Support Based Graph Framework for Effective Intrusion Detection and Classification

Rahul B Adhao (badhaor@gmail.com)
College of Engineering Pune (COEP) India
Vinod K Pachghare
College of Engineering Pune (COEP) India

Research Article

Keywords: Normalization, Features selection, Optimization, Support value measure, Classification

Posted Date: November 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1035364/v1

SUPPORT BASED GRAPH FRAMEWORK FOR EFFECTIVE INTRUSION DETECTION AND CLASSIFICATION

[*1]Rahul B Adhao, [2] Vinod K Pachghare
[*1] *Researcher at Department of Computer Engineering, College of Engineering Pune (COEP) India.*
[2] *Associate Professor at Department of Computer Engineering, College of Engineering Pune (COEP) India.*
**Email Id: badhaor@gmail.com*

ABSTRACT
Intrusion Detection System is one of the worthwhile areas for researchers for a long. Numbers of researchers have worked for increasing the efficiency of Intrusion Detection Systems. But still, many challenges are present in modern Intrusion Detection Systems. One of the major challenges is controlling the false positive rate. In this paper, we have presented an efficient soft computing framework for the classification of intrusion detection dataset to diminish a false positive rate. The proposed processing steps are described as; the input data is at first pre-processed by the normalization process. Afterward, optimal features are chosen for the dimensionality decrease utilizing krill herd optimization. Here, the effective feature assortment is utilized to enhance classification accuracy. Support value is then estimated from ideally chosen features and lastly, a support value-based graph is created for the powerful classification of data into intrusion or normal. The exploratory outcomes demonstrate that the presented technique outperforms the existing techniques regarding different performance examinations like execution time, accuracy, false-positive rate, and their intrusion detection model increases the detection rate and decreases the false rate.

***Key words:* -** Normalization, Features selection, Optimization, Support value measure, Classification

Declarations

Conflicts of interest/Competing interests

Compliance with ethical standards Conflict of interest The authors declare that they have no conflict of interest.

1. INTRODUCTION

With the progress of the Internet in the modern world, security threats to computer systems and the network have improved a lot. The security threats influence system security administrations. To control security threats number of innovations are established and organized in administrations, for example, firewall, anti-virus software, message encryption, secured software protocols, and so on. Likewise, this Intrusion Detection is a significant innovation that has existed for a long time [1-3]. Intrusion detection is one of the significant presentations of outlier identification that is utilized to recognize the system attacks by opponents. Intrusion Detection Systems (IDSs) are fundamental to guarantee system security [4, 5]. The generally utilized methodologies for intrusion detection are the anomaly and signature dependent methodologies [6]. Anomaly-dependent IDSs gain proficiency with the benchmark for the system conduct and any occasion that falls outside the accepted behavior is confirmed as a malicious event. The signature dependent IDSs acquire the normal and anomalous events to identify the attacks with their types [7].

The Intrusion detection system (IDS) is a significant component of secure information systems [8, 9]. Intruders in the network are attempting to access the unapproved resources in the system [10]. It is highly required to screen and examine the actions of the user and framework behaviors. Essentially, by adjusting the arrangement of the system parameters, the

conduct of the framework could be unpredictable. Subsequently, the framework must be furnished with the highlights for the intermittent observing and its conduct standards both for normal and abnormal activities [11-13]. Machine learning techniques can be effective in detecting intrusions. Numerous Intrusion Detection Systems are displayed dependent on machine learning strategies [14].

Machine learning is a common term for depicting a lot of optimization and processing techniques that are lenient of roughness and vulnerability. Currently, a machine learning framework has been stretched out for executing a successful interruption location framework. Machine learning approaches are exceptionally useful and enhanced in current intrusion detection [15-17]. Precisely, support vector machines, neural networks, decision trees to have powerful important plans in anomaly detection structures to improve the characterization execution and speed [18]. The key components of machine learning procedures are Fuzzy Logic (FL), Artificial Neural Networks (ANNs), Probabilistic Reasoning (PR), and Genetic Algorithms (GAs). The idea behind the application of soft computing techniques and particularly ANNs in implementing IDSs is to include an intelligent agent in the system that is capable of disclosing the latent patterns in abnormal and normal connection audit records and to generalize the patterns to new connection records of the same class [19, 20].

- Optimum features are selected from the normalized information using krill herd optimization for the features dimensionality reduction.
- Effective feature selection using krill herd optimization enhances the classification accuracy by diminishing the false positive rate. Support value is estimated for effectually selected features.
- Support based graph is constructed for the effective classification of data into intrusion or normal.

The structure of the manuscript is sorted out as pursues: Section 2 reviews the literature works in regards to the proposed strategy. In section 3, a short discussion about the proposed system is given, section 4 examines the exploratory outcomes, and section 5 finishes up the paper.

2. RELATED WORK

Majjed Al-Qatf *et al.* [21] proposed a powerful deep learning method STL-IDS dependent on the self-trained learning (STL) system. The presented methodology was utilized to feature learning and dimensionality reduction and improves the prediction accuracy of support vector machines (SVM) concerning attacks. The presented approach was assembled utilizing the sparse auto-encoder system, which was an effective learning algorithm for reconstructing a new feature representation in an unsupervised method. After the pre-processing step, the new feature is fed into the SVM to enhance its prediction capacity for intrusion and classification accuracy.

Farrukh Aslam Khan *et al.* [22] presented a novel two-stage deep learning (TSDL) dependent on a stacked auto-encoder for proficient system intrusion detection. The model contains 2 decision steps: the first step is responsible for classifying network traffic as normal or abnormal using a probability score value. Secondly, it was utilized in the final decision step as an additional. The presented model was able to learn useful feature representations from large amounts of unlabelled data and classifies them automatically and efficiently.

Chuanlong Yin *et al.* [23] presented an intrusion detection system-dependent based on deep learning, and we propose a deep learning approach for intrusion detection using recurrent neural networks (RNN-IDS). Besides, they investigated the performance of the design in binary classification and multiclass classification, and the number of neurons and different learning rate impacts on the performance of the proposed model. They analyzed the presented strategy with existing soft computing techniques presented by previous researchers on the benchmark data set.

Jie Gu *et al.* [24] presented a methodology for intrusion detection using an SVM ensemble with feature extension. Precisely, the logarithm marginal density ratios transformation was implemented on the original features to obtain new and better-quality transformed training data; the SVM ensemble was then used to build the intrusion detection model. Exploratory outcomes demonstrate that the presented technique can achieve a good and robust execution.

Haipeng Yao *et al.* [25] presented an MSML framework that incorporates components such as, pure cluster mining, pattern discovery, fine-grained classification, and model updating. In the pure cluster module, they presented knowledge of pure cluster formation and presented a hierarchical semi-supervised k-means calculation mean to discover all the unadulterated clusters. In the pattern discovery model, they defined the unknown pattern and apply a cluster-based technique intending to locate those unknown patterns. At that point, a test was sentenced to mark known examples or unlabelled unknown patterns. The fine-grained classification module can achieve fine-grained classification for those unknown pattern samples.

3. PROPOSED METHODOLOGY

In this paper, we have proposed an efficient methodology for the classification of intrusion detection data. The input information is first pre-processed by the normalization procedure. Afterward, the finest features are chosen from the normalized information for the highlights dimensionality decrease utilizing krill herd optimization. Here, the viable feature determination enhances the classification precision by diminishing the false positive rate. At that point, support value is estimated for effectually chosen features, and then a support based graph is constructed for the effective classification of data into intrusion or normal. The block diagram of the presented methodology is given in figure 1.

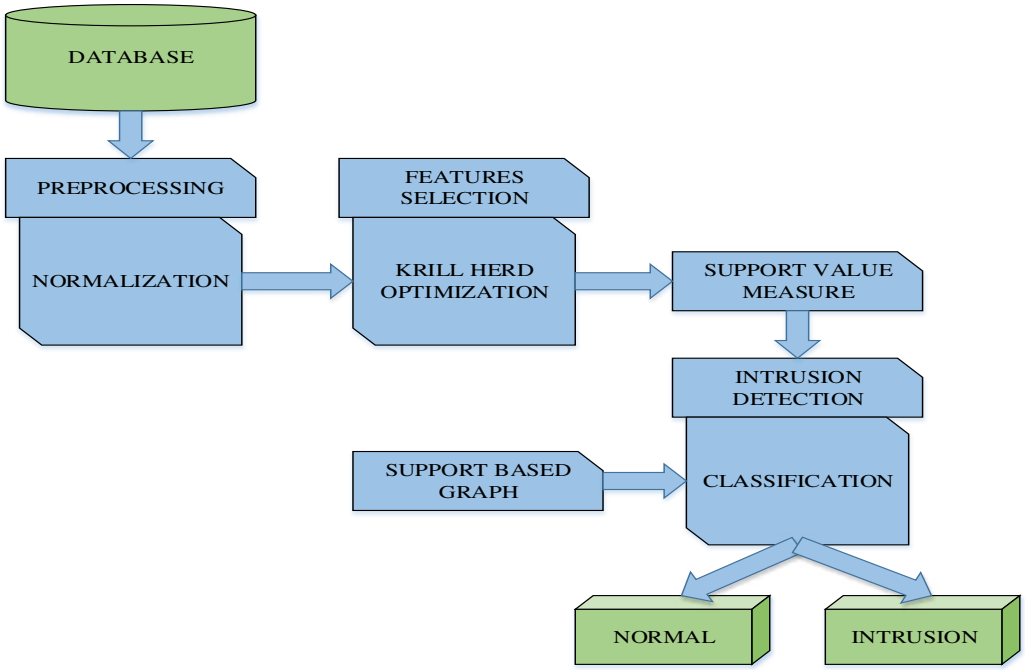

Figure: 1 Block diagram of the proposed methodology

3.1 DATA COLLECTION

3.1.1 DATASET 1: KDD CUP 99 dataset [26]

KDD'99 has been the most generally utilized dataset for the assessment of anomaly detection techniques. This dataset is created dependent on the information captured in the DARPA'98 IDS evaluation program. DARPA'98 is around 4 gigabytes of packet information

of 7 weeks of network traffic, which is handled into around 5 million association records, each with around 100 bytes. The two weeks of test information about 2 million association records. KDD training dataset comprises roughly 4,900,000 single association vectors every one of which comprises 41 features and is marked as either normal or an attack, with precisely one specific attack type.

3.1.2 DATASET 2: CIC IDS 2017 dataset [27]

CIC IDS 2017 comprises 5 days of information accumulation with 225,745 packages with more than 80 features and gathered over seven days of network activity. In the CIC 2017 dataset, the attack simulation is isolated into seven classes including Brute Force Attack, Heart Bleed Attack, Botnet, DoS Attack, DDoS Attack, Web Attack, and Infiltration Attack.

3.1.3 DATASET 3: ISCX IDS 2012 dataset [28]

The UNB (University of New Brunswick) ISCX 2012 dataset signifies powerfully created information that reflects system traffic and interruptions. Different multi-stage attack scenarios are carried to stream the anomalous segment of the dataset. Normal background traffic is given by performing client profiles that were artificially produced at arbitrary synchronized times making profile based client behavior.

3.1.4 DATASET 4: CICDDOS 2019 dataset [29]

Distributed Denial of Service (DDoS) attack is a menace to network security that aims at exhausting the target networks with malicious traffic. CICDDoS2019 dataset contains benign and the most up to date common DDoS attacks, which resembles the true real-world data.

3.2 PREPROCESSING

3.2.1 Normalization

Normalization performs the direct change of input information to fit into a particular range. Here, Min-max normalization is utilized for the standardization of data which linearly transforms the data. Min-Max normalization is regularly done through the accompanying condition,

$$Y = \frac{Y - Y_{\min}}{Y_{\max} - Y_{\min}} \tag{1}$$

Where, $Y_{\min}$ and $Y_{\max}$ are the minimum and maximum values in Y, and Y is the set of values in the dataset.

3.3 FEATURE SELECTION

Feature selection is portrayed as a technique whereby particular features are chosen from a set of features, which have high discrimination ability among class labels. It is a significant and regularly utilized method in numerous fields for dimension reduction. Feature determination is imperative in enhancing proficiency and besides for decreasing dimensions. In the proposed technique, krill herd optimization is utilized for the viable feature selection.

3.3.1 Krill Herd optimization algorithm

This is an iterative heuristic technique involved in the inalienable phenomenon of the krill herd [31]. This is primarily utilized for resolving optimization issues. The pseudo-code of krill herd optimization is represented in algorithm 1.

Begin

Define the size of the populace (S') and cycle ($\hat{C}_{\max}$)

Initialization
Set cycle $C'=1$;
Initialize the cluster information as input and population data
$\tilde{S}=1,2,3,.....S'$ Of krill arbitrarily.
Fitness assessment
Evaluate each krill as specified by the krill location
While $C'<\hat{C}_{\max}$ **do**
Class the populace/krill from finest to extremely worst.
for $i=1:S'$ do
Perform the accompanying motion calculations,
1) *Movement actuated by the krill*
2) *Foraging action*
3) *Physical dispersion*
Update the krill location in the inquiry space.
Evaluate each krill according to its location.
end for i
Categorize the krill from the finest to the poorest and locate the present best.
$\hat{C}_{\max}=C'+1$.
End while
Estimate the krill finest result.
End

Algorithm 1: Pseudo-code of krill herd optimization algorithm

Algorithm description

The presented krill herd optimization results in the effective chosen of features through the associated steps:

Step 1

The optimization begins with the initialization of normalized data.

Step 2

Fitness esteem is evaluated dependent on the krill individual positions.

Step 3

Next, the fundamental iteration begins by positioning the krill from the finest to the observably bad individual.

Step 4

From that point onwards, motion updates are processed for every krill using the going with conditions,

a) The searching update is finished by,

$$\overline{F}_z(\hat{t}+1)=S_f\beta_x+\omega_i\overline{F}_z(k') \tag{9}$$

$$\beta_z=\beta_z^{food}+\beta_z^{best} \tag{10}$$

Where, S_f denotes the foraging speed, ω_i denotes the inertia weight, β_z^{best} denotes the finest result of the z^{th} krill individual.

b) The induced motion identifies with the density maintenance of data is represented as,

$$\overline{M}_z(\hat{t}+1)=\overline{M}_{\max}\alpha_z+\omega_i+\overline{M}_z(\hat{t}) \tag{11}$$

$$\alpha_z=\alpha_z^{total}+\alpha_z^{target} \tag{12}$$

Where, $\overline{M}_{\max}$ denotes the most extreme activated speed, ω_i denotes the inertia weight, α_z^{total} denotes the nearby effect of the z^{th} krill individual has on its neighbors, α_z^{target} is the finest result of the z^{th} krill.

c) The final movement update is matching the physical distribution through irregular action and is represented as,

$$\overline{D}_y(\hat{t}+1) = \overline{D}_{\max}(\frac{1-i}{i_{\max}})\delta \tag{13}$$

Where $\overline{D}_{\max}$ denotes the greatest diffusion speed, δ denotes the random directional vector between -1 and 1.

Step 5

Because of the previously showed developments, using unique parameters of movement amidst the time, the location of the y^{th} krill amidst an opportunity to $\hat{t}+\Delta\hat{t}$ passed on by the associated condition and it is utilized to compute a krill individual location.

$$\overline{K}_z(\hat{t}+\Delta\hat{t}) = \overline{K}_z(\hat{t}) + \Delta\hat{t}\frac{d\overline{K}_z}{d\hat{t}} \tag{14}$$

Where $\Delta\hat{t}$ denotes a fundamental constant. Hereby utilizing the reference condition, the krill individual's position is refreshed and the finest result is acquired.

Step 6

Toward the end, the stopping condition is utilized for the fulfillment of function assessments. Though the stopping condition is not reached once more, categorize the krill populace from the finest to the poorest and assess the finest node individual location. The flow chart of krill herd optimization is represented in figure 3.

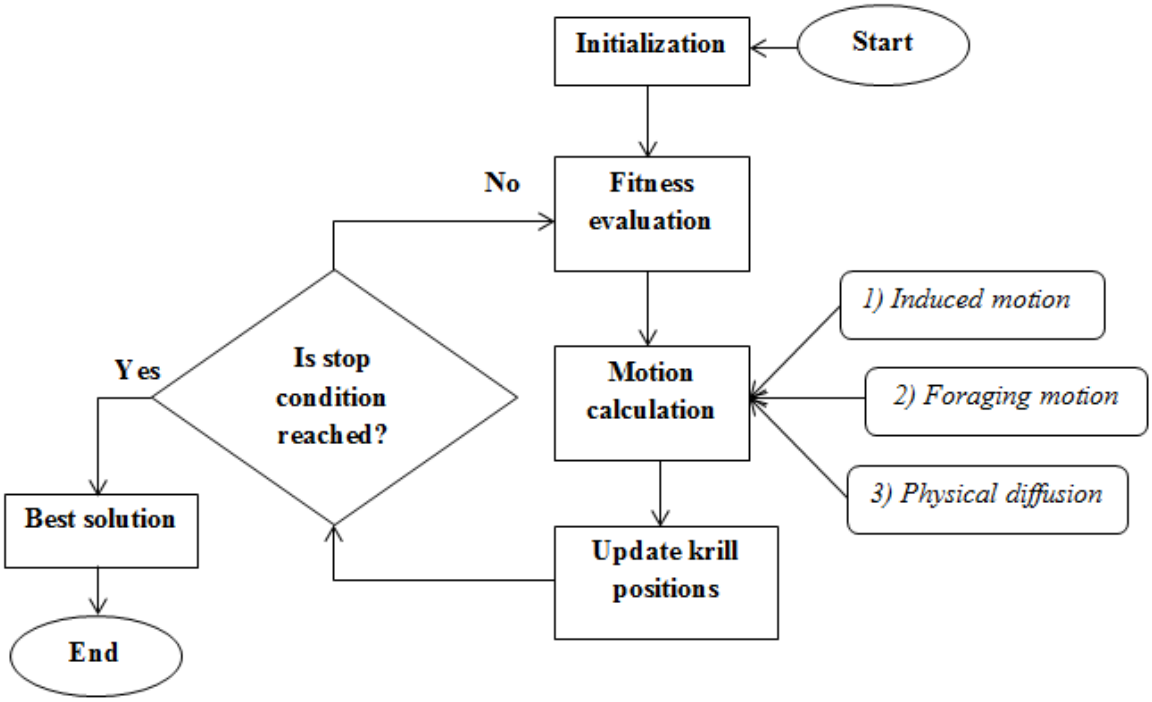

Figure 2: Flow diagram of krill herd optimization

3.4 Support value-based graph classification

In this section, a support value-based graph is utilized for the successful categorization of information into normal or intrusion. Support values are estimated for the chosen feature set and afterward, the average is computed from the support values. Consequently, the Median

support value is kept as a threshold for the successful classification of information into normal or intrusion.

3.4.1 Support value measure

In this section, input information is sorted reliant on the support value of features. Here, the Support value evaluation reliant on certain features is represented in condition (8).

$$\tilde{S}_{value} = (f_1 + f_2 + ...f_n) \Big/ (f_1 * f_2 * ....f_n) \tag{8}$$

Where $f_1, f_2, ...f_n$ denotes the selected optimal features set, $\tilde{S}_{value}$ denotes the support value.

3.4.2 Median support value

The median value is determined for all the support values after the estimation of the support value for the chosen features. In the presented classification, this support value measure is taken as a threshold for the significant categorization of information into normal or intrusion. The Median support value measure is processed by the condition (9).

$$M(\tilde{S}_{value}) = \frac{(N+1)_{th}}{2} \tag{9}$$

Where, M is the whole quantity of support value measures. The support value-based graph generation depends on the Median support value measure. The algorithm of the proposed support value-based graph classification is indicated in algorithm 2.

Input : $D =< f_1, f_2, f_3, ...f_k >$// Selected features dataset
Output: Support value-based graph classes

For every $f_k \in D$ do
  Estimate support value $\tilde{S}_{value}$ of f_k utilizing condition (8)
  Compute median support value ($M(\tilde{S}_{value})$) of f_k from condition (9)
Threshold $T \leftarrow Median$ sup $port\ value$
max $node \leftarrow T$
End for
For $i = 1$ to k do
If (max $node$ ==node $M(\tilde{S}_{value})$) then
Return Header node, H_{node}
Else
If (node $M(\tilde{S}_{value}) > Threshold$ max $node$)
$EG \leftarrow L(node\, e(set),)$
Else
$EG \leftarrow R(node\, e(set))$
End
End
End for
Result in the support value-based graph

Algorithm 2: Pseudo-code of support value-based graph classification

In support of value-based graph generation for the categorization of data into normal or intrusion, the input $Y =< y_1, y_2, y_3, ...y_k >$ is processed information and it is taken as an input. The support value-based graph generation in algorithm 2 yields the support value-based graph. At first for every data $y_k \in Y$ determines the support value utilizing condition (8) and the Median support value is then processed by utilizing the condition (9). The Median support value is taken as a threshold for the classification of information. The sample representation for the classification of data into a normal or intrusion utilizing support value-based graph is depicted in figure 3.

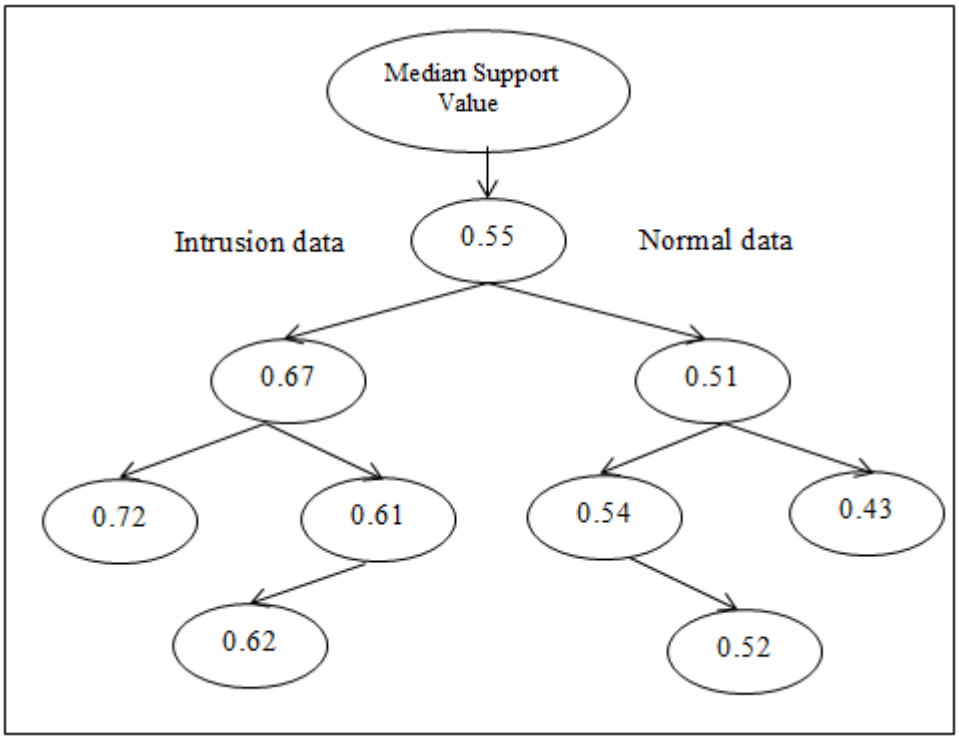

Figure 3: Sample representation of support value-based graph classification

In this classification, the support value of chosen feature esteem is higher than the threshold value then the data is located on the left side of the graph i.e.) intrusion data and if the support value of the selected feature value is lesser than the threshold value is placed in the right side correspondingly i.e.) normal data.

4. RESULTS AND DISCUSSION

The proposed support value-based graph classification was implemented in the working platform of MATLAB. In this section, the experimental outcomes accomplished for the presented technique are specified. The openly accessible KDD-CUP 99, CIC IDS 2017, ISCX IDS 2012, and CICDDOS2019 dataset was utilized to assess the classification of data into a normal or intrusion utilizing support value-based graph classification. The performance of the presented support value-based graph classification is contrasted with the existing Support vector machine (SVM) [24], Naive Bayes [23], and Random forest [23] classifiers for the accuracy, sensitivity, specificity, precision, recall, and F-measure, FPR, FNR, Kappa and Rank sum. Moreover, the presented work is analyzed with the existing optimization techniques such as genetic algorithm (GA) [18] and Particle swarm optimization (PSO) [30]. Statistical measures to examine the performance of the presented work are given in the subsequent section.

4.1 PERFORMANCE ANALYSIS

The statistical metrics of sensitivity, specificity, and accuracy can be expressed in terms of TP, FP, FN, and TN esteem. The performance of our presented work is analyzed by utilizing the statistical measures are mentioned in this section,

Accuracy

Accuracy is determined as the quantity of every single right prediction $(TN + TP)$ divided by the absolute number of a dataset $(TN + TP + FN + FP)$. It quantifies the degree of accurateness of information classification. Accuracy is ascertained by utilizing the condition (10),

$$\text{Accuracy} = \frac{(TN + TP)}{(TN + TP + FN + FP)} \tag{10}$$

Where TN is a true negative, TP is the true positive, FP is the false positive, and FN is the false negative.

Sensitivity

Sensitivity is the number of true positives that are viably recognized by a classification test. It demonstrates how extraordinary the test is at classifying the information. Sensitivity is computed by utilizing the condition (11).

$$\text{Sensitivit y} = TP/(TP + FN) \tag{11}$$

Specificity

Specificity is the number of true negatives effectively-recognized through classification tests. It recommends how great the test is at distinguishing normal data. Specificity is computed by utilizing the condition (12).

$$Specificity = TN/(TN + FP) \tag{12}$$

False-positive rate (FPR)

FPR is ascertained as the proportion among the number of negatives mistakenly measured as positives and the total amount of real negatives. False Positive Rate is computed by utilizing the condition (13).

$$FPR = \frac{FP}{FP + TN} \tag{13}$$

False-negative rate (FNR)

FNR is the degree of positives that provides negative test outcomes. The false-negative rate is computed by utilizing the condition (14).

$$FNR = \frac{FN}{FN + TP} \tag{14}$$

F-measure

It is an estimate of a test's precision. The F measure picks up its best value at 1 accompanied by the most unpleasant at 0. It is determined by the condition (15).

$$F = \frac{2TP}{2TP + FP + FN} \tag{15}$$

Receiver Operating Characteristics (ROC) curve

It is a graphical depiction tool that exhibits the intrusion detection precision against the FPR. The ROC is seen as one of the effective metrics utilized to assess the exhibition of IDSs

successfully. In the ROC curve, the best identification performance is 0% FPR and 100% TPR. Furthermore, the area under the curve of the ROC reflects detection accuracy.

4.2 DATASET 1: KDD CUP 99 dataset

Comparison table 2 delineates the performance of the presented classifier with the existing classifiers utilizing the KDD-CUP 99 dataset. It is depicted that the proposed system outcomes are highly improved than the existing classifications regarding the accuracy, sensitivity, specificity, precision, recall, F-measure, FPR, FNR, Kappa, and Rank sum test.

Table 2: Comparison table presented with existing classifiers utilizing KDD-CUP 99 dataset

Measures	Proposed	SVM	Naive Bayes	Random forest
Accuracy	**99.2**	88.1	82.7	84
Sensitivity	**98.68**	81.47	98.26	74.57
Specificity	**99.63**	95.54	76.37	97.78
Precision	**99.55**	95.35	62.83	98
Recall	**98.68**	81.47	98.26	74.57
F measure	**99.11**	87.86	76.65	84.70
FPR	**0.36**	4.45	23.62	2.21
FNR	**1.31**	18.52	1.73	25.4
Kappa	**98.53**	77.82	67.44	69.96
Rank sum	**85.75**	0.05	0.04	2.08

The features chosen of the presented work utilizing the krill herd optimization is contrasted with the existing GA, PSO optimization algorithms, and the features selection without optimization. The comparison examination provided in table 3 demonstrates that the performance of the proposed feature selection is enhanced than existing techniques.

Table 3: Comparison table of presented optimization with existing optimization utilizing KDD-CUP 99 dataset

Measures	Proposed	PSO	GA
Accuracy	99.2	87.5	78.8
Sensitivity	98.68	79.56	70.20
Specificity	99.63	97.31	91.37
Precision	99.55	97.34	92.25
Recall	98.68	79.56	70.20
F measure	99.11	87.56	79.73
FPR	0.36	0.26	8.62
FNR	1.31	20.43	29.79
Kappa	98.53	76.67	59.82
Rank sum	85	63	28

The classification performance outcomes are given in figure 4 and the feature selection utilizing different optimization techniques is provided in figure 5 proves that the proposed work accuracy is much greater than the existing techniques.

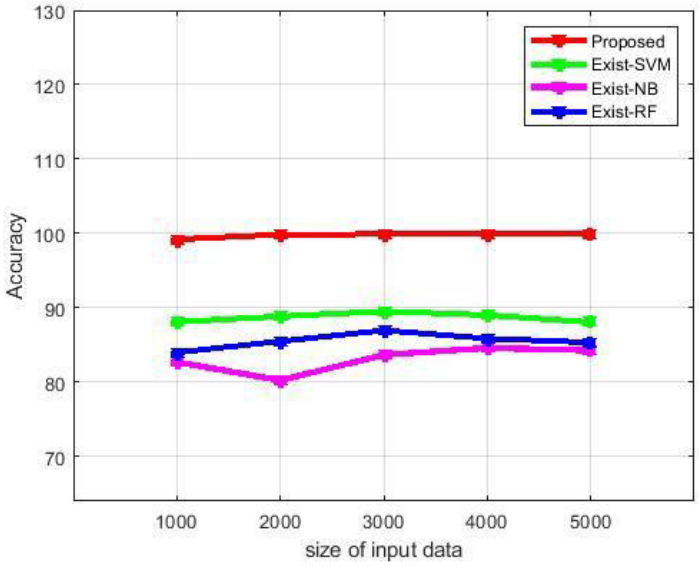

Figure 4: Comparison graph regarding accuracy with various classifiers

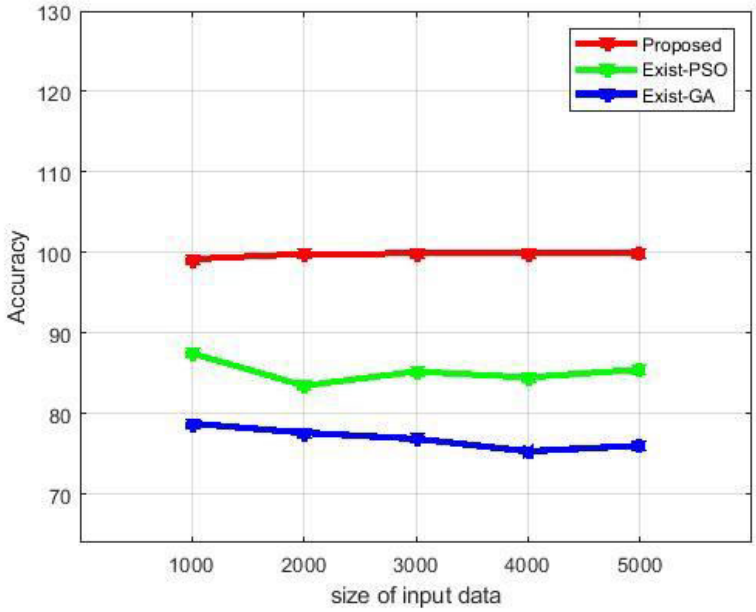

Figure 5: Comparison graph regarding accuracy with various optimizations

The comparison graph regarding accuracy portrays that the accuracy of the proposed classifier and the presented feature selection technique. The classification accuracy of the proposed technique is 99.2% and the accuracy of proposed feature selection utilizing optimization is 99.2% when utilizing KDD Cup 99 dataset. The ROC curve is made out of FPR and TPR. It is depicted in figure 6.

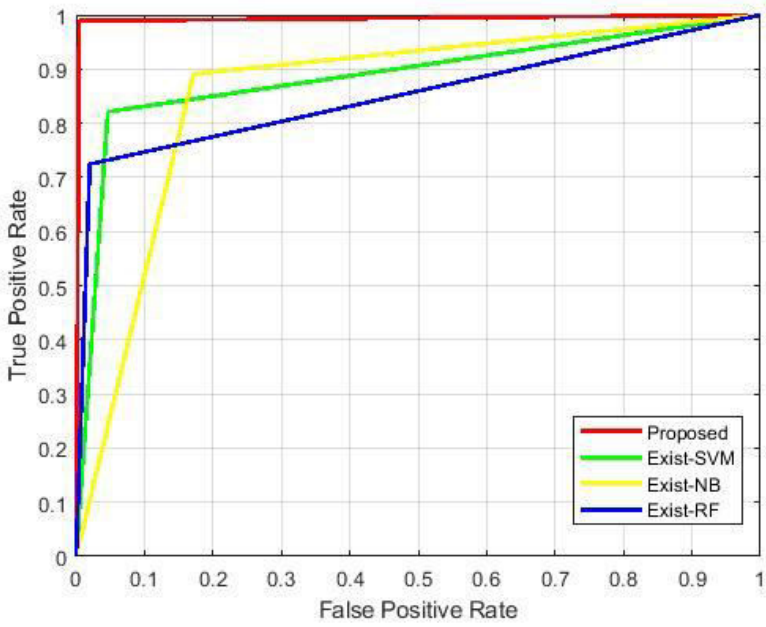

Figure 6: ROC curve with various optimizations using KDD-CUP 99 dataset

The comparison graph in terms of the ROC curve is analyzed with various existing techniques in figure 6. The ROC curve proves that the detection performance of the proposed intrusion detection is superior to the existing techniques.

4.3 DATASET 2: CIC IDS 2017 dataset

The classifier performance of the proposed methodology with existing classifiers utilizing the CIC IDS 2017 dataset is mentioned in table 4 delineates that the performance of the presented work is improved in different measures.

Table 4: Comparison table of proposed with existing classifiers utilizing CIC IDS 2017 dataset

Measures	Proposed	SVM	Naive Bayes	Random forest
Accuracy	**99.5**	87.2	82.5	86
Sensitivity	**99.4**	85.22	80.60	84.22
Specificity	**99.59**	89.4	84.64	87.97
Precision	**99.6**	90	85.6	88.6
Recall	**99.40**	85.22	80.60	84.22
F measure	**99.5**	87.54	83.02	86.35
FPR	**0.40**	10.59	15.35	12.02
FNR	**0.59**	14.77	19.39	15.77
Kappa	**99**	74.4	65	72
Rank sum	**96.43**	21.04	16.55	24.49

The performance of the proposed work features selection is contrasted with the different existing optimization techniques utilizing CIC IDS 2017 dataset is in table 5 and the obtained results prove that the proposed work performance is more prominent than the existing techniques in every performance measure.

Table 5: Comparison table of proposed optimization with existing optimization utilizing CIC IDS 2017 dataset

Measures	Proposed	PSO	GA

Accuracy	99.5	87.2	80.5
Sensitivity	99.4	86.32	82.37
Specificity	99.59	88.11	78.82
Precision	99.6	88.4	77.6
Recall	99.4	86.32	82.37
F measure	99.5	87.35	79.91
FPR	0.40	11.88	21.17
FNR	0.59	13.67	17.62
Kappa	99	74.4	61
Rank sum	96.43	59.16	19.45

The comparison graphs in figure 7 and figure 8 portray that the presented technique is superior to the number of previous optimization algorithms regarding accuracy.

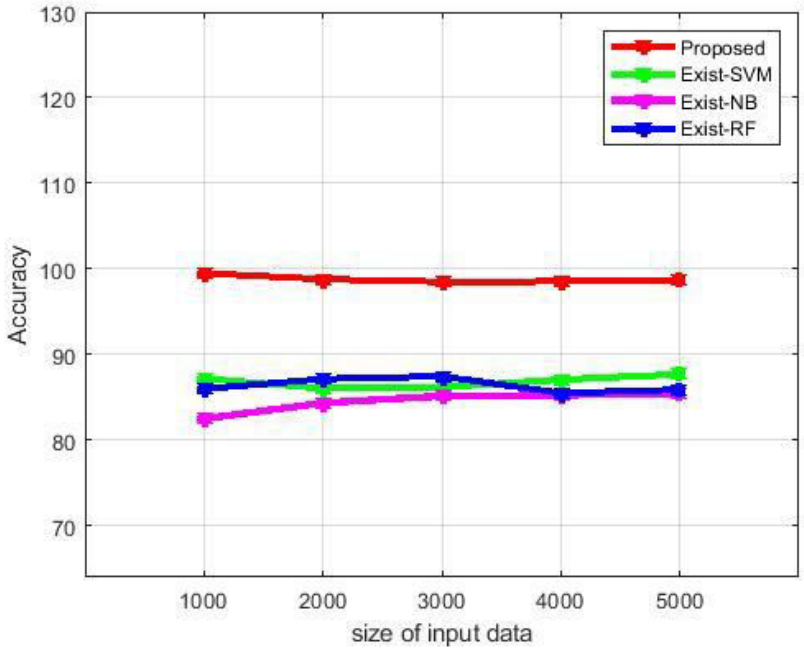

Figure 7: Comparison graph in terms of accuracy with various classifiers

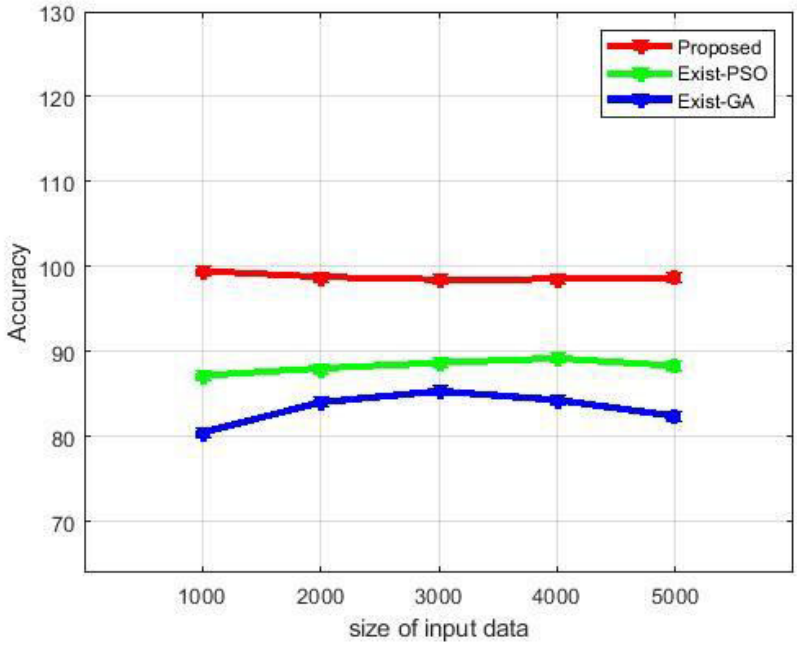

Figure 8: Comparison graph in terms of accuracy with various optimizations

The classification accuracy of the proposed technique is 99.5% and the accuracy of proposed feature selection utilizing optimization is 99.5% for CIC IDS 2017 dataset.

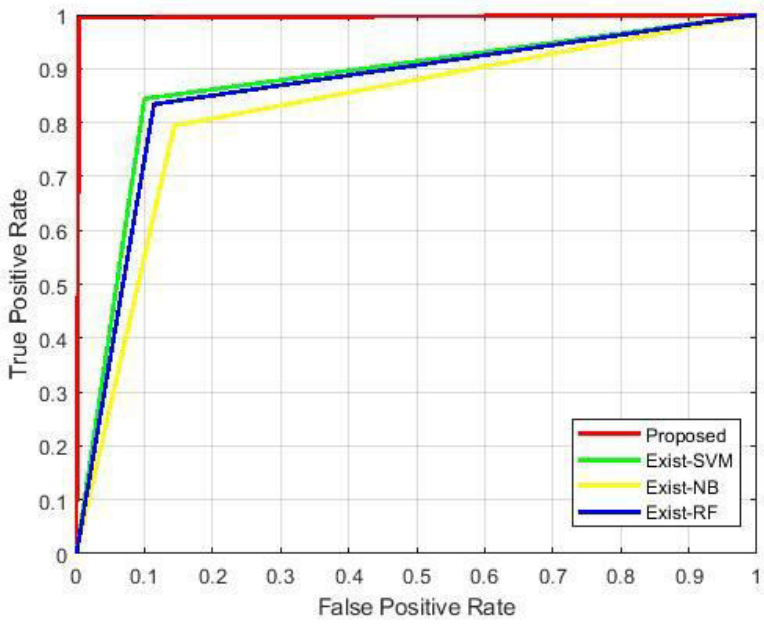

Figure 9: ROC curve with different optimizations utilizing CIC IDS 2017 dataset

The ROC curve comparison of the presented work with existing techniques utilizing the CIC IDS 2017 dataset is depicted in figure 9. It depicts that the proposed work detection performance is better than the existing technique through the ROC curve.

4.4 DATASET 3: ISCX IDS 2012 dataset

The performance of presented work features selection is contrasted with different existing classification techniques utilizing ISCX IDS 2012 dataset is in table 6 and provided results prove that the proposed work performance is greater than the existing techniques for every performance measure.

Table 6: Comparison table of proposed with existing classifiers utilizing ISCX IDS 2012 dataset

Measures	Proposed	SVM	Naive Bayes	Random forest
Accuracy	**99.5**	88.2	81.9	83.9
Sensitivity	**99.5**	92.09	88.35	90.39
Specificity	**99.49**	71.57	54.92	59.33
Precision	**99.87**	93.25	89.12	89.37
Recall	**99.50**	92.09	88.35	90.39
F measure	**99.68**	92.67	88.73	89.88
FPR	**0.50**	28.42	45.07	40.66
FNR	**0.49**	7.9	11.64	9.6
Kappa	**97.53**	56.42	41.34	45.71
Rank sum	**86.65**	57.26	69.37	61.79

The comparison table 7 delineates the proposed work feature selection utilizing krill herd optimization results in a better outcome than the existing techniques for different execution measures. Here, the performance of the proposed work is examined with the ISCX IDS 2012 dataset.

Table 7: Comparison table of proposed optimization with existing optimization utilizing ISCX IDS 2012 dataset

Measures	Proposed	PSO	GA
Accuracy	99.5	87.1	83.6
Sensitivity	99.5	91.67	90.25
Specificity	99.49	68.20	58.57
Precision	99.87	92.25	89.12
Recall	98.50	91.67	90.25
F measure	99.68	91.96	89.68
FPR	0.50	31.79	41.42
FNR	0.49	8.3	9.74
Kappa	97.53	53.49	45.04
Rank sum	86.65	77.89	57.97

The proposed work accuracy for ISCX IDS 2012 dataset is examined with different classifiers and optimization algorithms than the existing techniques are demonstrated by figure 10 and figure 11. It displays that the accuracy of the proposed work with the ISCX IDS 2012 dataset is vastly improved than the existing techniques.

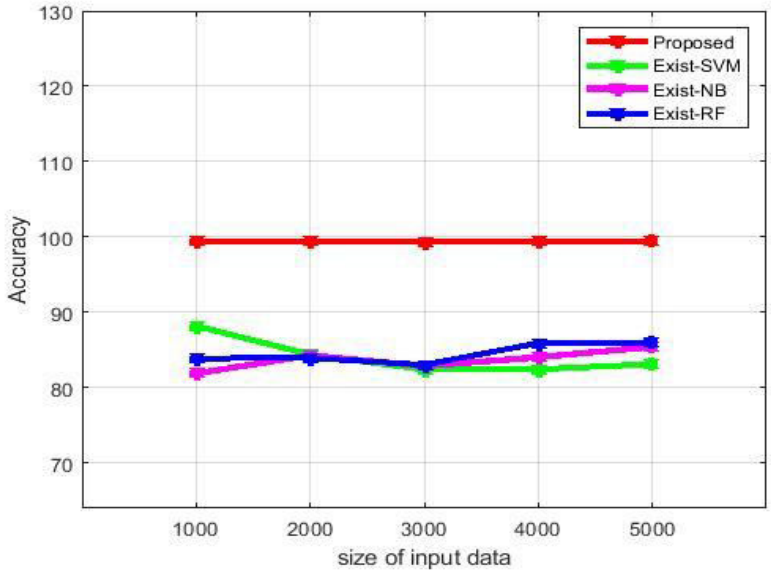

Figure 10: Comparison graph in terms of accuracy with various classifiers

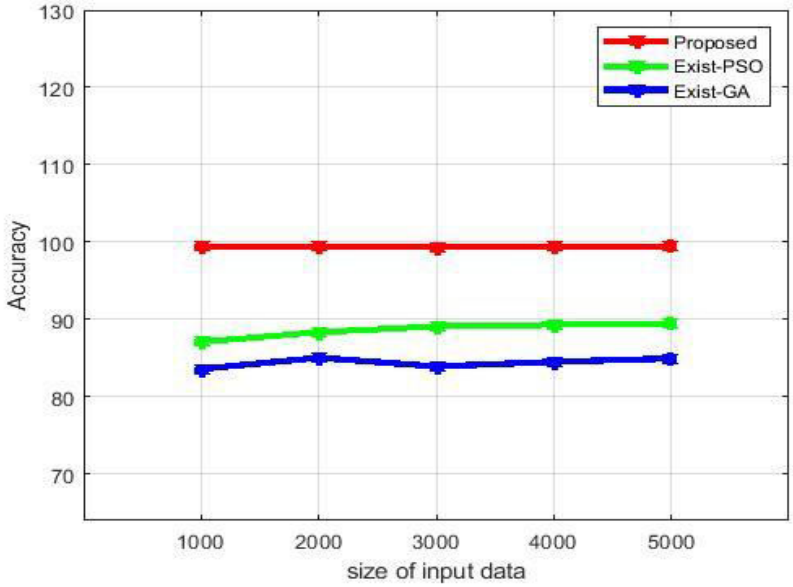

Figure 11: Comparison graph in terms of accuracy with various optimizations

The classification accuracy of the proposed technique is 99.5% and the accuracy of proposed feature selection using optimization is 99.5% when utilizing ISCX IDS 2012 dataset.

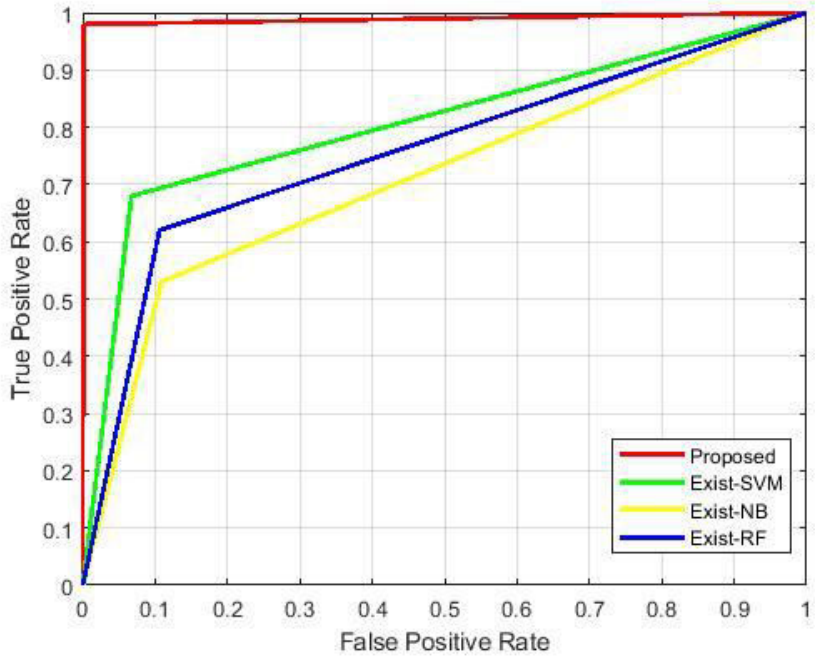

Figure 12: ROC curve with different optimizations utilizations ISCX IDS 2012 dataset

The performance of the presented system with the ISCX IDS 2012 dataset is examined by the ROC curve in figure 12. The examination proves that the prediction performance of the proposed strategy is enhanced than the existing methods.

4.5 DATASET 4: CIC DDOS 2019 dataset

Comparison table 8 depicts the performance of the proposed classifier with the existing classifiers using the CICDDOS2019 dataset. It is shown that our proposed technique performance is much better than the existing classifications in terms of accuracy, sensitivity, specificity, precision, recall, F-measure, FPR, FNR, Kappa, and Rank sum measures.

Table 8: Comparison table of proposed with existing classifiers using CICDDOS2019 dataset

Measures	Proposed	SVM	Naive Bayes	Random forest
Accuracy	**99.6**	90.3	83.6	87.4
Sensitivity	**99.2**	79.76	67.91	74.03
Specificity	**99.73**	93.94	88.55	92.04
Precision	**99.2**	82	65.2	76.4
Recall	**99.2**	79.76	67.91	74.03
F measure	**99.2**	80.86	66.53	75.19
FPR	**0.26**	6.05	11.44	7.95
FNR	**0.8**	20.23	32.08	25.96
Kappa	**99.46**	86.17	75.44	81.65
Rank sum	**100**	71.9	60.32	68.12

The proposed work features selection using the krill herd optimization is compared with the existing optimization algorithms in table 9 and the comparison analysis proves that the performance of the proposed work is improved than the existing techniques.

Table 9: Comparison table of proposed optimization with existing optimization using CICDDOS2019 dataset

Measures	Proposed	PSO	GA
Accuracy	99.6	87.5	81.6
Sensitivity	99.2	74.13	62.89
Specificity	99.73	92.17	88.03
Precision	99.2	76.8	64.4
Recall	99.2	74.13	62.89
F measure	99.2	75.44	63.63
FPR	0.26	7.82	11.96
FNR	0.8	25.86	37.10
Kappa	99.46	81.81	72.03
Rank sum	100	64.41	75.77

The classification performance and the optimization algorithms performance using the CICDDOS2019 dataset in figure 13 and figure 14 proves that the proposed work accuracy is much greater than the existing techniques.

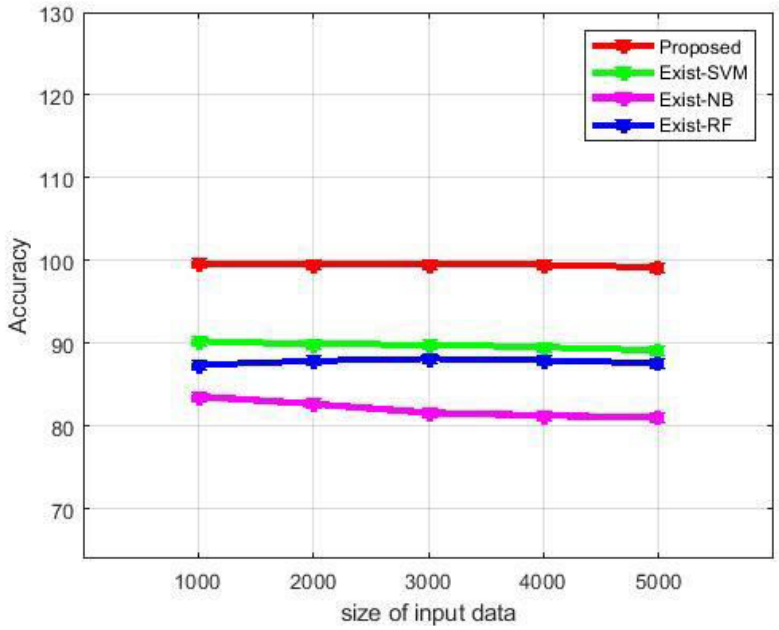

Figure 13: Comparison graph regarding accuracy with various classifiers

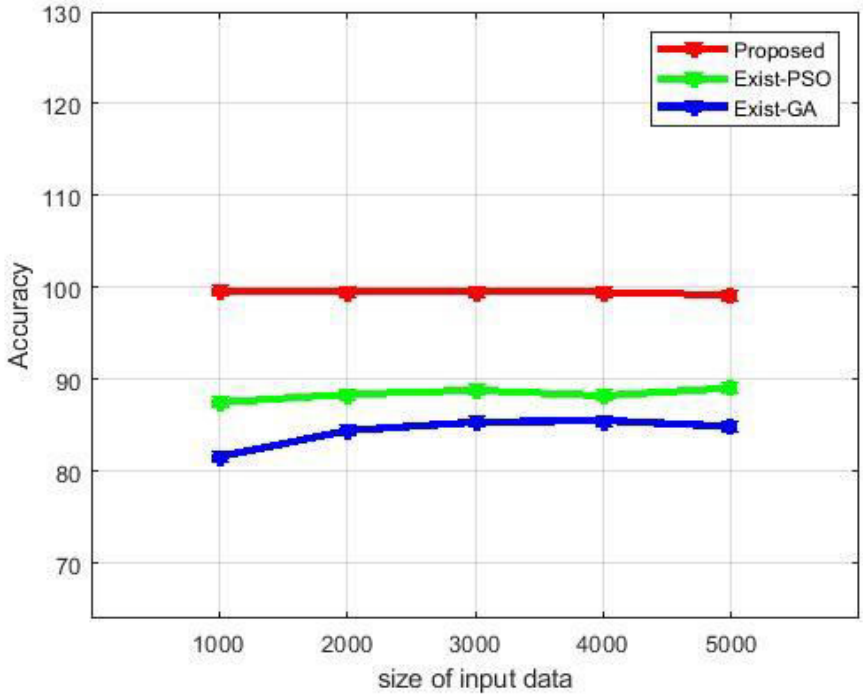

Figure 14: Comparison graph regarding accuracy with various optimizations

The accuracy measure for the proposed and the existing techniques are compared in the figure depicts that the classification accuracy of the proposed technique is 99.6% and the accuracy of proposed feature selection using optimization is 99.6% when using the CICDDOS2019 dataset.

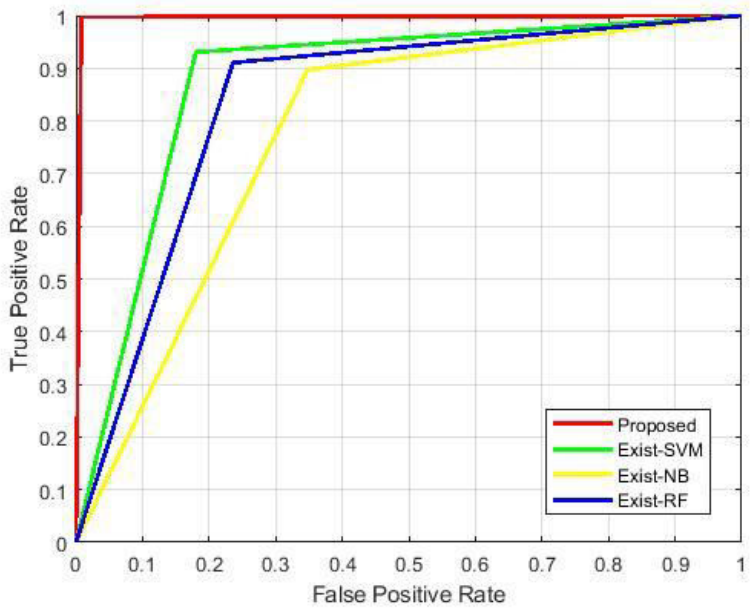

Figure 15: ROC curve with various techniques using CICDDOS2019 dataset

The ROC curve with various existing techniques in intrusion detection is given in figure 15 is analyzed with the CICDDOS2019 dataset. It displays that the detection performance of the proposed support value-based classification is better than the existing classifications.

5. CONCLUSION

In this paper, we have presented a support value-based graph classification for the categorization of data into normal or intrusion. Moreover, an optimal feature selection utilizing krill herd optimization yields superior outcomes for effectively choosing the features. In the presented technique, the input data is pre-processed and the features are ideally chosen to utilize optimization. Lastly, an effective support value-based graph classification is efficiently categorized the data into normal or intrusion. The exploratory outcomes exhibit that our presented classification outperforms the existing SVM, Naive Bayes, and random forest classifiers concerning performance metrics such as accuracy, sensitivity, specificity, precision, recall, F-measure, FPR, FNR, Kappa, and Rank sum measures.

REFERENCES

[1] Gnanaprasanambikai, L., and Nagarajan Munusamy. "Data Pre-Processing and Classification for Traffic Anomaly Intrusion Detection Using NSLKDD Dataset" Cybernetics and Information Technologies 18, no. 3 (2018): 111-119.

[2] Vinutha, Poornima, A Survey - Comparative Study on Intrusion detection System, "International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, 2015.

[3] Binbusayyis, Adel, and Thavavel Vaiyapuri. "Identifying and Benchmarking Key Features for Cyber Intrusion Detection: An Ensemble Approach." IEEE Access 7 (2019): 106495-106513.

[4] Anwar, Shahid, Jasni Mohamad Zain, Mohamad Fadli Zolkipli, Zakira Inayat, Suleman Khan, Bokolo Anthony, and Victor Chang. "From intrusion detection to an intrusion response system: fundamentals, requirements, and future directions." Algorithms 10, no. 2 (2017): 39.

[5] Ashfaq, Rana Aamir Raza, Xi-Zhao Wang, Joshua Zhexue Huang, Haider Abbas, and Yu-Lin He. "Fuzziness based semi-supervised learning approach for intrusion detection system." Information Sciences 378 (2017): 484-497.

[6] Aziz, Amira Sayed A., E. L. Sanaa, and Aboul Ella Hassanien. "Comparison of classification techniques applied for network intrusion detection and classification." Journal of Applied Logic24 (2017): 109-118.

[7] Upasani, Nilam, and Hari Om. "A modified neuro-fuzzy classifier and its parallel implementation on modern GPUs for real time intrusion detection" Applied Soft Computing (2019): 105595.

[8] Ben-Asher, Noam, and Cleotilde Gonzalez "Effects of cyber security knowledge on attack detection." Computers in Human Behavior 48 (2015): 51-61.

[9] Manzoor, Ishfaq, and Neeraj Kumar. "A feature reduced intrusion detection system using ANN classifier." Expert Systems with Applications 88 (2017): 249-257.

[10] Kevric, Jasmin, Samed Jukic, and Abdulhamit Subasi. "An effective combining classifier approach using tree algorithms for network intrusion detection." Neural Computing and Applications 28, no. 1 (2017): 1051-1058.

[11] Selvakumar, B., and K. Muneeswaran. "Firefly algorithm based feature selection for network intrusion detection." Computers & Security 81 (2019): 148-155.

[12] Aljawarneh, Shadi, Monther Aldwairi, and Muneer Bani Yassein. "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model" Journal of Computational Science 25 (2018): 152-160.

[13] Ambusaidi, Mohammed A., Xiangjian He, Priyadarsi Nanda, and Zhiyuan Tan. "Building an intrusion detection system using a filter-based feature selection algorithm" IEEE transactions on computers 65, no. 10 (2016): 2986-2998.

[14] Anjum Khan, Anjana Nigam,"Analysis of Intrusion Detection and Classification using Machine Learning Approaches", International journal of online science, 2015.

[15] Bhuyan, Monowar H., Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. "Network anomaly detection: methods, systems and tools." IEEE communications surveys & tutorials16, no. 1 (2014): 303-336.
[16] Ahmed, Mohiuddin, Abdun Naser Mahmood, and Jiankun Hu. "A survey of network anomaly detection techniques" Journal of Network and Computer Applications 60 (2016): 19-31.
[17] Shahreza, M. Lotfi, D. Moazzami, B. Moshiri, and M. R. Delavar. "Anomaly detection using a self-organizing map and particle swarm optimization." Scientia Iranica 18, no. 6 (2011): 1460-1468.
[18] Aslahi-Shahri, B. M., Rasoul Rahmani, M. Chizari, A. Maralani, M. Eslami, M. J. Golkar, and A. Ebrahimi. "A hybrid method consisting of GA and SVM for intrusion detection system" Neural computing and applications 27, no. 6 (2016): 1669-1676.
[19] Srinivas Mishra, Sateesh Kumar Pradhan, Subhendu Kumar Rath, "Performance Analysis of Network Intrusion Detection System using Back Propagation for Feed Forward Neural Network in MATLAB/SIMULINK" International Journal of Computational Engineering Research (IJCER), Vol. 8, 2018.
[20] Sunita, Swain, Badajena J. Chandrakanta, and Rout Chinmayee. "A hybrid approach of intrusion detection using ANN and FCM." European Journal of Advances in Engineering and Technology 3, no. 2 (2016): 6-14.
[21] Al-Qatf, Majjed, Yu Lasheng, Mohammed Al-Habib, and Kamal Al-Sabahi. "Deep Learning Approach Combining Sparse Autoencoder With SVM for Network Intrusion Detection" IEEE Access 6 (2018): 52843-52856.
[22] Khan, Farrukh Aslam, Abdu Gumaei, Abdelouahid Derhab, and Amir Hussain. "A Novel Two-Stage Deep Learning Model for Efficient Network Intrusion Detection." IEEE Access 7 (2019): 30373-30385.
[23] Yin, Chuanlong, Yuefei Zhu, Jinlong Fei, and Xinzheng He "A deep learning approach for intrusion detection using recurrent neural networks." Ieee Access 5 (2017): 21954-21961.
[24] Gu, Jie, Lihong Wang, Huiwen Wang, and Shanshan Wang. "A novel approach to intrusion detection using SVM ensemble with feature augmentation" Computers & Security (2019).
[25] Yao, Haipeng, Danyang Fu, Peiying Zhang, Maozhen Li, and Yunjie Liu. "MSML: A Novel Multilevel Semi-Supervised Machine Learning Framework for Intrusion Detection System." IEEE Internet of Things Journal 6, no. 2 (2018): 1949-1959.
[26]https://github.com/defcom17/NSL_KDD/blob/master/Original%20NSL%20KDD%20Zip.zip
[27] https://www.unb.ca/cic/datasets/
[28] http://ids-hogzilla.org/dataset/
[29] https://www.unb.ca/cic/datasets/ddos-2019.html
[30] Almomani, Omar. "A Feature Selection Model for Network Intrusion Detection System Based on PSO, GWO, FFA and GA Algorithms." Symmetry 12.6 (2020): 1046.
[31] Abualigah, Laith Mohammad Qasim. Feature selection and enhanced krill herd algorithm for text document clustering. Berlin: Springer, 2019.