[1]Mr. Kiran S Pawar

[2] Dr. Babasaheb J Mohite

An Article: Effective Machine Learning Methods for Feature Selection.

Journal of Electrical Systems

Abstract: - With the prompt increase of internet connections and newly developed applications, there is an increase in the risk of damage that can be accomplished through launching attacks. Meanwhile, Network Intrusion Detection systems (NIDS) are essential defence tools against intricate and developed network attacks. Due to the shortage of suitable datasets, anomaly-based methodologies in intrusion detection systems are difficult for precise deployment, analysis, and assessment. There are several such datasets such as NSL-KDDCup 1999, NSL-KDD, SNMP-MIB, CICIDoS 2017, UNSW-NB15, UKM-IDS20, CICIDS 2018, CICIDS 2019 that the researchers have used to assess the performance of offered intrusion detection approaches. The paper underlines the importance of feature selection and classification. Whereas, study discusses the challenges of the complexity of dataset, stability, and scalability issue, the feature selection with efficient classification algorithms overcome the defined problems. However, this study demonstrates the ability of IDS to identify novel and different types of attacks using validated datasets. Feature reduction depends on the various models that are available and fits to unique problem's specific subset of features using various measures.

Keywords: Network Intrusion detection System (IDS), Feature selection (FS), Machine learning (ML) algorithms, Evaluation Metrices, Datasets.

I. INTRODUCTION

The progression of the digital network is globally increased with each day. The attacks on the network also increase simultaneously due to the exposure present in the global network. A cyberattack [1] is an attempt to disable services of computers in the network to steal, destroy data, or use breached computer for an additional attack. According to [2], next coming few years, the cost of cybercrime will increase by fifteen percent yearly, reaching $10.5 trillion Us dollars yearly by 2025 from $3 trillion USD. The world needs to protect 200 zettabytes of data from cybercriminals by 2025, as Steve Morgan reports in their Cybersecurity Statistics Magazine. To stop external intrusions, an IDS is employed, and other components (such as a firewall and cryptographic methods) are implemented to intercept cyberattacks [3]. Most IDSs developed adaptive intrusion detection systems using feature selection to reduce computing burden and identify assaults effectively using machine learning and fuzzy classifiers. Designing and creating intelligent security systems that can recognize new vulnerabilities is possible with the help of the machine learning toolkit's classifiers and feature selection algorithms. The network's collected traffic is made up of several features connected to instances.

This article has a novel technical review for appreciative the efficient IDS and their types, different standard attacks behaviour datasets used to analysed the considered NIDS, feature drop techniques, and a model for Machine learning based classification. Moreover, the NIDS model is displayed as an evidence of concept based on literature. The addition of the review paper are as follows:

- Explain the different types of NIDS using efficient machine learning algorithms to recognize the difference between them.
- The explore the benchmark datasets used to perform the effective NIDS.
- Discussed the significance of feature optimization performances to understand the differentiation between them.
- Demonstrate the importance of determining the effectiveness and reliability of machine learning (ML) based classification procedures for detecting intrusion.

The paper is organized by section 1 literature review based on IDS with feature reduction section 2 Different FS Methods. Section 3 emphasizes on benchmarked datasets. Section 4 represent the different feature selection

[1] *Corresponding author: Research Scholar,
[2] Associate Professor

technique. Section 5 explores the machine learning algorithms and approaches. Section 6 comparative analysis of methods used in IDS. Section 7 conclusion of the article.

II. LITERATURE SURVEY

The growth of DDoS attacks, combined with the failure of traditional network- based detection procedures, usually requires the development of efficient attack recognition systems. Several data mining methods and machine learning (ML) based algorithms have been settled for the detection and prevention of different DDoS attacks [4]. The research [5] concentrates on the reduction of the features using machine learning algorithms to select small-time detection and keep the precision detection. The proposed system uses the REP Tree, Random Forest, Random Tree, Decision Stump, and PART classification methods. The PART classifier performed the detection rate of 99.77% on CICDS2017 dataset and validated it on CICDS2019 dataset. The article [6] has developed the model to build Intrusion detection system using the ML algorithms for attack detection. The proposed model uses the backward elimination approach and reduces the seven features from the 42 features from UNSW-NB 15 dataset. The experiment performed on the binary classification with the Multi-Layer Perceptron (MLP), Gradient Boosting Tree (GBT), Random Forest (RF), and Decision Tree (DT) ML algorithms using chi-square. The performance analysis revealed the Decision Tree (DT) is the best classifier in terms of extreme accuracy with minimum false alarm rate on UNSW-NB 15.

The article [7] analysed the execution of feature selection by removing the noisy (NaN values) features from UKM-IDS20 dataset. The relief performs the accuracy of 99.9690% with 1.94 seconds with 30 features using FURIA classifier.

The proposed [8] Ensemble techniques (EFFST) obtain the 24 features out of 85 attributes with DoS, DDoS, Botnet, infiltration attacks. The ensemble method achieves the 99.9909% detection rate. The ensemble feature selection technique uses the one-fourth split from the top ranked feature and created the compact subset. The threshold technique applied on selected subsets; the system selects the feature which is existing in at least two subsets. The system is implemented and validated on CICDS2017 dataset with J 48 classifier to detects the web attacks. The [6] uses the Backward Elimination technique for feature reduction using Chi-square. The independent seven features higher than 0.5 p value are removed from the UNSW-NB 15 data. The article [9] experiment obtained the accuracy value up to 90%. The research [10] proposed the ensemble artificial bee colony (En-ABC) on multiclass NSL-KDD dataset for anomaly detection. The model used for optimization using Unscented Kalman and Restricted Boltzmann Machine to filter non-linear properties of attributes. The system in validated on NSL KDD dataset and the model has reached the accuracy of 97.62%. The research has not been used the evolutionary algorithm and hybrid approaches to evaluate the performance.

The [11] model with information gain technique for feature selection using J48 with bagging on NSL-KDD dataset. The model performs the higher accuracy of 84.25% with 2.79 False alarm rate using 35 reduced features to detect the attacks. The research [12] model proposed the info gain(IG) and Correlation(CR) feature reduction method with ANN classifier on KDD Cup99 dataset. The system performed the better recall rate of 93.8% using 25 reduced feature to detect the DOS attacks. The suggested [13] model for feature reduction uses the combination of info gain(IG), Gain ratio(GR), ReflifF, and Chi-squared on NSL-KDD dataset.

The model produces the accuracy of 99.67% with lower False Alarm Rate(FAR) of 0.4% with 13 reduced feature using J48 classifier to detects the DoS types of attacks. The study [14] proposed the info gain(IG), Gain ratio(GR), and chi-squared feature selection method with J48 classifier on NSL-KDD set. The model produces the 99.98% of detection perecentage with low False Alarm Rate(FAR) of 0.12 % to detects the attacks. The proposed model [15] removes the feature that more than 0.5 correlation with other features. Also removes the feature variance are lower than 0.5 from the set. The model evaluated on logistic regression with ensemble approach. The system performed the accuracy of 99.5% with 0.6% lowest false alarm rate on NSL KDD dataset.

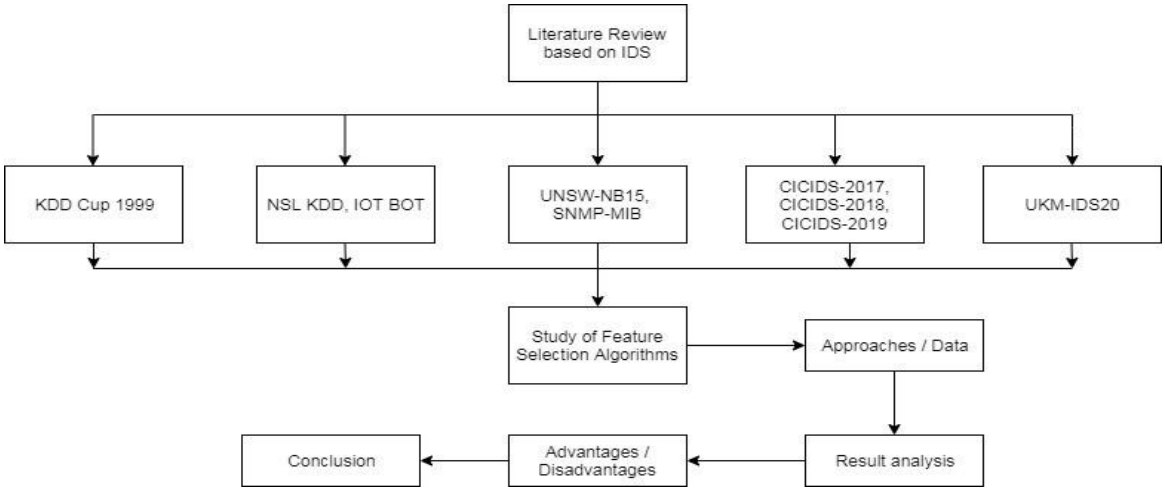


Figure 1: Proposed literature review flow

III. DATASET DESCRIPTION

The proficiency of IDS to detect the novel and various types of attacks through the validated datasets. Although, unavailability of various dataset because of privacy and its standard issues. Some network traffic datasets available publically for researcher such as NSL-KDDCup 1999 [16] NSL-KDD [17], CICIDoS 2017[18], UNSW-NB15[19], UKM-IDS20 [20] CICIDS 2018[21], CICIDS 2019[22], SNMP-MIB [23] dataset for development. The section focuses in brief on the benchmark network datasets.

Table1: Description of Datasets

Dataset	**Developed by**	**#of Features**	**# Instances**	**#Attacks**
NSL-KDD [24]	University of California	41	1,25,973	U2R, R2L, Probe, and DoS
ISCXIDS2012[25]	Canadian Institute of Cybersecurity (CIC)	546	24,50,324	DoS, DDoS, HTTP, and Secure Shell (SSH) bruteforce
UNSW-NB15[24]	Australian Centre for Cyber Security	47	175,341	DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms
SNMP-MIB(23)	Information Technology Department at Mu'tah University Jordan	34	4,998	DoS attacks and BruteForce attack
CICIDS 2017 [26]	Canadian Institute of Cybersecurity (CIC)	80	170,366	DoS, DDoS,Botnet, port scan and Web attacks
CICIDS 2018 [21]	Canadian Institute for Cybersecurity (CIC)	82	-----	Brute Force, DoS, DDoS,Botnet, and Web attacks.
CICIDS 2019 [27]	Canadian Institute for Cybersecurity (CIC)	82	172839	Reflected and ExploitedDDoS
UKM-IDS20 [20]	University of Kebangsaan Malaysia	46	12887	ARP Poisoning , Dos, Scan, Exploits Attacks

3.1 KDD 1999:

The dataset (Irvine 1998-99) updated from the DARPA98 by analysing the tcpdump section It includes many attacks, including buffer overflow, Neptune-DoS, pod-DoS, and Smurf-DoS [28]. (University of California, 2007). In a simulated setting, the attack and benign traffic are combined. The test results were skewed due to the significant number of redundant records and data corruptions in this dataset.

3.2 NSL KDD:

The NSL-KDD is enhanced version of KDD Cup99 and used by majority researcher for proposed IDS. The author stated [29] in article that the benchmark KDD99 or NSL-KDD dataset do not represent the new traffic behaviours with recent novel attacks.

3.3 BOTNET:

This dataset [30] has been developed by Elaheh Biglar Beigi et.al in University of New Brunswick, Canada.

3.4 ISCXIDS2012:

The ISCXIDS2012 [31] dataset developed in the (Canadian institute of cybersecurity) New Brunswick. The dataset generated and collected by Ali Shiravi et. Al and has the real network traffic of HTTP, S IMAP, MTP, FTP, POP3, and SSH protocol. The 21 network workstation connected for setting up the testbed architecture to identification of the vulnerabilities through various services and servers. It has collected from combination of the data generated from the seven days in June 2010. The dataset consists of 68,792 network attacks and 2,381,532 normal traffic instances.

3.5 SNMP-MIB:

The SNMP- MIB dataset is generated and collected by Alkasassbeh et.al in labs of Information Technology Department (Mu'tah University). The Author [23] generated and collected dataset of the Simple Network Management Protocol (SNMP) from the design test bed network. The SNMP- MIB has 34 features and 4998 total instances of attacks. The test bed environment captures the realistic management information (MIB) based attack traffic as well as normal traffic. The author uses the LAnTrafficV21 for normal traffic generation using UDP, TCP, SCTP, or ICMP protocols. The experiment used the THC –Hydra 5.22 tool for attack traffic generation protocol like FTP, HTTP, telnet and SSH etc. The HyenaeFE3 and HttpDosTol4.0 7 tool used for DoS, DDoS tack generation. Simultaneously the DOSHHTP4 2.5.1, Sloworis script5 and Activeperl language6 for software development.

3.6 UNSW-NB 15:

The [6] 42 features of UNSW-NB 15 dataset contain the normal and anomalous attacks. The UNSW-NB 15 dataset is an imbalanced set contains categorical, integer and binary format. The [32] dataset consists of Fuzzer, Backdoor, Dos, Analysis, Exploit, shellcode, Generic, Worms and Reconnaissance types of attacks. The dataset is in binary format where '0' represented the normal traffic and '1' denoted as attacks traffic for labelling the output. The pre-determined [27] divides UNSW train and UNSW test have 1,75,341 and 82,332 samples. Total 257673 samples taken for distribution in UNSW-NB 15train and test set.

3.7 CICIDS 2017:

he CICIDS2017 dataset consist of 170,366 instances of Benign and web types of attacks. The [26] performed the analysis by relabelling to the CICIDS 2017 dataset. The article concluded the imbalance ratio 0.001% of the dataset to detects the Heartbleed, sql injection, web attacks.

3.8 STA2018:

The STA2018 dataset was produced by converting the ISCX2012 network traffic into appropriate format for machine learning process. This has been [33] This dataset uses 193 basic features to describe each link, and a portion of Onut's feature classification schema applied to these features to make a complete of 550 features. The five feature were removed from the dataset because possibility the overfitting. It consists of 545 features including normal network traffic. The UNB ISCX 2012's network traffic was converted into a format that is appropriate for ML tasks to create the STA2018 dataset.

3.9 CICIDS 2018:

DDoS, Brute-force, Botnet, DoS, Web attacks, Heartbleed, and internal network infiltration are all seven [21] types of cyber-attacks. The organisation involves 420 computers, 30 servers and 50 machines that make up the attacking network environment. The dataset contains 80 features of benign and attack types that

CICFlowMeter-V3 extracted. These dataset [34] features shows backward directions, network flow and packet forward. The CICIDS-2018 dataset is bigger than CICIDS-2017 dataset in terms of size, coming in at over 400GB. The table1 of dataset description provides information about the total instances.

3.10 CICIDS 2019:

The selected [27] CICIDS 2019 dataset contains 172839 instances. This dataset [35] have various recent reflective attacks like portmap, Netbios, LDAP, MSSQL, UDP, SNMP, DNS, SYN, and DDoS. The benign and most recent DDoS assaults are included in CICDoS2019, which closely mirrors actual real-world statistics (PCAPs). The test bed architecture uses 25 users to build the behaviour abstract on email, SSH, FTP, HTPPS and HTTP protocols.

3.11 UKMIDS-20:

The UKM-IDS20 [20] dataset developed in University of Kebangsaan Malaysia. It has 46 features containing the four types attacks such as Scan, DoS, ARP Poisoning and exploits types of attacks. The network traffic generation and capturing through testbed configuration via Hyper-V for network simulation and Tshark for traffic analysation tools. The metasploit framework used for penetration testing and Nmap for scanning the network for malicious attacks. The UKM-IDS dataset consist of normal, (Mass HTTP, TCP flood, UDP data flood) of ARP Poisoning, Scans, and (BeFF HTTP exploits, Metasploit exploits) of exploits attacks. The dataset has 10308 training and 2579 testing set of instances.

Table 2: Dataset Instances information

Dataset	**Class**	**# Instances**	**#Total Instances**
NSL-KDD [24]	Benign	67,343	1,25,973
	DoS, R2L, U2R,Probe	58,630	
	Benign	56,000	175,341
UNSW-NB15 [24]	Fuzzer, Backdoor, Dos, Analysis, Exploit, shellcode, Generic, Worms, reconnaissance	119,241	
ISCXIDS2012[25]	Benign	2,381,532	24,50,324
	DoS, DDoS, HTTP, and Secure Shell (SSH) brute force	68,792	
KDD 99 [36]	Benign	12,24,608	4,898,431
	DoS, Remote-to-local, User-to-root, Probing	36,73,823	
CICIDS 2019 [35]	Benign	43938	7,97,532
	Port map	9348	
	LDAP	95260	
	MSSQL	285944	
	SYN	190313	
	NetBIOS	172729	
CICIDS 2018 [34]	Benign	13390249	1,61,37,183
	DdoS-LOIC-HTTP,UDP,HOIC	1263933	
	DoS- slowHTTPTest, Goldeneye,Slowloris	654300	
	Bruteforce- XSS, Web,FTP,SSH	381784	
	SQLi	87	
	Infiltration	160639	
	Bot	286191	
CICIDS 2017 Web Attack[8]	Benign	168,186	170,366
	XSS	652	
	SQLi	21	
	Brute Force	1507	
	Normal	8909	

UKM-IDS20[20]	ARP	592	12,887
	DoS	1742	
	Scan	597	
	Exploited	1047	

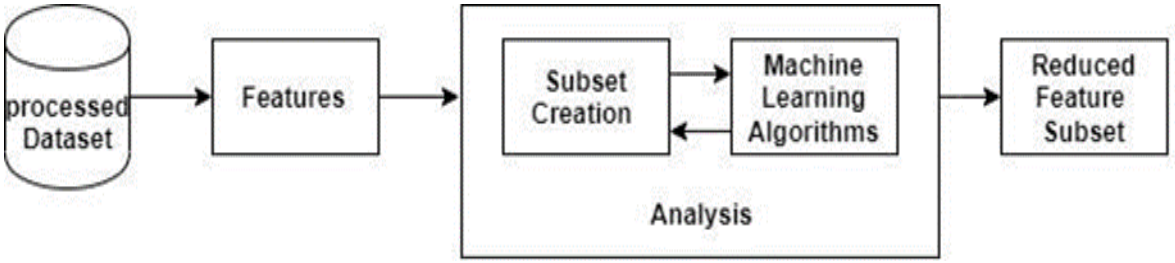
IV. FEATURE SELECTION METHODS

Feature selection or reduction is the process of choosing the finest subset of relevant variables while building the model. The ease of variable selection has been significantly researched in the beginnings of machine learning. Even though several techniques proposed to deal with the difficulty of variable selection the usage of a number of techniques, it's miles tough to pick out a particular technique because the maximum outfitted one regarding the feature subset selection issue. Overcoming the issue of complexity many and large feature selection helps to reduce the reduction from the many and large features. It reduces the overall data complexity issues of machine learning model. The techniques categorized into supervised and unsupervised feature selection.

4.1 Wrapper selection:

The wrapper feature selection method is primarily based totally on a selected machine learning set of rules that we're seeking to fit on a given dataset. The [37] method includes a selected learning algorithm so that it will be followed to assess the accuracy overall performance of a feature subset which result in higher solutions. However, they may be difficult to insert in the presence of big data as heavy computational burden required while using learning algorithms. The wrapper methods involve in training of machine learning methods with different combination of features so there is higher changes of overfitting by selecting right set of features.

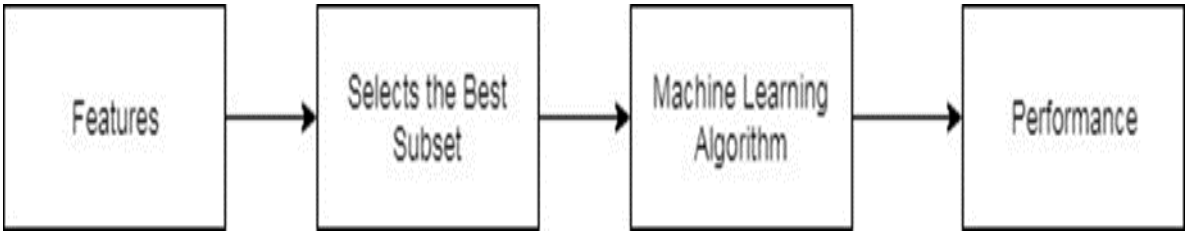
Figure 2: Wrapper feature selection



4.2 Filter Based:

The filter technique makes use of constitutive properties [38] along with separation or correlation from the dataset for sorting the features. The filter based feature selection qalgorithms are a few pertinent factors used to create a threshold. Features that do not meet the threshold criteria are omitted. The filter method and the wrapper method are equivalent. The classifier's independence from the filter technique is the only distinction, as the classifier is a component of the valuation phase for choosing the features. On the fundamental goal function, it chooses or discards characteristics.

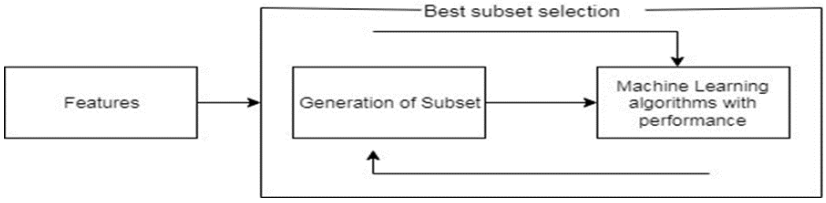
Figure 3: Filter based feature selection



4.3 Evolutionary/Embedded:

In general, Evolutionary feature selection method is a process of biological evolution, such as reproduction, mutation, recombination, and selection based on genetic algorithm imitate the natural process.

Figure 4: Evolutionary/Embedded based feature selection



The following table classify the variation between the wrapper and the evolutionary methods with an example of classification algorithms.

Table 2: Wrapper vs. Filter vs. Evolutionary Methods

Wrapper	**Filter**	**Evolutionary/Embedded**
Evaluates on the algorithm to get reduced subset of features	Generic set of methods to find reduced subset offeature	embed the feature during building process. By observing the iteration training phase of model
Measure usefulness based on classifier performance	Measure relevance based on univariate statistics on cross validation performance	an evolutionary algorithm is Natural evolution to optimise computationalprocess.
High computation cost	More faster in terms of time complexity	Fits in between wrapper and filter FS in terms of timecomplexity
Higher chances of overfitting by selecting right set of features	Less chances of over fitting	Reduced over fitting
e.g. Forward, stepwise selection, backward, recursive elimination. Genetic Algorithms (GA) etc.	e.g.Information Gain(GA), RelifF, Chi-square, Gain Ratio, correlation etc.	e.g. Combination of filter and wrapper.

V. FEATURE SELECTION ALGORITHMS

5.1 Information Gain (IG):

The algorithm chooses which parts of a training dataset from among a specified collection of characteristics are best for discriminating between the classes that need to be learnt [39]. This is based on information theory. The IG works with the help of using the uncertainty related to identifying the class attributes whilst the value of the features are unknown [13]. It is created on information theory that's utilized in rating and choosing top features to minimize the individual variable size.

The entropy [40] can be defined for variable X is as follows.

$$H(X) = -\sum_{i} P(x_i)\log_2(P(x_i)), \quad (1)$$

In eq. P (x i) represents the value of former possibilities of X.

$$H(X|Y) = -\sum_{j} P(y_j)\sum_{i} P(x_i|y_j)\log_2(P(x_i|y_j)), \quad (2)$$

The following entropy of X when analyzing value of variable Y is defined above:

$$IG(X|Y) = H(X) - H(X|Y). \quad (3)$$

P (x i |y j) is the posterior possibility of X given Y in. The IG is described as the quantity through which the entropy of X reductions to reflect an extra information approximately X provided through Y and is described as above: This

will be used to select essential features, if IG (X/Y) > IG (Z/Y) then that features of Y and X are correlated than the features of Y and Z.

5.2 Gain Ratio (GR):

The gain ratio [41] biases the decision tree towards thinking about attributes with a big number of distinct values. A change of the IG that solves the problem of bias toward features with a bigger set of values, exhibited through. It is the ratio among the Split of Intrinsic value and the information gain [42] as shown in example.

$$\text{Gain Ratio } (y,\ x) = \frac{\text{Information Gain}\,(y,\ x)}{\text{Intrinsic Value}(x)}, \tag{4}$$

Where,

$$\text{Intrinsic Value } (x) = -\sum \frac{|S_i|}{|S|} * \text{Log}_2 \frac{|S_i|}{S} \tag{5}$$

Where, |S| is the possible values of x, and |Si | is the actual values of x.

5.3 Chi-squared:

When the test statistic is chi-squared distributed under the null hypothesis, a chi- squared test (also known as a X2 test) is a viable statistical hypothesis test to do. The [43] individuality of features from the class is measured in feature selection by the x 2 statistic. Before calculating a score. This [44] can be defined as:

$$\chi^2\,(r,\ c_i) = \frac{N[P\,(r,\ c_i)\,P\,(\bar{r},\bar{c}_i) - P\,(r,\bar{c}_i)\,P\,(\bar{r},\ c_i)]^2}{P(r)P\,(\bar{r})\,P\,(c_i)\,P\,(\bar{c}_i)}, \tag{6}$$

where N stands for the complete dataset, r stands for a feature's existence (r¯ absence) and Ci stands for a class. P (r, Ci) is the likelihood that feature r appears in class c I while P (r, ci) is the likelihood that the feature r does not. P (r, c¯i) and P (r¯, c¯i) are additional probability that the characteristics exist or do not exist in a class that is not labelled Ci respectively. The chance that a feature will exist in the dataset is P(r), whereas the probability that it won't appear in the dataset is P (r¯). The probability of classifying a dataset as belonging to class Ci or not are P(ci) and P (c¯i).

5.4 RelifF:

Multiclass datasets with noisy data might be problematic, but the ReliefF technique can manage them. Based on Relief, ReliefF [45] uses a random sample of instances to determine its K nearest hits Hj through the similar class and its K nearest loses Mj(Class) from a separate class. For all features A, it adjusts the weight,

$$W[A_i] = W[A_i] - \sum_{j=1}^{K} diff\left(A_i, X_m, H_j\right)/(T \times K) + \sum_{C \neq Class(X_m)} \left[\frac{P(C)}{1-P(C)} \sum_{j=1}^{K} diff\left(A_i, X_m, \mathrm{M}_j(\mathrm{C})\right)\right]/(T \times K) \tag{7}$$

W[Ai] based on, Mj(Class), and Hj.

where P(C) is the class C prior probability. The number of iterations that may be made is T. The class of instance Xm is Class(Xm).

5.5 Symmetric Uncertainty (SU):

Although IG [46] is an effective filtering method, it has the drawback of favouring characteristics with higher values. This disadvantage also applies to mutual information. The entropy of features with class label are used to normalise mutual information. Given by, features fi and y's symmetric uncertainty Ent is the entropy used to categorise a dataset's observation. Contrary to the linear correlation approach, which can only forecast dependencies, symmetric uncertainty is able to predict both linear and non-linear relationships between the characteristics. Another benefit of SU is symmetric in nature, time complexity with the dataset's dimensionality, making it suitable for large datasets. Additionally, parallel processing allows for speedier computation.

$$SU(f_i, y) = \frac{2 \times MI(f_i, y)}{Ent(f_i) + Ent(y)} \quad (8)$$

VI. MACHINE LEARNING CLASSIFIERS

Weka: Weka [47] is well known mathematically proven developed by university of Waikato. It is an open-source tool to apply machine learning algorithms for binary classification with 10-fold cross validation for testing.

Figure 5: classification task.

Classification task y=f(x)

The train model is applied to categorize unknown data into normal classes or intrusion. As a model, the resulting classifier then predicts the class to which the input data may belong from a set of attributes. A usual method for utilizing classification algorithms is shown in Figure 5 classification of task. This article discusses many metrics that may be used to assess a classifier's success in terms of its capability to expect the appropriate class. There are various varieties of classification models. The following classifiers are discussed as follows.

- Boosting: AdaBoost, XGBoost, SMOTEBoost, Boosting, LightGBM, OSBoost, Gradient Tree.
- Rules based: RIPPER, CBA, Decision set, PART, OneR, ZeroR, DT, JRip, FURIA etc.
- Trees based Classifiers: J48, Decision trees, Random Forest, REP Tree, Decision Stump, etc.
- Bayes: Naïve Bayes, BayesNet, etc
- Functions: Logistic, Multi layer Perceptron , etc
- Meta: AdaBoost, Bagging, Vote, etc.

6.1 ANALYSIS OF EVALUATION MATRICES:

6.1.1 Evaluation Matrices:

The metrics mentioned below have been used to evaluate the outcomes. (1,2,3,4) The parameters used for prediction based on matrix.

- True Positives (TP).
- True Negatives (TN).
- False Positives (FP).
- False Negatives (FN).

6.1.2 Accuracy:

The percent of data that are correctly classified is known as accuracy. It is used to evaluate the performance of classification methods. The following is the expression for calculative accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

6.1.3 Precision and Recall (Sensitivity):

Precision is defined as the proportion of records correctly classified as positive out of the all numbers of positive records recognized. The recall [5] is the percentage of records correctly classified as positive from all number of sincerely positive records. These are suitable assessment metrics for classification problems with a wide range of class label density. Recall is well-known as detection rate (DR).

$$\text{Precision} = \frac{TP + TN}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP + TN}{TP + FP} \quad (3)$$

6.1.4 F1 score:

The F1 score is calculated using precision and accuracy. It is only high when both are high. As a result, it is objective. The precision [5] and recall are given equal weightage.

$$F1 = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

6.1.5 False Positive or alarm rate:

It is also known as false positive rate. It is equal to the number of incorrect positive predictions divided by the dataset's true negative values.

$$FAR = \frac{FP}{TN + FN} \quad (5)$$

6.1.6 Incorrectly instances (ICI):

$$\text{Error rate} = \frac{FP + FN}{TP + TN + FP + FN} \quad (6)$$

It is equal to number of incorrectly classified instances divided by all prediction instances.

Table 3: Analysis of feature selection methodologies

Ref. No.	In Year	Dataset	Classifier	Method	Accuracy/ observation	Features
[48]	2019	CICIDS- 2017	J48	deleting one feature at a time (wrapper)	96.36%	15
[49]	2021	IoT-BoT dataset	JRip	IG, CR, GR, CHI, SU	99.9994%	36
[50]	2020	UNSW- NB15	DNN	Split 70% Training 30% Testing	99.2 %	8
[51]	2017	SNMP-MIB	Bayes Net	Info gain, RelifF	99.80%	34
[52]	2018	KDD Cup 99	FURIA	CFS and BFS	99.58%	11
[53]	2021	KDD Cup 99	J48	Combination of IG CR	99.9964%	16
[53]	2021	CICIDS 2017	J48	Combination of IG CR	99.9569% Portmap	36
[20]	2021	UKM-IDS20	HOE-DANN	RS-DCFA	96.46%	NA
[54]	2022	NSL-KDD	DT, RF, XGBoost	selectKbest	99.9%	20

[55]	2022	NSL-KDD	Random Forest	K-NN in python	99.5%	NA
[56]	2020	CICIDS 2017	J48	Info Gain (IG)	99.86%,	22
[56]	2020	CICIDS 2017	RF	Info Gain (IG)	99.87%	52
[57]	2021	CICIDS 2019	DNN	Removed manually having 'NaN' Values	99.97%	69
[58]	2020	CICIDS 2017	Bayesian-rough set	Feature Probability Estimation (FPE)	TPR of 56.34%,	40
[59]	2019	NSL-KDD	Random Tree	SU-Genetic algorithm	99.8109% DR	17
[60]	2020	Hping3DDoS	SVM KNN, ANN, and NB	ReliefF, Sequential forwardfloatingselection, Lasso	98.30%	6
[61]	2019	CICIDS DDoS 2017	AdaBoost	SMOTE With R	81.83%	25

VII. CHALLENGES IN SELECTION OF METHOD AND FUTURE DIRECTION

Based on the lesson learnt and finding above, the following are some challenges and future direction for further research. Most of the researchers use the NSL- KDD, KDD Cup 199, UNSW-NB, CICIDS 2018 and CICIDS 2017 dataset. These datasets have generated the old network traffic. It has been suggested to use the latest network traffics generated data for experimental research to invent the finest suitable model methodologies for feature selection.

7.1 Dataset Selection:

The selection of the best dataset is challenge to researcher for developing effective machine learning based IDS. The CAIDA 2007, NSL-KDD, KDD Cup 99, UNSW-NB15 data are inadequate and consist of old network traffic and attack behavior. So, obtaining the dynamically generated recent network traffic can solve this problem from recent datasets like CICIDS 2019, UKM-IDS20 for effective Network intrusion detection system. These datasets contain the latest types of reflected and exploited attacks.

Table 4: dataset classification

Dataset	**year**	**Packet capture**	**Real traffic**	**Labeled**	**IoT set**
KDD Cup	1999	Yes	Yes	Yes	No
NSL KDD	2009	Yes	Yes	Yes	No
ISCXIDS	2012	Yes	Yes	Yes	No
Botnet	2014	Yes	Yes	Yes	Yes
UNSW-NB	2015	Yes	Yes	Yes	No
SNMP-MIB	2016	Yes	Yes	Yes	No
CICIDS 17	2017	Yes	Yes	Yes	No
STA2018	2018	Yes	Yes	Yes	No
CICIDS 18	2018	Yes	Yes	Yes	No
CICIDS 19	2019	Yes	Yes	Yes	No
UKMIDS20	2020	Yes	Yes	Yes	No

7.2 Feature selection:

The selection of filter-based feature selection method is based on machine learning algorithms. The approach explored by kshirsagar et.al [8] is obtain feature subsets that perform threshold for detection of web attacks. However, ensemble approaches need more study to get promising performance.

A. Classifier Choice

The classification algorithm is to be selected from the suit of the different types of available classifiers. The tree-based classifiers J48 [8] given better performance on intrusion detection model proposed by the researcher on network traffic dataset. The high configuration machine will help to build the model in a minimum time frame.

B. Detection of attacks:

Most of the work is on the specific TCP, HTTP, UDP based attacks for the detection of network and transport layer attacks. There is a need to include latest types of vulnerability present in the network traffic for efficient intrusion detection. The system Selects the dynamically generated dataset for research would detects the recent types of reflected, exploited, ARP, Scan and DDoS attacks present in network traffic.

C. Scalability:

The article [62] presented about the Nature-inspired Algorithms (NIA) uses to solves the feature selection problem. The study investigated the 34 different operators and found that chaotic maps are more popular operators. whereas integration of Nature-inspired Algorithms (NIA) with a classifier is the most widely used technique for feature selection i.e. Hybridization. The research addressed about the scalability gap of datasets and better feature selection technique will enhance d the performance of binary dataset.

D. Complexity

The system's performance suffers these numerous characteristics, and the network is put under more strain. Identification of characteristics that are significant or irrelevant in IDS is crucial. The system's performance is enhanced with the lowest build-up time through the selection of essential network traffic. The selected numerous characteristics and set of number features reduces the complexity of various types of network layer, transport layer,

presentation layer-based attacks. The feature selection techniques are important steps in data analytics process, that reduce the complexity and the generation time of the model.

CONCLUSION

The ensemble technique with threshold approach presented a better performance on machine learning algorithms to detect the different types of novel attacks. The feature selection technique suggestion is based on the set of input variables or machine learning algorithms. Instead, it is on your specific problem using systematic experiments on the set of variables. It is up to different available models that fits to different subset of features selected through different measures that fits to your specific problem.

The paper underlined the importance of feature selection and classification. Whereas discussing the challenges about complexity of dataset, stability and scalability issue, the feature selection with efficient classification algorithms will overcome the problems defined. However, the section of method depends on the researcher who combines with innovative approaches to get the finest method for the specific objective. Try different models that fit to subset of selected features through various statistical measures.

In future, Use the hybrid approach with combination of the feature selection technique with machine learning algorithms for reduced the feature up to minimum subset to perform the superior model for intrusion detection.

REFERENCES

[1] What is a cyber-attack? Recent examples cyber-attacks|Unisystem Online. https://www.unisys.com/glossary/cyber-attack/example.html. Accessed 29 April, 2023

[2] Cybersecurity Prediction and statistics for 2021 to 2025. https://cybersecurityventures.com /top-5-cybersecurity-facts-figures-predictions-and-statistics- for-2021-to-2025/. Accessed on 29 April, 2023.

[3] Tsai CF, Hsu YF, Lin CY, Lin WY (2009) Intrusion detection by machine learning: A review. Expert Syst. Appl. 36:11994–12000.

[4] S. Sen, K. D. Gupta, and M. M. Ahsan, “Leveraging machine learning approach to setup software-defined network (SDN) controller rules during DDoS attack,” in Proceedings of International Joint Conference on Computational Intelligence, 2020, pp. 49–60.

[5] M. I. Kareem and M. N. Jasim, "DDOS Attack Detection Using Lightweight Partial Decision Tree algorithm," 2022 International Conference on Computer Science and Software Engineering (CSASE), 2022, pp. 362-367, doi: 10.1109/CSASE51777.2022.9759824.

[6] R. A. Disha and S. Waheed, "A Comparative study of machine learning models for Network Intrusion Detection System using UNSW-NB 15 dataset," 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), 2021,pp.1-5, doi:10.1109/ICECIT54077.2021. 9641471.

[7] Pawar, K., Mohite, B., & Kshirsagar, P. (2022). Analysis of Feature Selection Methods for UKM- IDS20 Dataset. In International Conference on Computing in Engineering & Technology (pp.461-467). Springer, Singapore.

[8] Kshirsagar, D., & Kumar, S. (2022). Towards an intrusion detection system for detecting web attacks based on an ensemble of filter feature selection techniques. Cyber-Physical Systems, 1- 16.

[9] De la Hoz, E., De La Hoz, E., Ortiz, A., Ortega, J., & Prieto, B. (2015). PCA filtering and probabilistic SOM for network intrusion detection. Neurocomputing, 164, 71-81.

[10] Garg, S., Kaur, K., Batra, S., Aujla, G. S., Morgan, G., Kumar, N., & Ranjan, R. (2020). En-ABC: An ensemble artificial bee colony based anomaly detection scheme for cloud environment. Journal of Parallel and Distributed Computing, 135, 219-233, (2020).

[11] N. T. Pham, E. Foo, S. Suriadi, H. Jeffrey, and H. Lahza, "Improving performance of intrusion detection system using ensemble methods and feature selection," In Proceedings of theAustralasian Computer Science Week Multiconference, pp. 1-6. 2018

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

[13] O. Osanaiye, H. Cai, K. Choo, A. Dehghantanha, Z. Xu, and M. Dlodlo, "Ensemble-based multi- filter feature selection method for DDoS detection in cloud computing," EURASIP Journal on Wireless Communications and Networking, vol. 2016, no. 1,pp. 130, 2016

[14] Bharot, Nitesh, et al. "Distributed denial-of-service attack detection and mitigation using feature selection and intensive care request processing unit." Arabian Journal for Science and Engineering 43 (2018): 959-967.

[15] Pranto, Md Badiuzzaman & Ratul, Md. Hasibul & Rahman, Md & Jahan, Ishrat & Zahir, Zunayeed-Bin. (2022). Performance of Machine Learning Techniques in Anomaly Detection with Basic Feature Selection Strategy - A Network Intrusion Detection System. 13. 36-4436. 10.12720/jait.13.1.36-44.

[16] KDD Cup 1999 Data. http://kdd.ics.uci.edu/databases /kddcup99/kddcup99.html. Accessed on 29 April, 2023. 2.30 PM.

[17] NSL-KDD | Datasets | Research | Canadian Institute for Cybersecurity | UNB. https://www.unb.ca /cic / datasets/nsl.html. Accessed on 24 April, 2023. 10.15 AM.

[18] CICIDoS 2017 Datasets | Research | Canadian Institute for Cybersecurity |UNB. https://www.unb.ca /cic /datasets/ dos-dataset.html (accessed April. 29, 2023) 12.30 PM.

[19] The UNSW-NB15 data set. https://www.unsw.adfa.edu. au/unsw- canberra-Cyber /cybersecurity /ADFA-NB15-Datasets/. Accessed on 24 April, 2023. 3.00 PM.

[20] Al-Daweri, Muataz Salam, Salwani Abdullah, and Khairul Akram Zainol Ariffin. "An adaptive method and a new dataset, UKM-IDS20, for the network intrusion detection system." Computer Communications 180 (2021): 57-76.

[21] CICIDoS 2018 Datasets | Research | Canadian Institute for Cybersecurity |UNB https:/www.unb.ca /cic/datasets/ids-2018.html (accessed April. 30, 2023) 04.30 PM.

[22] CICIDoS 2019 Datasets | Research | Canadian Institute for Cybersecurity |UNB. https://www.unb.ca/cic/datasets/ddos-2019.html (accessed April. 29, 2023) 04.30 PM.

[23] Alkasassbeh, Mouhammd & Al-Naymat, Ghazi & Hawari, Eshraq. (2016). Towards Generating Realistic SNMP-MIB Dataset for Network Anomaly Detection. International Journal of Computer Science and Information Security ISSN 1947 5500. Vol. 14. (pp. 1162-1185).

[24] Rashid, M., Kamruzzaman, J., Imam, T., Wibowo, S., & Gordon, S. (2022). A tree-based stacking ensemble technique with feature selection for network intrusion detection. Applied Intelligence, 1-14.

[25] Kanna, P. Rajesh, and P. Santhi. "Unified deep learning approach for efficient intrusion detection system using integrated spatial–temporal features." Knowledge-Based Systems 226 (2021): 107132.

[26] Panigrahi R., Borah S. "A detailed analysis of CICIDS dataset for designing intrusion detection systems" Int. J. Eng. Technol., pp. 479-482, 7 (2018).

[27] S. Rajagopal, P. P. Kundapur and H. K. S., "Towards Effective Network Intrusion Detection: From Concept to Creation on Azure Cloud," in IEEE Access, vol. 9, pp. 19723-19742, 2021, doi:10.1109/ACCESS.2021.3054688.

[28] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. ICISSp, 1, 108-116.

[29] Aldweesh, A., Derhab, A., & Emam, A. Z. (2020). Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. Knowledge-Based Systems, 189, 105124.

[30] E. Biglar Beigi, H. Hadian Jazi, N. Stakhanova and A. A. Ghorbani, "Towards effective feature selection in machine learning-based botnet detection approaches," 2014 IEEE Conference on Communications and Network Security, 2014, pp. 247-255, doi: 10.1109/CNS.2014.6997492.

[31] Shiravi, Ali, et al. "Toward developing a systematic approach to generate benchmark datasets for intrusion detection." computers & security 31.3 (2012): 357-374.

[32] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." In 2015 military communications and information systems conference (MilCIS), pp. 1-6. IEEE, 2015.

[33] Al Tobi, Amjad M., and Ishbel Duncan. "Improving intrusion detection model prediction by threshold adaptation." Information 10.5 (2019): 159.

[34] Kim, J., Shin, Y., & Choi, E. (2019). An intrusion detection model based on a convolutional neural network. Journal of Multimedia Information System, 6(4), 165-172.

[35] Sharafaldin, I., Habibi Lashkari, A., Hakak, S., & Ghorbani, A.A. (2019). Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy. 2019 International Carnahan Conference on Security Technology (ICCST), 1-8.

[36] Al-Zewairi, Malek, Sufyan Almajali, and Moussa Ayyash. "Unknown security attack detection using shallow and deep ANN classifiers." Electronics 9.12 (2020): 2006.

[37] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," 2016 International Conference on Engineering & MIS (ICEMIS),2016, pp. 1-5, doi:10.1109/ICEMIS.2016. 7745366.

[38] Cateni S, Colla V, Vannucci M (2017) A fuzzy system for combining filter features selection methods. Int Journal of Fuzzy System19(4):1168–1180.

[39] Singh, Khundrakpam Johnson and De, Tanmay. "Efficient Classification of DDoS Attacks Using an Ensemble Feature Selection Algorithm" Journal of Intelligent Systems, vol. 29, no. 1, 2020, pp. 71-83. https://doi.org/10.1515/jisys-2017-0472

[40] L Yu, H Liu, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003). Feature selection for high-dimensional data: A fast correlation-based filter solution (Springer, Washington DC, 2003), pp. 856–863.

[41] Ibrahim, H. E., Badr, S. M., & Shaheen, M. A. (2012). Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems. arXiv preprint arXiv:1210.7650.

[42] Priyadarsini, R. P., Valarmathi, M. L., & Sivakumari, S. (2011). Gain ratio based feature selection method for privacy preservation. ICTACT Journal on soft computing, 1(4), 201-205.

[43] L Devi, P Subathra, P Kumar, Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO-2015). Tweet sentiment classification using an ensemble of machine learning supervised classifiers employing statistical feature selection methods (Springer, 2015), pp. 1–13.

[44] Nissim, N., Moskovitch, R., Rokach, L., & Elovici, Y. (2012). Detecting unknown computer worm activity via support vector machines and active learning. Pattern Analysis and Applications, 15(4), 459-475.

[45] Cui, X., Li, Y., Fan, J., & Wang, T. (2022). A novel filter feature selection algorithm based on relief. Applied Intelligence, 52(5), 5063-5081.

[46] Nagarajan, G., & Babu, L. D. (2022). Missing data imputation on biomedical data using deeply learned clustering and L2 regularized regression based on symmetric uncertainty. Artificial Intelligence in Medicine, 123, 102214.

[47] Weka tool, March 2023, [online] Available: http://www.cs.waikato.ac.nz/ml/weka/.

[48] R. B. Adhao and V. K. Pachghare, "Performance-Based Feature Selection Using De-cision Tree," 2019 International Conference on Innovative Trends and Advances in Engineering and Technology (ICITAET), 2019, pp. 135-138, doi: 10.1109/ICITAET47105.2019.9170235.(2019).

[49] Nimbalkar P., Kshirsagar D. "Analysis of Rule-Based Classifiers for IDS in IoT". In: Shukla S., Unal A., Varghese Kureethara J., Mishra D.K., Han D.S. (eds) Da-ta Science and Security. Lecture Notes in Networks and Systems, vol 290. Sprin-ger, Singapore. https://doi.org/ 10.1007/978-981-16-4486-3_51(2021).

[50] Choudhary S, Nishtha K "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 datasets using deep learning in IoT". Procedia ComputSci 167:1561–1573 (2020).

[51] Alkasassbeh, Mouhammd. "An empirical evaluation for the intrusion detection fea-tures based on machine learning and feature selection methods." arXiv preprint ar-Xiv:1712.09623 (2017).

[52] Gündüz, S. Y., & ÇETER, M. N. (2018). Feature selection and comparison of classification algorithms for intrusion detection. Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering, 19(1), 206-218.

[53] Kshirsagar, D., Kumar, S. "A feature reduction based reflected and exploited DDoS attacks detection system." J Ambient Intell Human Comput (2021).

[54] Rashid, M., Kamruzzaman, J., Imam, T., Wibowo, S., & Gordon, S. (2022). A tree-based stacking ensemble technique with feature selection for network intrusion detection. Applied Intelligence, 1-14.

[55] Pranto, M. B., Ratul, M. H. A., Rahman, M. M., Diya, I. J., &Zahir, Z. B. (2022). Performance of Machine Learning Techniques in Anomaly Detection with Basic Feature Selection Strategy- A Network Intrusion Detection System. Journal of Advances in Information Technology Vol, 13(1).

[56] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Idris, A. M. Bamhdi and R. Budiarto, "CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection," in IEEE Access, vol. 8, pp. 132911-132921, 2020, doi: 10.1109/ACCESS.2020.3009843.

[57] Cil, A. E., Yildiz, K., &Buldu, A. (2021). Detection of DDoS attacks with feed forward based deep neural network model. Expert Systems with Applications, 169, 114520.

[58] M. Prasad, S. Tripathi, and K. Dahal, "An efficient feature selection based Bayesian and Rough set approach for intrusion detection," Applied Soft Computing, vol. 87, pp. 105980, 2020.

[59] C. Wang, H. Yao, and Z. Liu, "An efficient DDoS detection based on SU-Genetic feature selection," Cluster Computing, vol. 22, no. 1, pp. 2505-2515, 2019.

[60] H. Polat, O. Polat, and A. Cetin, "Detecting DDoS Attacks in Software-Defined Networks Through Feature Selection Methods and Machine Learning Models," Sustainability, vol. 12, no. 3,pp. 1035, 2020.

[61] A. Yulianto, P. Sukarno, and N. Suwastika, "Improving adaboost-based intrusion detection system (IDS) performance on CIC IDS 2017 dataset," In Journal of Physics: Conference Series, vol. 1192, no. 1, p. 012018, 2019.

[62] Pawar, Kiran S. & Mohite, Babasaheb J. (2024) Performance analysis of UKM-IDS20 dataset on machine learning algorithms, Journal of Statistics and Management Systems , 27:5, 997–1008, DOI: 10.47974/JSMS-1296.

[63] Abu Khurma R, Aljarah I, Sharieh A, Abd Elaziz M, Damaševičius R, Krilavičius T. A Review of the Modification Strategies of the Nature Inspired Algorithms for Feature Selection Problem. Mathematics. 2022;10(3):464.https://doi.org/10.3390/ math100304 64.