysis of IoT networks using artificial neural network intrusion detection system. [C] *Proceedings of the International Symposium on Networks, Computers and Communications (ISNCC)*, 2016: 1-6.

[25] JAVAID A, NIYAZ Q, SUN W, *et al.* A deep learning approach for network intrusion detection system. *Eai Endorsed Transactions on Security*, 2016, 3(9): e2.

[26] RGHIOUI A, KHANNOUS A, and BOUHORMA M. Denial-of-Service attacks on 6LoWPAN-RPL networks: Threats and an intrusion detection system proposition. [J] *Journal of Advanced Computer Science*, 2014, 3(2): 143-153.

[27] XIANG C, CHONG M, and ZHU H. Design of mnitiple-level tree classifiers for intrusion detection system. [C] *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems*, 2004, 2: 873-878.

[28] WU Y, KE Y, CHEN Z, *et al.* Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping. [J] *Catena*, 2020, 187, 104396.

[29] RAJASEGARAR S, LECKIE C, and PALANISWAMI M. Anomaly detection in wireless sensor networks. [J] *IEEE Wireless Communications*, 2008, 15(4): 34-40.

[30] RAZA S, WALLGREN L, and VOIGT T. SVELTE: Real-time intrusion detection in the Internet of Things. [J] *Ad Hoc Networks*, 2013, 11(8): 2661-2674.

[31] KUMAR K S, NAIR S A H, ROY D G, *et al.* Security and privacy-aware artificial intrusion detection system using federated machine learning. [J] *Computers*, 2021, 96, 107440.

[32] ATUL D J, *et al.* A machine learning based IoT for providing an intrusion detection system for security. [J] *Microprocessors*, 2021, 82, 103741.

[33] LAN T, ZHANG C, LI L, *et al.* Perimeter security system based on fiber optic disturbance sensor. [C] *Advanced Sensor Systems and Applications*, SPIE, 2007, 6830: 107-112, https://doi.org/10.1117/12.756541.

[34] XIANG C, YONG P C, and MENG L S Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees. [J] *Pattern Recognition Letters*, 2008, 29(7): 918-924.

[35] HAN X, XU D, and LIU Y. Application of principal components analysis in condenser fault diagnosis. [C] *Proceedings of the 6th World Congress on Intelligent Control and Automation*, 2006, 2: 5666-5669.

[36] ALLWOOD G, WILD G, and HINCKLEY S. Optical fiber sensors in physical intrusion detection systems: A review. [J] *IEEE Sensors Journal*, 2016, 16(14): 5497-5509.

[37] IQBAL S, HUSSAIN I, SHARIF Z, *et al.* Reliable and energy-efficient routing scheme for underwater wireless sensor networks (UWSNs). [J] *International Journal of Cloud Applications Computing*, 2021, 11(4): 42-58.

[38] HALL M, FRANK E, HOLMES G, *et al.* The WEKA data mining software: an update. [J] *ACM SIGKDD Explorations Newsletter*, 2009, 11(1): 10-18.

[39] AHMAD T, and MUCHAMMAD K. L-SCANN: Logarithmic subcentroid and nearest neighbor. [J] *Journal of Telecommunications Information* Technology, 2016, 4: 71-80.

[40] KHADIJA M A, WIDYAWAN S, and NUGROHO I L E. Detecting network intrusion by combining DBSCAN, principle component analysis and ranker. [C] *Proceedings of the International Seminar on Research of Information Technology and Intelligent Systems*, 2019: 165-170.

参考文:

[1] WANG B, YAO X, JIANG Y, SUN C, 和 SHABAZ M. 基于改进遗传算法的火电厂烟尘实时监测系统设计。[J]

医疗工程杂志，2021, 2021, 文章编号 7212567: 1-10, https://doi.org/10.1155/2021/7212567

[2] SEVILLA F R S，等。最先进的数据收集、分析和全球传输公用事业的未来需求，以应对传感数据的持续增长。[J] 国际电力能源系统杂志，2022, 137, 文章编号 107772.

[3] SHARIF Z、JUNG L T 和 AYAZ M。移动边缘计算的基于优先级的资源分配方案。[C] 第二届国际计算与信息技术会议（国际商会）论文集，2022：138-143。

[4] LULLA G、KUMAR A、POLE G 和 DESHMUKH G. 基于物联网的智能安全和监控系统。[C] 新兴智能计算和信息学(ESCI)国际会议论文集，2021：385-390。

[5] SICARD F、ZAMAÏ É、FLAUS J-M 和 SAFETY S。一种基于行为模型和临界状态距离概念的方法，用于提高工业控制系统的网络安全。[J] 可靠性工程, 2019, 188: 584-603.

[6] JAHROMI A N、KARIMIPOUR H、DEHGHANTANHA A 和 CHOO K-K R。在支持物联网的网络物理系统中检测和归因网络攻击。[J] IEEE 物联网杂志, 2021, 8(17): 13712-13722.

[7] SHARIF Z、JUNG L T、AYAZ M、YAHYA M 和 KHAN D. 通过物联网边缘计算平台实现智能家居自动化。国际高级计算机科学应用杂志, 2022, 13(4): 474-484.

[8] LIN K-S、YEH K-H、CHIANG Y-J 和 WANG L. 在每个防御区域使用光纤激光腔进行光纤周界入侵检测。[J] IEEE 传感器杂志, 2018, 18(20): 8352-8360.

[9] AJOUH H H、JAVIDAN R、KHAYAMI R 等。物联网骨干网络中基于异常的入侵检测的两层降维和两层分类模型。[J] IEEE 计算新兴主题汇刊, 2016, 7(2): 314-323.

[10] SHARMA J、GIRI C、GRANMO O-C 和 GOODWIN M。具有额外的树特征选择、极限学习机集成和软最大聚合的多层入侵检测系统。[J] 欧亚知识产权信息安全期刊, 2019, (1): 1-16.

[11] SHARIF Z 等。使用智能手机的基于物联网的家庭自动化系统。[J] 赛义德爵士大学工程技术研究杂志, 2021, 11(2): 70-76.

[12] JAN S U、AHMED S、SHAKHOV V 和 KOO I。面向物联网的轻量级入侵检测系统。[J] IEEE 访问, 2019, 7: 42450-42471.

[13] AYAZ M、AMMAD-UDDIN M、SHARIF Z 等。基于物联网(物联网)的智能农业：让田野对话。[J] IEEE 访问, 2019, 7: 129551-129583.

[14] SHARIF Z、JUNG L T、RAZZAK I 和 ALAZAB M。移动边缘计算中高效资源利用的自适应和基于优先级的资源分配。[J] IEEE 物联网杂志，2021：1-15，https://doi.org/10.1109/JIOT.2021.3111838。

[15] CHEN P、YOU C 和 DING P. 在光纤周界入侵检测系统中使用改进的基于概率神经网络的樽海鞘群算法进行事件分类。[J] 光纤技术, 2020, 56, 文章编号 102182, https://doi.org/10.1016/j.yofte.2020.102182

[16] SAHEED Y K、ABIODUN AI、MISRA S 等。基于机器学习的入侵检测，用于检测物联网网络攻击。[J] 亚历山大工程学报, 2022, 61(12): 9395-9409.

[17] KOZIK R、CHORAŞ M、FICCO M 等。一种可扩展的分布式机器学习方法，用于边缘计算环境中的攻击检测。并行学报, 2018, 119: 18-26.

[18] TSIKALA VAFEA M 等。用于研究、诊断和治疗新冠肺炎患者的新兴技术。蜂窝, 2020, 13(4): 249-257.

[19] UDDIN M、AYAZ M、MANSOUR A 等。用于智能农业的云连接飞行边缘计算。[J] 点对点网络应用, 2021, 14(6): 3405-3415.

[20] OTOOM M、OTOUM N、ALZUBAIDI M A 等。用于早期识别和监测新冠肺炎病例的基于物联网的框架。[J] 生物医学信号处理, 2020, 62, article id 102149.

[21] KUMAR S、RAUT R D 和 NARKHEDE B。在新冠肺炎大流行情况下针对医护人员使用人工智能物联网(物联网)提出的协作框架。[J] 国际医疗管理杂志，2020, 13(4): 337-345.

[22] 金峰, 陈敏, 张伟, 等．通过结合对数比过采样、异常值检测和度量学习对车联网进行入侵检测。[J] 信息科学, 2021, 579: 814-831.

[23] DIRO AA 和 CHILAMKURTI N. 使用深度学习方法的物联网分布式攻击检测方案。[J] 下一代计算机系统, 2018, 82: 761-768.

[24] E. HODO 等。使用人工神经网络入侵检测系统对物联网网络进行威胁分析。[C] 网络、计算机和通信国际研讨会 (ISNCC) 论文集，2016：1-6。

[25] JAVAID A、NIYAZ Q、SUN W 等。一种用于网络入侵检测系统的深度学习方法。EAI 认可的安全交易，2016，3(9)：电子 2。

[26] RGHIOUI A、KHANNOUS A 和 BOUHORMA M. 6LoWPAN-RPL 网络上的拒绝服务攻击：威胁和入侵检测系统命题。[J] 计算机科学学报, 2014, 3(2): 143-153.

[27] XIANG C、CHONG M 和 ZHU H. 入侵检测系统的多层树分类器设计。[C] IEEE 控制论和智能系统会议论文集，2004，2：873-878。

[28] WU Y, KE Y, CHEN Z, 等。交替决策树与 AdaBoost 和套袋集成在滑坡敏感性映射中的应用。[J] 连锁, 2020, 187, 104396.

[29] RAJASEGARAR S、LECKIE C 和 PALANISWAMI M. 无线传感器网络中的异常检测。[J] IEEE 无线通信, 2008, 15(4): 34-40.

[30] RAZA S、WALLGREN L 和 VOIGT T. SVELTE：物联网中的实时入侵检测。[J] 自组织网络, 2013, 11(8): 2661-2674.

[31] KUMAR K S、NAIR S A H、ROY D G 等。使用联合机器学习的安全和隐私感知人工入侵检测系统。[J] 计算机, 2021, 96, 107440.

[32] ATUL D J，等。基于机器学习的物联网，用于提供安全入侵检测系统。[J] 微处理器, 2021, 82, 103741.

[33] LAN T，ZHANG C，LI L，等。基于光纤干扰传感器的周界安全系统。[C] 高级传感器系统和应用，SPIE，2007，6830：107-112，https://doi.org/10.1117/12.756541。

[34] XIANG C、YONG P C 和 MENG L S 使用贝叶斯聚类和决策树设计入侵检测系统的多级混合分类器。[J] 模式识别快报, 2008, 29(7): 918-924.

[35] HAN X, XU D, 和 LIU Y. 主成分分析在冷凝器故障诊断中的应用。[C] 第六届世界智能控制与自动化大会论文集, 2006, 2: 5666-5669.

[36] ALLWOOD G、WILD G 和 HINCKLEY S. 物理入侵检测系统中的光纤传感器：综述。[J] IEEE 传感器杂志, 2016, 16(14): 5497-5509.

[37] IQBAL S、HUSSAIN I、SHARIF Z 等。用于水下无线传感器网络(UWSN)的可靠且节能的路由方案。[J] 国际云应用计算学报, 2021, 11(4): 42-58.

[38] HALL M、FRANK E、HOLMES G 等人。维卡数据挖掘软件：更新。[J] 美国计算机学会 SIGKDD 探索通讯, 2009, 11(1): 10-18.

[39] AHMAD T 和 MUCHAMMAD K. 扫描仪：对数子质心和最近邻。[J] 电信信息技术学报, 2016, 4: 71-80.

[40] KHADIJA MA、WIDYAWAN S 和 NUGROHO I L E. 通过结合数据库扫描、主成分分析和排序器检测网络入侵。[C] 信息技术与智能系统研究国际研讨会论文集, 2019: 165-170.