INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Semi-supervised Clustering Based Feature Selection with Multiobjective Genomic Search Class-based Classification Method for NIDPS

P Poobalan[1]*, **S Pannirselvam**[2]

1 Ph.D., Research Scholar, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, 638009, India
2 Associate Professor (Rtd), Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, 638009, India

OPEN ACCESS

Received: 04-02-2022

Accepted: 13-04-2022

Published: 27-05-2022

Citation: Poobalan P, Pannirselvam S (2022) Semi-supervised Clustering Based Feature Selection with Multiobjective Genomic Search Class-based Classification Method for NIDPS. Indian Journal of Science and Technology 15(19): 948-955. https://doi.org/10.17485/IJST/v15i19.297

* **Corresponding author**.

poobalanmca@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The purpose of semi-supervised clustering-based feature selection with the multiobjective genomic search class based classification process is to extract the intrusion features from the content of the unbalanced class field and structures in the dataset. **Method:** A class-based taxonomy with Multiobjective Genomic Search Method (SCMGSM) is semi-supervised Clustering Based Feature Selection designed to avoid inappropriate classification and to generate class based efficient authenticity for detecting intrusions. The proposed SCMGSM operates on a cluster class basis to reduce the overhead distribution by utilizing the subset classifications based on labelling process. For cluster class based feature selection there is not necessary to retain the interconnected features, so the proposed taxonomy to detect the attack features for each subset classification improves the performance. SCMGSM protects the data source using the potential blocking and minimizes the hassle of unauthorized users from changing features. **Findings:** The experiments focused on a set of data to identify and classify features based on each class. The proposed SCMGSM handles enhanced classification accuracy with 5.37% normal class completeness, reduces false positive rate to 2.78%, improves detection rate to 6.75% and reduces the backup issue by 10.65% compared to overhead classification. **Novelty:** SCMGSM operates based on multi objective genomic search based cluster class taxonomy and offers high degree of accuracy, classification and low false-positive rating compared to the detection of exploitation, Fuzzers, Generic and Renaissances attack classes with in a specific subset or a group of features.

Keywords: Classification; Cluster; Feature selection; Intrusion detection; Network attacks

1 Introduction

In recent years, network technology has enhanced the essential security for communications and businesses of the network. Nevertheless, modern web-based data development structures are always subject to various attacks, causing severe losses due to various vulnerabilities that cause many cyber security issues in the system. Therefore, the internet security of the device for various threats has become very important to ensure the availability and reliability of the system. Data security is one of the most important factors in keeping a person's or a company's digital data.

Authors Ahmed Akbar Mekandara and Tohari Ahmed (2021) follow a precautionary technique that detects intrusion and determines the type of their attacks, but privacy measures are not taken into account. SCMGSM collects similar attack features, protects their privacy and prevents intrusion by encroaching on privacy by protecting data through digital media[(1)]. Intruders are local or remote, local intruders' network clients, attempting to some extent in access attempts to elevate their access levels by abusing unauthorized privileges.

At the same time, Ahmed, Z, Shahid Khan, A, Y Xiang, C, Abdullah, J, Ahmed, F (2021) introduce IDS to confirm talking to clients to remote intruders and attempting to gain illegal access with the external network device[(2)]. Ali Metiaf, Qianhong Wu and Yazan Aljeroudi (2019) cited efforts to reduce infiltration detection (ID) as detection systems such as privacy, network availability or monitoring of networks[(3)]. However, SCMGSM works on the attack features with clustering could provide better surveillance with sensor features without any additional computing work. Genetic algorithms are an important contributing factor in adjusting performance accuracy.

Chandrasekhar G., and Sahin F. have invented a new kind of system that requires regular updates of the rules, and signatures that give a false positive ratio (FPR) are unpredictable and are not capable of detecting unknown attacks. The proposed Intrusion detection (IDS) gathers and analyzes records from masses of systems and network assets for signs of intrusions[(3)].Early researchers used crowdsourcing data to detect and mesh anomalous accounts to maintain network security by detecting intruders. Nevertheless, it confirmed that 80% of users click on a malicious link before it is blocked by a blacklist. At the same time, these approaches take too much processing and need a supervised model of attention and involvement in the active recognition of information[(4)]. The flood of network traffic suffocates the target initiates as of spread of sources.

Cheng J. R.., Li M.Y., Tang X.Y, reported in their studies that the offender's details were intended to disrupt the host's services connected to the network and to create network services and resources that were not available to its intended users through infiltration. SCMGSM Eliminates the flexibility to thwart an attack by blocking one source of the attack, having to detect the attack, attack detection techniques compare the present action of the target network to a record of known signatures. On the opposite hand, these techniques make it smarter to identify new attacks[(5)]. Detection techniques for unknown attacks, intrusion credentials and feature selection are crucial[(6)]. However, the proposed hybrid method uses feature selection as a dimension reduction technique that attempts to detect low dimensional representation of the data.

By removing irrelevant features Diro A.A., and Chilamkurti N, while keeping the original feature space intact[(7)]. Geeta Kocher and Gulshan Kumar and Janarthanam S., Sukumaran S., and Shanthakumar M to perceive unidentified attacks by linking the contemporary action of the target system to a known usual profile[(8,9)]. To provide better detection for intrusion classification, extracting features from network structures help to improve intrusion detection methods[(10)]. To stop breach attacks the SCMGSM presented several countermeasures for diagnosis, protection, as well as a trackback. With all of these countermeasures, prediction methods aim to find the contrast, between the existing tasks of the location system to a data source of recognized attack signatures. However, the proposed hybrid methods are challenging to find new attacks. The diagnosis methods present to find unidentified attacks by contrasting the existing task of the location system. Maker learning-based diagnosis strategies have adhered to limits. Not enough classified information of monitored discovering strategies trigger reduced diagnosis rate, as well as weird initialization of not being watched discovering specifications brings about neighborhood ideal or even unsatisfactory diagnosis impact[(11)]. Excessive functions in discovering methods trigger "the curse of dimensionality", as well as unusual activities establish trigger unsatisfactory detection performance.

To come over the stated restrictions, this paper suggests a lite weight hybrid multi-objective class-based classification method with a semi-supervised feature selection method to categorize and apply gene expression to analysis progress, the data dimensionality the presence of noise or irrelevant features, directly using these high-dimensional data may result in reduced time efficiency and improving classification performance.A predicted class after all accessible features consuming distinct and nonnegative IDS analysis to influence the discriminative evidence for sorting with a consistent set of constraints to the labeled data[(12)] for achieving the effective goal. Class-based classification with genomic search accomplishes by combining discriminative data with term-specific feature assortment. Each session signifies whether or not an exact feature for the class with a learned indicator vector.

2 Methodology

Methodology intends to improve intrusion detection as well as cluster-based semi-supervised feature selection. Develop the ability to manage cluster-based objective functions and manage management to deal with irrelevant issues. To drive the clustering process towards a standard solution, integrate them into a single process as a standard optimization. The goal of the proposed multi-target genetic optimization is to capture the best classification for collection rate utilization and predictive error reduction.

2.1 Feature Selection

Feature selection is the process of decreasing dimensionality in contrivance wisdom. It is vital for several causes: First, the total calculation is a summary if we can moderate with the dimensionality. Second, all features that may not be useful for classifying information can be redundant and inappropriate from a classification point of view. Therefore, it is necessary to determine its subset of functionality. Semi-supervised clustering for the feature selection technique encodes several features with the cluster centers in the usage preset. Based on the attacks and services used transformation set during the operations are informative and well-divided groups. The optimum clusters acknowledged the performance of the multi-objective optimization typically related to the single-objective optimization models.

For decision-making, the semi-supervised clustering algorithm uses the cost function with a set of operational rules that will help us identify the most effective set of clustering capabilities. To collect the optimal distance produced by the cluster using equation (1)

$$\sum_{k=1}^{K} \sum_{c_i=k} \sum_{c_i=k'} \sum_{j=1}^{p} \left(x_{ij} - x_{ji'}\right)^2 \tag{1}$$

The observations xi and xi ′ come from the selected features, and x$_{ij}$ denotes the importance of the jth feature for i. To reduce the objective function, the clustering algorithm tries to deliver the individual observation to a group. The total of clusters K and the cluster i was assigned to monitor is C$_i$, whereas $1 \leq$ C$_i$ $\leq$K. Dataset of all variables is quantified as well as calculated between observations using squared Euclidean eyes as shown in Equation (1), also known as the ”Best Number of Inner Clusters”(BNIC), using (2)and (1) is written as (3):

$$d\left(x_i, x_i\right) = \sum_{j=1}^{p} \left(x_{ij} - x_{ji\,i'}\right)^2 \tag{2}$$

$$\sum_{k=1}^{K} n_k \sum_{C_i=k} \sum_{j=1}^{p} \left(\mathrm{x}_{ij} - \overline{\mathrm{x}}_{jk}\right)^2 \tag{3}$$

Where n$_k$denotes the number of interpretations in cluster k and $x\boxtimes_{jk}$ denotes the mean value of feature j of the cluster k.

The following strategies to obtain variations of the proposed method for feature selection of optimized clusters.

- Assign each sample to the first cluster randomly.
- Calculate the significance of j in cluster k, for each characteristic j and k.
- Assign a new group to each review i as shown in (4),

$$C_i = \arg\min k \sum_{j=1}^{p} \left(\mathrm{x}_{ij} - \overline{\mathrm{x}}_{kj}\right)^2 \tag{4}$$

- Continue in steps 2 and 3 until the algorithm is combined.

The above procedure helps to update the weight as it travels between clusters using varying numbers, allowing it to lift and stack. The size of the gap between the two groups is called the ”cosmos sum” mentioned in (5)

$$G_k = E\left|\log\left(W_k\right)\right| - \log\left(W_k\right) \tag{5}$$

When K = k, make W$_k$ as in BNIC of (2) shows that W$_k$ decreases as K increases, so there is no way to choose the value of K that decreases W$_k$. Under the reference distribution, the expected value of E [log (W$_k$)] is determined. Individual clusters with the same parameters were chosen as data items of interest from reference distribution.

2.2 Genomic Search Procedure

The purpose of selecting related features is to make classification simple and accurate help to connect the required system. Identity selection[13] seeks to promote the purpose of identification selection. The selected option is to find the best feature set for the configuration[14]. The main objective of this optimization process is to reduce the number of features based on the class score as well as to reduce network detection errors and improve the accuracy of the test prediction. Proponents have redesigned top-notch options to protect against different types of network intrusion (NID). Each Network Prevention Control Panel selects individuals in the current test dataset based on clustering[15].

2.3 Proposed Method design

The solution for data cluster classifications based on the class has an optimization problem of the admin panel is a system that controls the solution, namely a vector of two parts, identification, and processing. The aim is to reduce the number of unconnected identities by consuming the identification problem as a reduction problem, thus reducing the number of inaccuracies. In the next section, the recommended procedures are described in detail. The diagram of the proposed method is shown in Figure 1. Predictability is determined by finding the best solution in the search space of the objective function of the object, adjusting the speed as described in (6).

$$x(t+1) = A \times x(t) + v_{id}(t+1) \tag{6}$$

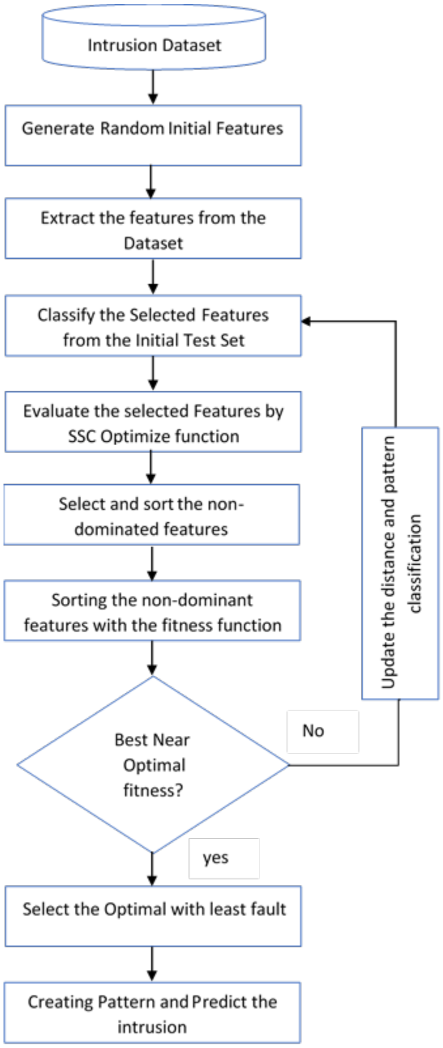
Intrusion Dataset
Generate Random Initial Features
Extract the features from the Dataset
Classify the Selected Features from the Initial Test Set
Evaluate the selected Features by SSC Optimize function
Select and sort the non-dominated features
Sorting the non-dominant features with the fitness function
Best Near Optimal fitness?
No
yes
Update the distance and pattern classification
Select the Optimal with least fault
Creating Pattern and Predict the intrusion

Fig 1. Representation of the proposed SCMGSM

The network offers the best selection in low-density areas of objective space compared to samples from cluster regions. The rapidity of the update function available in (7) is on the selected number of features of the feature as well as the self-expression in the binary search space and zero value to one.

$$v_{id}(t') = w \times v_{id}(t) + r_1 \times c_1 \times p'_{id}(t) + r_2 \times c_2 \times [REP(h) - x_{id}(t)] \tag{7}$$

The inertial weight w controls the analysis of the elements, and the number of parts, id denotes the number of parts. REP(h) is the maximum result for different species from its source and their selected index based on the value of the corresponding function, and r_1 and r_2 are the integers between zero and one (r_1,r_2∈[0,1])), and c_1 and c_2 are the acceleration constants used to control the individual's impulses completely.

3 Analysis Concerns and Discussions

Proposed Method SCMGSM uses semi-monitored cluster class-based feature selection for multi-objective gnomic search mode. The performance of the proposed hybrid system is compared with a set of measurements, compared with the current methods in the review literature, given in the following section. In this work, the UNSW_NB15 benchmark database is used with full-featured and filtered subgroups for testing and evaluation.

3.1 Performance Metrics

Metrics follow different aspects of IDS, but no single metric seems to be sufficient to evaluate this process thoroughly. The proposed system evaluates IDS capabilities to classify events accurately as normal or attack, as well as other performance objectives such as hardware architecture, vulnerability, and the ability to withstand intrusion attacks. Detection of intrusion and reduction of misclassifications of terminals in the network following criteria.

To do determine accuracy, classification rate (CR), detection rate (DR), and false-positive rate (FPR):

$$Detection\ rate\ \frac{TP}{TP+FN}$$

$$False\ Positive\ rate\ (FPR)\ =\frac{FP}{FP+TN}$$

$$Classfication\ Rate\ (CR)\ =\frac{TP+TN}{TP+\ TN+FP+FN}$$

$$Accuracy\ =\frac{TP+TN}{TP+TN+FP+FN}$$

Evaluation of this work carried on a hybrid of state-of-the-art programs and practices with the synthetic attack nature like Analysis, Backdoor, DOS, Exploits, Fuzzers, Generic, Normal, Reconnaissance, Shell Code, and Worms are the ten types listed in Table 1. Selection of cluster classes done by Jaccard scores in the UNSW-NB15 dataset performance based on a multi-objective optimization problem.

Table 1. Depiction of the UNSW-NB15 da taset

Class	Aggregate	Ratio (%)
Analysis	2,000	1.141
Backdoor	1,746	0.996
DoS	12,264	6.994
Exploits	33,393	19.045
Fuzzers	18,184	10.371
Generic	40,000	22.813
Normal	56,000	31.938
Reconnaissance	10,491	5.983
Shellcode	1,133	0.646
Worms	130	0.074
Total	175,341	100

Individuals from recent populations were measured uniquely and the Jaccard score was drawn. For experimental purposes, the dataset is divided into subsets namely S1, S2, and S3, the intrusion detection, and prediction capability are in Table 2.

Table 2. Accurateness evaluation with diverse feature subclasses

Feature Set	No. of features	Feature set
All Features	42	src_bytes’, ’dst_bytes’, ’count’, ’rerror_rate’, ’dst_host_ same_srv_rate’, ’dst_host_diff_srv_rate’,’dst_host_ same _src_port_rate’, ’dst_host_rerror_rate’, ’service_finger’, ’service_ftp_data’,’service_http’, ’service_private’, ’service_smtp’,’service_telnet’
S1	20	src_bytes’, ’dst_bytes’, ’wrong_fragment’, ’num_ compromised’, ’same_srv_rate’, ’dst_host_serror_ rate’, ’dst_host_srv_serror_rate’, ’service_ecr_i’
S2	17	’sbytes’, ’sttl’, ’sload’, ’sinpkt’, ’dinpkt’, ’stcpb’, ’tcprtt’, ’synack’, ’smean’, ’ct_srv_src’, ’ct_srv_dst’
S3	21	duration’, ’src_bytes’, ’dst_bytes’, ’hot’, ’num_failed_logins’, ’num_root’, ’num_access_files’, ’dst_host_count’, ’dst_host_srv_count’, ’dst_host_same_srv_rate’, ’dst_host_same_src_port_rate’,’dst_host_srv_diff_host_rate’, ’service_ftp_data’,’service_imap4’

In Table 3 Consider, the complete set of the features along with subset features based on attack classes the method selects only three subset of features for classification of intrusion to predicts the accuracy. Nevertheless, the proposed SCMGSM method consists of all 42 features but prediction accuracy is lower than the existing methods. Divide the dataset into subsets and consider the equal size of the feature set for experimental data selection the proposed SCMGSM performs better to predict the intrusion as shown and also attains the maximum prediction accuracy level of 98.81% of features designed as subset S_2.

Table 3. Performance of SCMGSM on Class-based feature set

Methods	Features/ Subset	Detection Rate	False Positive Rate	Classification
KNN [8] **RF** [2] **CFS-BA** [5] **DPEA** [9] **LOF** [1] **Proposed Method**	Full set of Features (42)	87.93 90.15 91.23 90.92 93.44 97.11	5.05 4.78 3.12 3.45 2.56 1.55	91.12 90.45 93.22 89.96 88.67 98.87
KNN [8] **RF** [2] **CFS-BA** [5] **DPEA** [9] **LOF** [1] **Proposed Method**	S1 (20)	90.45 89.67 92.54 93.15 95.00 97.86	4.78 3.95 3.24 2.75 2.64 1.89	90.98 92.13 91.34 92.30 92.56 93.77
KNN [8] **RF** [2] **CFS-BA** [5] **DPEA** [9] **LOF** [1] **Proposed Method**	S2 (17)	88.23 92.13 92.45 90.77 94.65 98.88	3.56 2.67 2.45 2.40 2.33 0.78	92.37 93.44 93.76 94.98 95.31 98.81
KNN [8] **RF** [2] **CFS-BA** [5] **DPEA** [9] **LOF** [1] **Proposed Method**	S3 (21)	90.23 91.43 89.45 90.77 95.35 96.98	4.56 3.63 2.11 2..10 1.33 1.10	84.37 89.44 90.76 92.98 94.39 95.89

To compute the classification results corresponding to Table 3 of the S_2 subset has a characteristic, the maximum detection rate also computes the attack classification. Table 3 shows the prediction results based on the process of attack classes. The proposed SCMGSM feature selection approach reduces recital faults in the sorting model and tests sample estimate. To find the optimum features mined in the recommended process by SCMGSM assessed at each step of the optimization algorithm iteration to improve the rapidity of cluster crusade to individual data compare along with existing methods also finds that the RF method for the subset, S2 provides a better result than the other existing method.

For Subset S2 the proposed SCMGSM provides better results for the S2 subset, a higher prediction, classification, and detection rate, and as much as compared with the minimum false positive rate of other attacks approximately compared with existing methods as shown in table 3. In attacks class, the proposed method responds more than 10.55% respectively on the detection, reducing the false-positive rate to 2.78% and the classification rate to 6.44%, of the KNN, For Comparing LOF the SCMGSM improves 4.23% of detection, 3.50% of classification rate and also reduces the false-alarm rate to 1.55%. For the full set of features, the proposed method makes only a minimum difference over the existing method. Against the KNN method, results were more than 9.81%, 7.75% of detection and classification rate, and false alarms reduced to 3.50%. CFS-BA provides better than other existing methods but not much more than the SCMGSM to 5.88% of detection rate, 5.65% of classification rate with a reduced false-positive alarm rate of 1.57%. Comparing the outperformance of the proposed system against the existing methods, the class-based classification is an innovative classification method for intrusion detection.

Figure 2 gives the comparison of the proposed SCMGSM against the existing method for class-based classification and shows the prediction accuracy of the semi-supervised based multi-objective genomic search method’s performance based on the attack feature class.

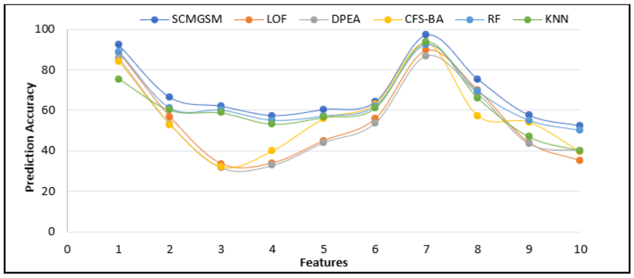

Fig 2. Prediction Accuracy Comparison of SCMGSM against Features

Figure 3 shows that the proposed method improves the accuracy of prediction on analysis as if attack classes Backdoor, Generic, reconnaissance, and worms attacks effectively comparing the Random Forest (RF) Method.

In addition, Figure.3 gives the proposed SCMGSM performance of the attack class classification on the subset $S_2$of the UNSW-NB15 Dataset.

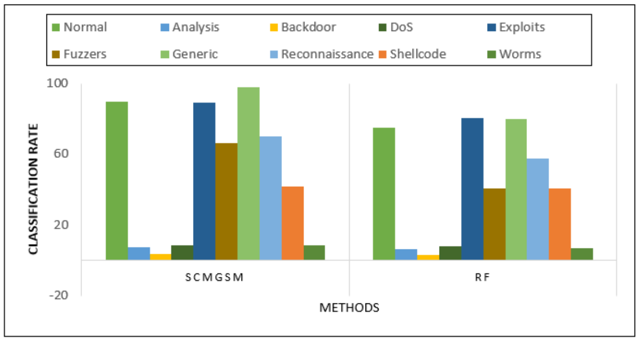

Fig 3. Performance of Attack Class-based features classification

Class-based detection rates of the attack feature are considered from the subset S2 of the UNSW-NB15 dataset. The SCMGSM gives much better than the RF for class-based detection on an individual cluster, attack detection details in Figure 4.

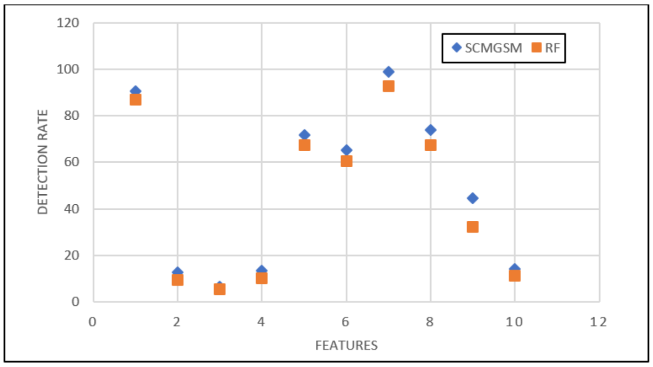

Fig 4. Class-based detection rate prediction based on attack class features of the S_2 dataset

4 Conclusion

The proposed approach performs a more extensive exploratory search by utilizing the concept of semi-supervised clustering. Using the given feature set of multi-objective optimization functions outperforms RF on several evaluation metrics while remaining competitive on all others. According to the proposed work, the SCMGSM method trusts to differentiate the attack classes and normal classes up to 98.81%, which is higher than the RF of 95.31%, considering the classification of 5.37%, the detection rate of 6.75% more than the RF also provides a minimum false positive rate than the RF by using S_2 subset on UNSW-NB15 dataset. In this work, the inter-IDS gathering pace is cast-off first to cluster parallel features generated by dissimilar IDSs composed in order to avoid dismissal. The presented work considers only subsets for class-based classifications to provide better accuracy, detection, and classification of features. In the next step, the multicast clustering technique with the base classifiers on the class-based classification of feature set along with optimization application enables wireless sensor networks.

References

1) Megantara AA, Ahmad T. A hybrid machine learning method for increasing the performance of network intrusion detection systems. *Journal of Big Data*. 2021;8(1):142–143. Available from: https://dx.doi.org/10.1186/s40537-021-00531-w.
2) Ahmad Z, Khan AS, Shiang CW, Abdullah J, Ahmad F. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*. 2021;32(1). Available from: https://dx.doi.org/10.1002/ett.4150.
3) Metiaf A, Wu Q, Aljeroudi Y. Searching With Direction Awareness: Multi-Objective Genetic Algorithm Based on Angle Quantization and Crowding Distance MOGA-AQCD. *IEEE Access*. 2019;7:10196–10207. Available from: https://dx.doi.org/10.1109/access.2018.2890461.
4) Chandrashekar G, Sahin F. A survey on feature selection methods. *Computers and Electrical Engineering*. 2014;40(1):16–28. Available from: https://dx.doi.org/10.1016/j.compeleceng.2013.11.024.
5) Cheng J, Li M, Tang X, Sheng VS, Liu Y, Guo W. Flow Correlation Degree Optimization Driven Random Forest for Detecting DDoS Attacks in Cloud Computing. *Security and Communication Networks*. 2018;2018:1–14. Available from: https://dx.doi.org/10.1155/2018/6459326.
6) Zhao C, Xin Y, Li X, Yang Y, Chen Y. A Heterogeneous Ensemble Learning Framework for Spam Detection in Social Networks with Imbalanced Data. *Applied Sciences*. 2020;10(3):936–936. Available from: https://dx.doi.org/10.3390/app10030936.
7) Diro AA, Chilamkurti N. Distributed attack detection scheme using deep learning approach for Internet of Things. *Future Generation Computer Systems*. 2018;82:761–768. Available from: https://dx.doi.org/10.1016/j.future.2017.08.043.
8) Kocher G, Kumar G. Analysis of Machine Learning Algorithms with Feature Selection for Intrusion Detection using UNSW-NB15 Dataset. *International Journal of Network Security &amp; Its Applications*. 2021;13(1):21–31. Available from: https://dx.doi.org/10.5121/ijnsa.2021.13102.
9) Janarthanam S, Sukumaran S, Shanthakumar M. Active Salient Component Classifier System on LocalFeatures for Image Retrieval. *Indian Journal of Science and Technology*. 2017;10(26):1–9. doi:10.17485/ijst/2017/v10i26/112405.
10) Liu H, Zhou M, Liu Q. An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*. 2019;6(3):703–715. Available from: https://dx.doi.org/10.1109/jas.2019.1911447.
11) Saxena AK, Sinha S, Shukla P. A review on intrusion detection system in mobile ad-hoc network. *2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE)*. 2017;p. 549–554.
12) Torabi M, Udzir NI, Abdullah MT, Yaakob R. A Review on Feature Selection and Ensemble Techniques for Intrusion Detection System. *International Journal of Advanced Computer Science and Applications*. 2021;12(5):538–553. Available from: https://dx.doi.org/10.14569/ijacsa.2021.0120566.
13) Vinayakumar R, Alazab M, Soman KP, Poornachandran P, Al-Nemrat A, Venkatraman S. Deep Learning Approach for Intelligent Intrusion Detection System. *IEEE Access*. 2019;7:41525–41550. Available from: https://dx.doi.org/10.1109/access.2019.2895334.
14) Wang Y, Wang YX, Singh A. A Theoretical Analysis of Noisy Sparse Subspace Clustering on Dimensionality-Reduced Data. *IEEE Transactions on Information Theory*. 2019;65(2):685–706. Available from: https://dx.doi.org/10.1109/tit.2018.2879912.
15) Zhang Q, Li R, Chu T. Kernel semi-supervised graph embedding model for multimodal and mixmodal data. *Science China Information Sciences*. 2020;63(1):119–204. Available from: https://dx.doi.org/10.1007/s11432-018-9535-9.