JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

Analysis of Features Selection Effects on Different Classification Algorithms with Performance Metrics Improvement based on PortScan-attack of CICDDoS2019 Dataset

Zainab A. Ibrahim [1], Imad J. Mohammed [2]

[1, 2] Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

zainabaziz889@gmail.com , Emad.J@sc.uobaghdad.edu.iq

Received 2022 April 02**; Revised** 2022 May 20**; Accepted** 2022 June 18**.**

Abstract

DDoS attack is a type of network security threat that aims to flood target networks with harmful traffic. Despite the fact that several statistical methods have been designed for DDoS attack detection, creating an offline or real-time detector with low cost is still one of the main concerns. The aim of this research using (ML) techniques is to categorize data traffic as either normal or malicious. This paper handles a general "PortScan attack CICDDOS2019 Data set" there are 79 attributes in total, and more than 200 thousand records, this dataset contains the normal and attack traffics. This study implements boruta, forward and backward, and variable significance algorithms using the RStudio tool to detect the most relevant attributes through selection feature and perform classification effectively. After the preprocessing and feature selection phases, the obtained dataset was classified by Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM) algorithms. The experimental results show that (RF) with forward and backward has a higher rate of accuracy than other algorithms 100%, Precision 0.99993, Recall 100%, F1-Measure 0.99994, Specificity 0.9999 with classification achievement.

1- Introduction

In almost every scientific discipline, the concept of dataset spreads, as data is the basis for research activities. Although the term dataset is frequently used in papers, reports, and articles, no precise definition for it [1]. Some datasets contain a large number of features, some of which are irrelevant. A feature selection approach is used to address this problem.

The most proper features can be chosen by applying feature selection algorithms. These algorithms can refine the prediction outcomes. Yet, the feature selection algorithms are best to extract the pertinent features and shun redundancy. So, it is favorable to use feature selection algorithms to shun the loss of significant data [2].

DoS attacks consume the target system's computing resources and network bandwidth by flooding it with malicious traffic, preventing it from providing regular services to rightful users. On a bigger scale, DDoS goes even further. DDoS attacks seize control of a large number of vulnerable systems [3]. Such as PortScan is the Internet hostile attack through open (ports) through which hackers gain access to computers [4], as illustrated in Figure 1.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

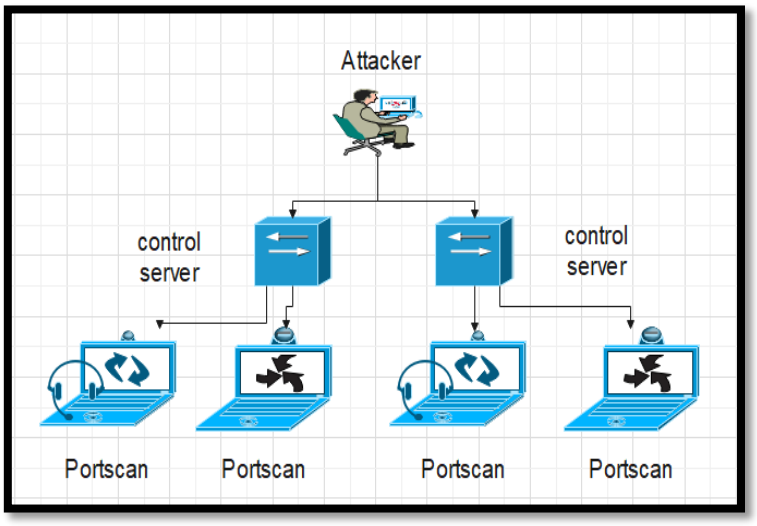

Figure (1): DDoS Attack with Portscan

2- Related Work

In this study [5], here, four datasets were used (ISCXIDS2012, KDDCUP99, CICIDS2017, and CICDDOS2019). The normalization procedure pre-processes the input data. Then, Use the features selected to minimize the dimensions. The dataset is categorized by a Support value-based graph. Finally, the data is categorized into a normal classification or intrusion. The results show that the classification used is superior to (NB), (SVM), and (RF) classifiers. The performance measures that were used are Accuracy, Precision, Sensitivity, Recall, Specificity, FPR, F1-measure, Kappa, FNR, and Rank sum measure.

In [6], a logistic regression classifier was applied to the two datasets. The first dataset consisted of a Portmap attack where binomial logistic regression was used and the second dataset consisted of LDAP and NetBIOS variant of a DDoS attack where polynomial logistic regression method was used to classify these two variables with normal data. Feature selection was performed on the dataset, correlation test was performed for each feature with the detection label to check the relevance. Portmap attack detection accuracy of 99.91% with an f1 score of 0.9913 while LDAP and NetBIOS attack detection accuracy of 99.94% with an f1 score of 0.9847.

In [7], Stacking-based, bagging-based, and boosting-based ensemble learning techniques are used at work. The CICDDoS2019 benchmark was chosen for the examination. Preprocessing data entails a number of stages. Two feature selection procedures are utilized to determine the most significant features: tree-based and Pearson's correlation coefficient. Performance is measured by the likelihood of detection, the likelihood of a false alarm, the likelihood of miss detection, and accuracy. According to the findings, ensemble learning stacking-based techniques outperform the other algorithms on all four assessment metrics.

In [8], two situations were used: in the first, IP flows were collected from SDN Floodlight controllers using Mininet emulation. The CICDDoS2019 dataset was used in the second scenario. The performance of the system is compared to that of other methods for detecting a DDoS attack and a Portscan attack in the first scenario. The methods used for classification in DL are LSTM_FUZZY and compared with ML methods are KNN, SVM, MLP, POS-DS and the DL method is LSTM-2. The performance measures used are: False-Positive rate, Recall, and Precision, the best result in LSTM_FUZZY, Recall 99.87, and Precision 99.74

In [9], the CICIDS2017 dataset was used, the proposed technique shows the detection of portscan attempts using deep learning and (SVM) algorithms, and the performance measures that were used are Recall, F1-score, Precision, and Accuracy, where the results showed the deep learning algorithm, it achieved better results with an accuracy of 97.80%.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

In [10], KDD 99 dataset was used, the suggested technique covers categorization of portscan attacks using the SVM algorithm, the Consistency Subset Evaluation algorithm, and the Best First search method were used to reduce features, with performance metrics of Acc, TPR, FPR, Pr, Rc, and computation time, and a rating accuracy of 99.9185%.

3- Proposed Methodology

When using the CICDDoS2019 dataset, it was found that it was not preprocessed, so the data was processed by removing irrelevant attributes to facilitate work performance, and the number of features was reduced using feature selection to reduce performance time and obtain better results. Figure 2 shows the steps of the methodology used in this research:

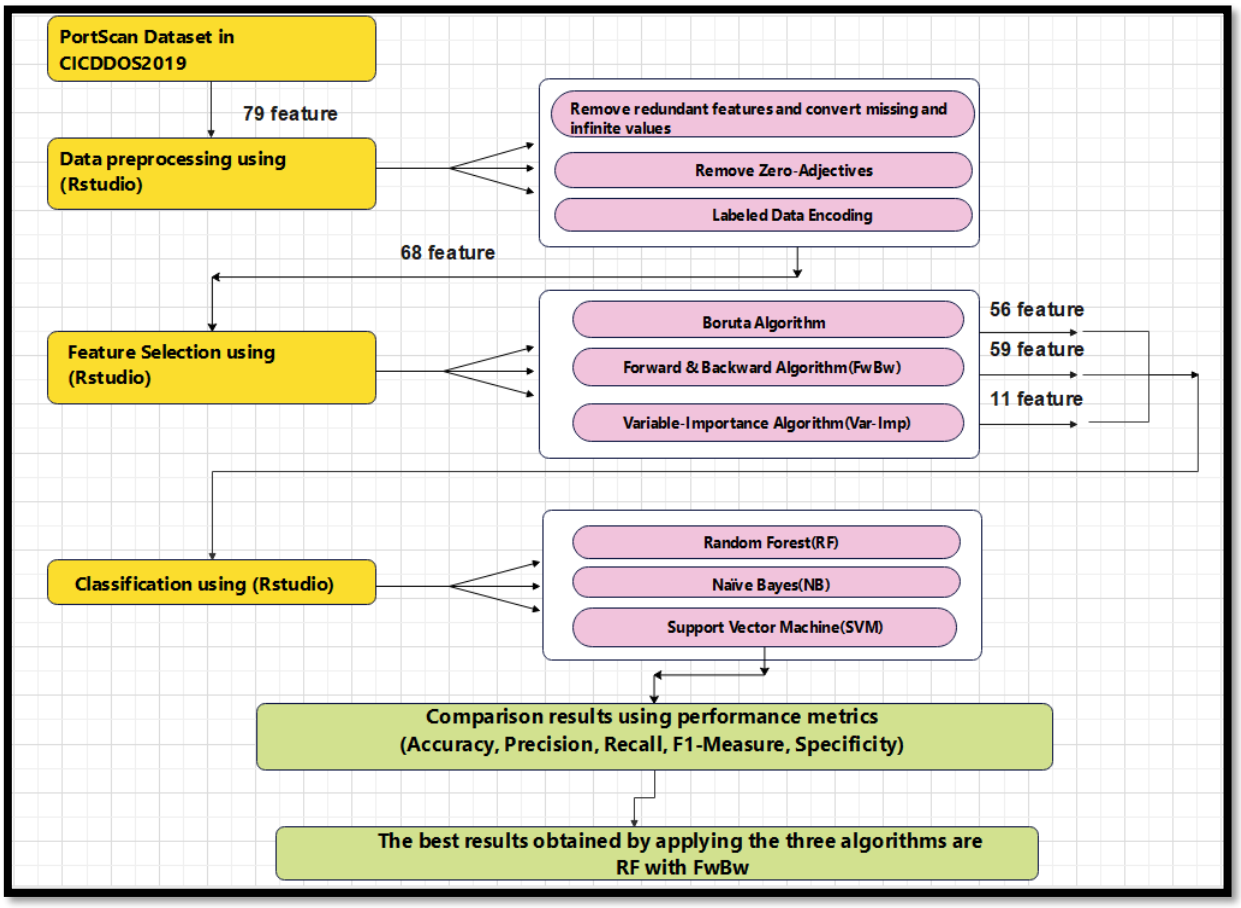

Figure (2): PortScan attack detection framework

4- Data and Methodology

4-1 CICDDoS2019 Dataset

CICDDoS2019 uses CICflowmeter-V3 with flows labeled based on the destination and source IPs, time stamp, protocols, destination, source ports, and attack, to include outcome analysis of network traffic, containing the most recent popular DDoS attacks. The attacks in this dataset include 12 attacks on the training day DNS, NTP, SSDP, MSSQL, LDAP, SNMP, NetBIOS, UDP-Lag, UDP, WebDDoS, TFTP, SYN, and 7 attacks during the day of the test, including LDAP, UDP, PortScan, NetBIOS, MSSQL, SYN, and UDP-Lag. (Dataset is publicly available at https://www.unb.ca/cic/datasets/ddos-2019.html). Over 80 traffic features were retrieved from raw data using the CICflowmeter-V3 and saved as a CSV file per machine. The capturing period for the training day on 2019 January 12th, which began at 10:30 a.m. and concluded at 17:15 p.m., and the testing day on 2019 March 11th, which began at 09:40

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452
a.m. and ended at 17:35 p.m. [11]. In the current work, the dataset is the PortScan traffic described in Table 1. The used dataset from CICDDOS2019 contains 79 features. The definition of extracted features is available in Table 2.

Table (1): the details of the Portscan Dataset

Dataset Name	CICDDoS2019
CSV File Used	Friday-WorkingHours-Afternoon-PortScan
Year Of Release	2019
Total Number Of Instances	286468
Number Of Attributes Used in This Paper	79
Number Of Class	2 class (Bening & PortScan)

Table (2): the explanation names of features in the CICDDOS2019 Dataset

No	Feature Name	No	Feature Name	No	Feature Name
1	Destination Port	28	Bwd IAT Std	55	Avg Bwd Segment Size
2	Flow Duration	29	Bwd IAT Max	56	Fwd Header Length
3	Total Fwd Packets	30	Bwd IAT Min	57	Fwd Avg Bytes/Bulk
4	Total Backward Packets	31	Fwd PSH Flags	58	Fwd Avg Packets/Bulk
5	Total Length of Fwd Packets	32	Bwd PSH Flags	59	Fwd Avg Bulk Rate
6	Total Length of Bwd Packets	33	Fwd URG Flags	60	Bwd Avg Bytes/Bulk
7	Fwd Packet Length Max	34	Bwd URG Flags	61	Bwd Avg Packets/Bulk
8	Fwd Packet Length Min	35	Fwd Header Length	62	Bwd Avg Bulk Rate
9	Fwd Packet Length Mean	36	Bwd Header Length	63	Subflow Fwd Packets
10	Fwd Packet Length Std	37	Fwd Packets/s	64	Subflow Fwd Bytes
11	Bwd Packet Length Max	38	Bwd Packets/s	65	Subflow Bwd Packets
12	Bwd Packet Length Min	39	Min Packet Length	66	Subflow Bwd Bytes
13	Bwd Packet Length Mean	40	Max Packet Length	67	Init_Win_bytes_forward
14	Bwd Packet Length Std	41	Packet Length Mean	68	Init_Win_bytes_backward
15	Flow Bytes/s	42	Packet Length Std	69	act_data_pkt_fwd
16	Flow Packets/s	43	Packet Length Variance	70	min_seg size_forward
17	Flow IAT Mean	44	FIN Flag Count	71	Active Mean
18	Flow IAT Std	45	SYN Flag Count	72	Active Std
19	Flow IAT Max	46	RST Flag Count	73	Active Max
20	Flow IAT Min	47	PSH Flag Count	74	Active Min
21	Fwd IAT Total	48	ACK Flag Count	75	Idle Mean
22	Fwd IAT Mean	49	URG Flag Count	76	Idle Std
23	Fwd IAT Std	50	CWE Flag Count	77	Idle Max
24	Fwd IAT Max	51	ECE Flag Count	78	Idle Min
25	Fwd IAT Min	52	Down/Up Ratio	79	Label
26	Bwd IAT Total	53	Average Packet Size		
27	Bwd IAT Mean	54	Avg Fwd Segment Size		

4-2 Data Preprocessing

Preprocessing is a procedure for transforming raw data into usable information. The data is cleaned to remove any poor data, filter out any inaccurate data, and reduce unnecessary data specifics [12]. By applying preprocessing to the source datasets, the first step transforms raw data into a format appropriate for analysis [13].

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

4-2-1 convert missing and infinite values

The CICDDOS2019 dataset contains 79 attributes, this data contains anomalies, which will be replaced by different values, such as missing values that can be replaced with their lowest attribute values and infinite values with their highest attribute values that can be resolved. In the PortScan dataset, for example, 'FlowBytes' has outliers like 'Infinity' and 'NaN', but 'Flowpackets' only have outliers like 'Infinity' [14].

4-2-2 Remove Zero-Adjectives

An attribute with a value of zero for all records is known as a zero- adjective. When the dataset attribute's minimum and maximum values are both zero, it is termed a zero- adjectives type. The examination of the CICDDOS2019 dataset revealed that it has 10 zero- adjectives with the same value for all entries, as shown in the suggested work, These attributes are: (32Bwd PSH Flags, 33Fwd URG Flags, 34Bwd URG Flags, 50CWE Flag Count, 57Fwd Avg Bytes/Bulk, 58Fwd Avg Packets/Bulk, 59Fwd Avg Bulk Rate, 60Bwd Avg Bytes/Bulk, 61Bwd Avg Packets/Bulk, 62Bwd Avg Bulk Rate). As a result, deleting the zero- adjectives is improving the model's accuracy [15].

4-2-3 Labeled Data Encoding

Finally, CICDDOS2019 contains nominal attributes. Data labels are replaced from nominal attributes with numerical values, because of the impossibility to train machine learning models with nominal data before any process [16]. We trained the model to classify input traffic into two categories ("Benign" and "PortScan"), with the benign 0 and PortScan 1 labels encoding the nominal value.

4-3 Feature Selection Methods

In a huge data set, feature selection is critical in detecting unnecessary and redundant features. It is a widely used preprocessing procedure for huge amounts of data [17]. The main area of knowledge discovery, pattern recognition, and statistical science is feature selection. The point of feature selection is to get rid of non-essential inputs. In addition, the accuracy of the classification algorithms depends on the use of appropriate feature selection algorithms to reduce the dimensions of the dataset [18]. There are three basic ways to feature selection, such as filtering, wrapping, and embedding methods. This work uses the wrapper method.

4-4-1 Boruta Algorithm

Boruta is a new selection feature technique implemented as an R package is a feature selection wrapper algorithm that takes into account all relevant features. By comparing the importance of original attributes to the importance that is achieved at random, the relevant features are determined [19].

4-4-2 Forward and Backward Algorithm

The searching procedure begins with a blank set of features in forward selection. Following that, a subset will be generated by adding one feature at a time to each stage. To select which feature should be added to the collection, all features are ordered according to a predetermined criterion, and the best option is picked. Backward selection begins with a collection containing all feasible features, which are then rejected one by one in subsequent phases of the search. Of course, this feature is no longer available from the feature set. Both elementary procedures are utilized alternatively in a mix of forward and backward selection. [20].

4-4-4 Variable Importance Algorithm

The variable importance value can be used to identify significant features. This allows the method to decrease a dataset's processing computational overheads while also improving detection rates [21].

4-4 Classification Methods

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

Dataset Classification by the model of machine learning is one of the main methods used for discovering each item of the dataset's known classes [22]. Some classification algorithms are tested on the CICDDOS2019 dataset, in the current paper. To verify the capability and the accuracy of these algorithms in the analysis of the dataset. Three classification algorithms are performed for attaining the purpose of current research: (SVM), (RF), and (NB). A description of algorithms will be given in the following.

4-4-1 Random Forest Classifier (RF)

Random forest classifier (RF) provides an algorithm missing value estimation, as well as the ability to do many sorts of data analysis, classification, and unsupervised learning. And it is robust to training data reduction and noise [23]. An (RF) is a community classifier that uses a subset of variables to construct a multi-decision tree [24].

4-4-2 Naive Bayes classifier (NB)

A naive Bayes classifier (NB) is a simple and very effective probabilistic classification method. The naive Bayes classifier is based on the premise of strong independence [25]. Naïve Bayes is used because of its simplicity and good performance in classification, and it is one of the popular classes in terms of accuracy and computational efficiency [26].

4-4-3 Support Vector Machines classifier (SVM)

(SVM) is a learning supervised model which used to distinguish two classes that are based on statistical learning theory. SVM solves problems related to classification, learning, and prediction [27]. For classification, SVM's core strategy is to identify the hyper-plane that offers the best separation between the two classes. Typically, in SVM, a set of data called the training dataset is used to develop the hyper-plane, and the generalizing capacity of the developed hyper-plane is confirmed using an independent subset termed the testing dataset [28].

4-5 Performance Metric

The term performance metrics refers to a collection of metrics based on the confusion matrix and are used to evaluate the performance of different combinations of machine learning and a dataset's features [29]. The confusion matrix's dimensions are 2*2 where the main diagonal indicates correct predictions and the secondary diameter indicates incorrect predictions, table 3 shows the following:

Table (3): confusion matrix

Predicted	**Actual**	
	Normal	Attack
Normal	True Negative (TN)	False Positive (FP)
Attack	False Negative (FN)	True Positive (TP)

Table 3 contain [30]:

True Positive (TP): this worth demonstrates that the attack packets were correctly classified as attacks.

True Negative (TN): this worth indicates that the normal packets were correctly classified as normal.

False Negative (FN): this worth indicates that the attack packet was incorrectly classified as normal.

False Positive (FP): this worth indicates that a normal packet was incorrectly classified as an attack.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

Among the performance metrics that were used in this study are: Accuracy (ACC), Precision (PR) or Positive Predictive Value, Recall (Rc) or Sensitivity, F1-Measure (F1), Specificity (Sp). Each metric will be explained.

Accuracy (ACC) [31]: is the most widely used metric for assessing performance a classifier's accuracy. Is what it's called the percentage of correct classifications, which are defined as equation (1):

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

Precision (PR) [32]: the relationship between true positive predicted values and complete positive predicted values, which is described as an equation is displayed (2):

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

Recall (Rc) [32]: It's the proportion between true positive prediction values to the total of predicted true positive and predicted false negative values, which is described by equation (3):

$$\text{Recall (Rc)} = \frac{TP}{TP+FN} \quad (3)$$

F1-Measure (F1) [32]: is an overall measure of the model's accuracy that combines precision and recall, which are defined as equation (4):

$$\text{F1-Measure (F1)} = \frac{2(PR*RC)}{PR+RC} \quad (4)$$

Specificity (Sp) [33]: the Sp measures the proportion of negative patterns being correctly recognized as being negative, which are defined as equation (5):

$$\text{Specificity (Sp)} = \frac{TN}{TN+FP} \quad (5)$$

5- Results and Discussion

After analyzing the outputs, this section discusses the stages of implementing the proposed work in addition to the experimental results. The dataset was classified using machine learning methods after the process of preprocessing and feature selection. The tested PortScan dataset CICDDOS2019 contains 79 features, after pre-processing them and removing redundant and irrelevant features, it became 68 of (79) features. After that, it moves to the next processing stage, which is the selection of the feature.

The following tables (4, 5, 6, and 7) show the results of performance metrics using three feature selection algorithms with three classification algorithms. The use of feature selection algorithms aims to make the number of attributes smaller and create a subset of important attributes. Where the Boruta algorithm minimizes the number of features to 56 features, the Forward and Backward algorithm reduces the number of features to 59 features, and the Variable Importance algorithm reduces the number of features to 11 features. After that, the used classification algorithms are implemented.

Table (4): the results of the Random Forest algorithm with three feature selection algorithms

Random Forest					
	ACC	PR	RC	F1	SP
Boruta	1	0.99991	1	0.99994	0.9999
Forward and Backward	**1**	**0.99993**	**1**	**0.99994**	**0.9999**
Variable Importance	0.9982	0.99674	0.99993	0.99829	0.9967

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

Table (4) and figure (3), show the use of the Random Forest (RF) classification algorithm with the three feature selection algorithms: Boruta, Forward and Backward, and Variable Importance. The best values of performance metrics derived from the (RF) with (FwBw), which gave Accuracy (1), Precision (0.99993), Re-call (1), F1-measure (0.99994), and Specificity (0.9999).

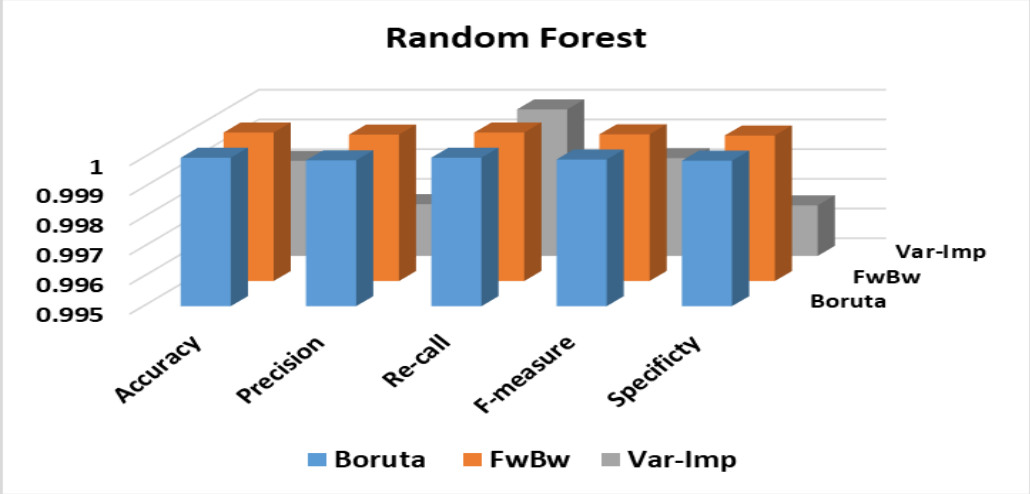

Figure (3): represents all performance metrics with a Random Forest classifier.

Table (5): the results of the Naïve Bayes algorithm with three feature selection algorithms

Naïve Bayes					
	ACC	PR	RC	F1	SP
Boruta	**0.9837**	**0.99488**	**0.97628**	**0.98541**	**0.9949**
Forward and Backward	0.954	0.99916	0.92419	0.96013	0.9992
Variable Importance	0.9352	0.9992	0.89614	0.94484	0.9992

Table (5) and figure(4), illustrate the best performance metrics values obtained from the Naive Bayes classifier appeared with Boruta, which gave Accuracy (0.9837), Precision (0.99488), Re-call (0.97628), F1-measure (0.98541), and Specificity (0.9949).

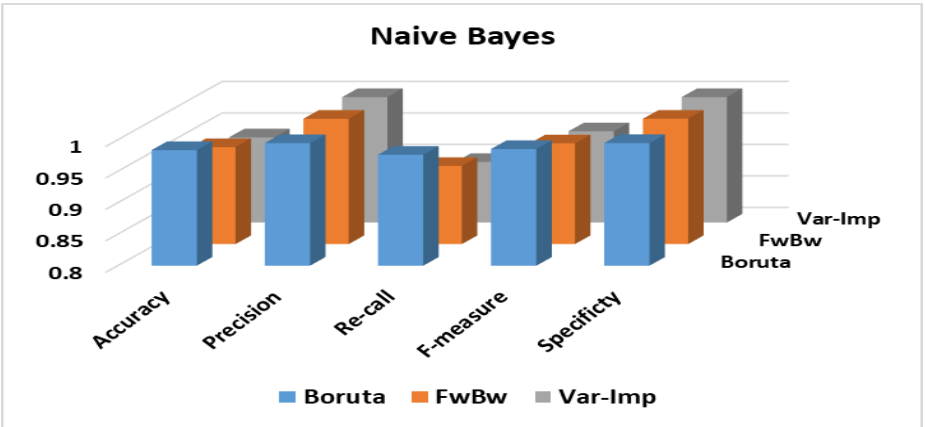

Figure (4): represents all performance metrics with a Naïve Bayes classifier.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

Table (6): the results of the Support Vector Machine algorithm with three feature selection algorithms

Support Vector Machine					
	ACC	PR	RC	F1	SP
Boruta	0.8862	0.7953	1	0.88598	0.7953
Forward and Backward	0.8869	0.79656	1	0.88672	0.7966
Variable Importance	**0.9962**	**0.99325**	**0.99987**	**0.99648**	**0.9933**

Table (6)and figure(5), illustrate the best performance metrics values obtained from the Support Vector Machine classifier appeared with Variable Importance, which was given Accuracy (0.9962), Precision (0.99325), Re-call (0.99987), F1-measure (0.99648), and Specificity (0.9933).

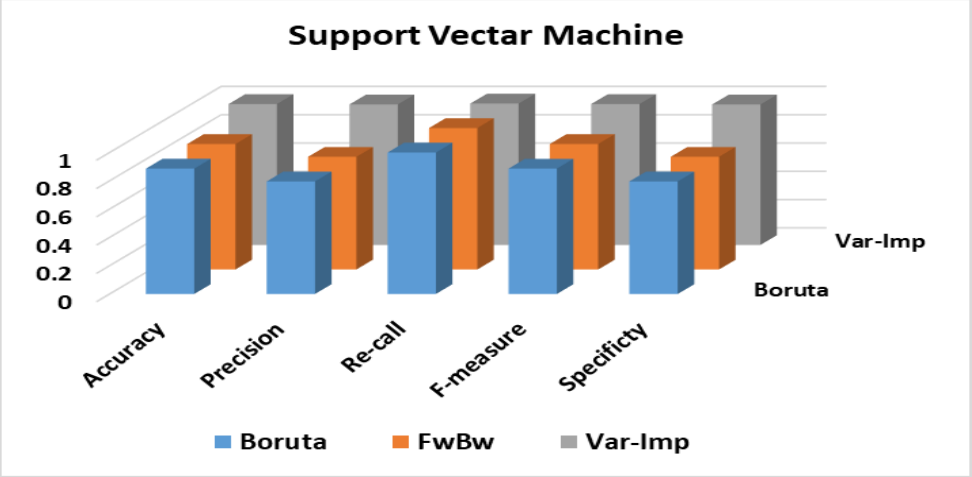

Figure (5): represents all performance metrics with a Support Vector Machine classifier

Table (7): The results of the Best work for each algorithm

Best work for each algorithm					
	ACC	PR	RC	F1	SP
Random forest / FwBw	**1**	**0.99993**	**1**	**0.99994**	**0.9999**
Naïve Bayes /Boruta	0.9837	0.99488	0.97628	0.98541	0.9949
Support Vectpr Machine /Var-Imp	0.9962	0.99325	0.99987	0.99648	0.9933

Table (7) and figure (6), show the results of the three classification algorithms are compared. It was found that the best performance was done using Random Forest with Forward and Backward, all the results of performance metrics were higher than the other algorithms used.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

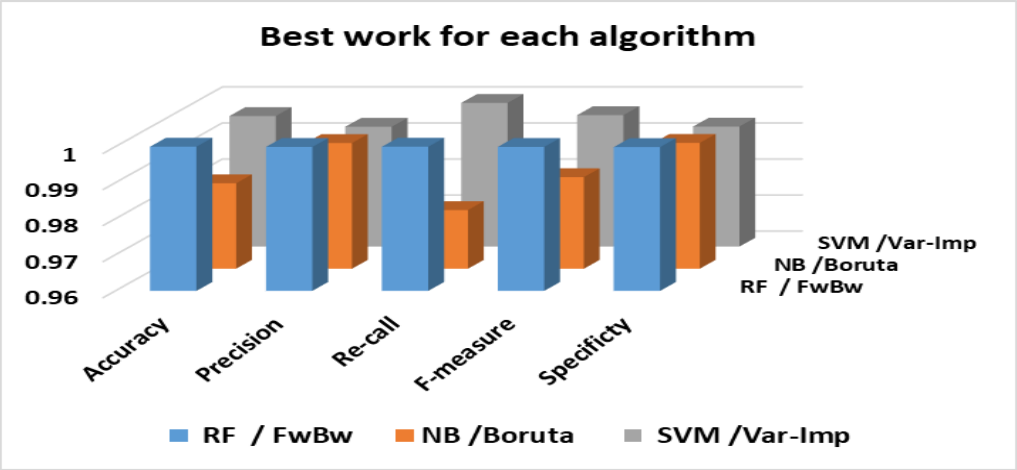

Figure (6): represents all performance metrics with the best work for each algorithm

In short, the forward and backward performance has been considered the best feature selection model, because all features are sorted and the best ones are chosen based on certain criteria. As for Random Forest, it enhances classification accuracy for multi-category classification tasks, Random Forest was deemed the best model.

6- The Comparison between this study and related work

Compared with the related works, this study was characterized by using three classification algorithms with three feature selection algorithms. Using the RF classification algorithm with the feature selection algorithm FwBw, which achieved the highest accuracy of 100%, Precision 0.99993, Recall 1, F1-Measure 0.99994, and Specificity 0.9999. Therefore, the proposed model in this study is highly efficient.

Table (8): The difference between related works

Related works reference No.	Dataset	Purpose	Feature selection methods	Classification algorithms	Performance measures	Description the result
5	ISCXIDS2012, KDDCUP99, CICIDS2017, and CICDDOS2019	Intrusion detection and classification	Krill Herd optimization	Support value-based graph, SVM, NB, and RF	ACC, Se, Sp, Pr, Rc, F1, FPR, FNR, Kap, and Rank sum	High ACC for CICDDOS 2019 equals 99.6% Pr and Rc 99.2%
6	CICDDoS2019 The first dataset is the Portmap attack. The second dataset is LDAP and NetBIOS attack	Detecting DDoS Attacks	correlation test	binomial logistic regression, polynomial logistic regression	Acc, Pr, Rc, and F1	The ACC in Portmap dataset is 99.91% and the F1 is 0.9913 The ACC in LDAP and NetBIOS dataset is 99.94% and the F1 is 0.9847
7	CICDDoS2019	For intrusion detection in smart grid	tree-based and Pearson's correlation coefficient	stacking-based, bagging-based,	TPR, FNR, FPR, and ACC	High ACC in the stacking-

				and boosting-based		based equals 93.4%
8	IP flows in SDN network	detection of portscan and DDoS attack	--	LSTM_FUZZY, KNN, SVM, MLP, POS-DS, and LSTM-2	Rc, Pr, and FPR	High Rc in the LSTM_FUZZY equals 99.87, and the Pr is 99.74
9	CICIDS2017	detection of portscan attempts	--	Deep learning and SVM	Acc, Pr, Rc, and F1	High ACC in deep learning equals 97.80%, Pr, Rc, and F1 is 0.99
10	KDD99	Classification of Portscan Attacks	Consistency Subset Evaluation algorithm and Best First search method	SVM	Acc, TPR, FPR, Pr, and Rc	High ACC equals 99.91%

7- Conclusion

Briefly, in the present paper, the CICDDOS2019 dataset has been processed, by reducing the number of features using feature selection algorithms and applying some appropriate classifiers. Each algorithm identifies certain features to determine the minimum number of features, generates precise outcomes based on performance metrics, and considers the best results for analyzing the dataset. The best results achieved in this study were when the feature selection algorithm is being used (FwBw) with (RF), which gave an ACC of 100%, PR 0.99993, RC 1, F1 0.99994, SP 0.9999.

References

[1] S. S. K. M. W. Allen H. Renear1, Definitions of Dataset in the Scientific and Technical Literature, 2010.
[2] K. S. M. A. H. S. S. H. R. Maryam Zaffar, A Study of Feature Selection Algorithms for Predicting Students Academic Performance, 2018.
[3] W.-W. L. 1. T.-T. N. Chin-Shiuh Shieh 1, Detection of unknown ddos attacks with deep learning and gaussian mixture model, 2021.
[4] C. R. E. S. Cynthia Bailey Lee, Detection and Characterization of Port Scan Attacks, 2003.
[5] V. K. P. Rahul B Adhao, Support Based Graph Framework for Effective Intrusion Detection and Classiication SUPPORT BASED GRAPH FRAMEWORK FOR EFFECTIVE INTRUSION DETECTION AND CLASSIFICATION, 2021.
[6] B. R. Bajracharya, DETECTING DDOS ATTACKS USING LOGISTIC REGRESSION Duplicate Bug Tracker View project Personality Based Music Recommendation System View project, 2020.
[7] N. K. G. A. Tala Talaei Khoei, Ensemble Learning Methods for Anomaly Intrusion Detection System in Smart Grid, 2021.
[8] L. F. C. J. L. MATHEUS P. NOVAES, Long short-term memory and fuzzy logic for anomaly detection and mitigation in software-defined network environment, 2020.
[9] M. A. A. Dogukan AKSU, Detecting Port Scan Attempts with Comparative Analysis of Deep Learning and Support Vector, 2018.
[10] M.Vidhya, Efficient Classification of Portscan Attacks using Support Vector Machine, 2013.
[11] A. H. L. A. A. G. Iman Sharafaldin, Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy, 2019.
[12] R. S. M. Jupri1, Data mining, fuzzy AHP and TOPSIS for optimizing taxpayer supervision, 2019.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 1712-1723
https://publishoa.com
ISSN: 1309-3452

[13] G. C. S. J. M. D. Yuyang Zhou, Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier, 2019.
[14] I. J. M. Aaya F. Jabbar, Botnet Detection Framework based on Five Features-Distance Measures supported by Comparisons of Four machine learning Classifiers using CICIDS2017 Dataset, 2019.
[15] I. J. M. 2. Aaya F. Jabbar 1, Development of an Optimized Botnet Detection Framework based on Filters of Features and Machine Learning Classifiers using CICIDS2017 Dataset, 2020.
[16] N.-A. L.-K. S. D. Mahmoud Said Elsayed, DDoSNet: A Deep-Learning Model for Detecting Network Attacks, 2020.
[17] V. R. S.Visalakshi, A Literature Review of Feature Selection Techniques and Applications, 2014.
[18] H. D. M. A. C. Huseyin Polat, Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods, 2017.
[19] B. N. S. K. Chithra Selvaraj, Empirical study of feature selection methods over classification algorithms, 2018.
[20] K. L. Izabela Rejer, Genetic algorithm and forward method for feature selection in EEG feature space, 2013.
[21] D. S. K. J. H. K. J. S. P. Sang Min Lee, Spam detection using feature selection and parameters optimization, 2010.
[22] D. T. M. D. M. V. D. G. V. Dr Gnanambal S, Classification Algorithms with Attribute Selection: an evaluation study using WEKA, 2018.
[23] B. G. J. R. M. C.-O. J. R.-S. V.F. Rodriguez-Galiano, An assessment of the effectiveness of a random forest classifier for land-cover classification, 2012.
[24] M. A. A. A. H. Z. Tuğba Aytaç, Detection DDOS attacks using machine learning methods, 2020.
[25] N. S. Dr. Saurabh Mukherjee, Intrusion Detection using Naive Bayes Classifier with Feature Reduction, 2012.
[26] W. I. A. H. T. S.L. Ting, Is Naïve Bayes a Good Classifier for Document Classification, 2011.
[27] B. B. S. A. A. A. M. H. Noreen Kausar, A Review of Classification A Review of Classification Approaches Using Support Vector Machine in Intrusion Detection, 2011.
[28] D. M. A. P. Thomas Oommen, An objective analysis of support vector machine based classification for remote sensing, 2008.
[29] A. H. S. S. S. M. M. R. Md. Shohel Rana, Evaluation of tree based machine learning classifiers for android malware detection, 2018.
[30] M. A. S. K. M. A. Mohammad Almseidin, Evaluation of Machine Learning Algorithms for Intrusion Detection System, 2017.
[31] S. W. Y. Z. C. T. Yangguang Liu, A Strategy on Selecting Performance Metrics for Classifier Evaluation, 2014.
[32] M. A. S. M. E. K. Ezz El-Din Hemdan, COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images, 2020.
[33] R. R. a. V. Palade, Optimized Precision - A New Measure for Classifier Performance Evaluation, 2006.