EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Impact of Features Reduction on Machine Learning Based Intrusion Detection Systems

Masooma Fatima[1,*], Osama Rehman[2] and Ibrahim M. H. Rahman[3]

[1]Systems Ltd, Karachi, Pakistan
[2]Department of Software Engineering, Bahria University, Karachi, Pakistan
[3]The Open Polytechnic of New Zealand, Wellington, New Zealand

Abstract

INTRODUCTION: As the use of the internet is increasing rapidly, cyber-attacks over user's personal data and network resources are on the rise. Due to the easily accessible cyber-attack tools, attacks on cyber resources are becoming common including Distributed Denial-of-Service (DDoS) attacks. Intruders are using enhanced techniques for executing DDoS attacks.
OBJECTIVES: Machine Learning (ML) based classification modules integrated with Intrusion Detection System (IDS) has the potential to detect cyber-attacks. This research aims to study the performance of several machine learning algorithms, namely Naïve Bayes, Decision Tree, Random Forest, and Support Vector Machine in classifying DDoS attacks from normal traffic.
METHODS: The paper focuses on DDoS attacks identification for which multiclass dataset is being used including Smurf, SIDDoS, HTTP-Flood and UDP-Flood. balanced datasets are used for both training and testing purposes in order to obtain biased free results. four experimental scenarios are conducted in which each experiment contains a different set of reduced features.
RESULTS: Result of each experiment is computed individually and the best algorithm among the four is highlighted by mean of its accuracy, detection rates and processing time required to build and test the classifiers.
CONCLUSION: Based on all experimental results, it is found that Decision Tree algorithm has shown promising cumulative performances in terms of the metrics investigated.

Keywords: DDoS attacks, Random Forest, Naïve Bayes, SVM, WEKA, IDS.

Received on 03 February 2022, accepted on 31 March 2022, published on 13 April 2022

doi: 10.4108/eetsis.vi.447

*Corresponding author. Email: masoomafatima69@gmail.com

1. Introduction

Intrusion is defined when a user tries to access any kind of information whose access is not authorized to use, making the user an intruder who can be an internal or external. Internal intruders have limited access but overstep legitimate access of rights. In comparison, external intruders don't have rights to access the systems, but do so through illegitimate ways. In the present world, cyber-attacks can occur at any time and over any system comprising a cyber-component connecting it to the Internet or even to a local network. The universal use of computers and computer networks in today's culture has made computer network security a universal issue [1].

The risk of cyber-attacks increases with internet connection and recently it was suggested that more sensitive data have higher probability to become the target to cyber- attacks, such as banks that have millions of customer records [2]. In such attacks, the attackers steal sensitive data and attempt to blackmail organizations as

well as individuals. The concept of Intrusion Detection System (IDS) was introduced by Anderson in 1980 with the purpose to assist a network to identify cyber-attacks. IDS has become an important area of research to detect undesired intrusion to a sensitive system. Intrusion is any set of actions that intimidates the reliability, accessibility, or privacy of a network resource. Whereas intrusion detection is the process of monitoring and analysing the occurring events and activities in the system or network to find out unusual behaviours.

As existing network systems have gaps and weaknesses, similarly the conventional IDS also lags in performance and fails on many occasions to discriminate Distributed Denial-of-Service (DDoS) attacks from normal traffic. The main purpose of DDoS attacks is to deny services availability to legitimate users by increasing the rate of requests to the server. The following points summarize the need for a more sophisticated ID to tackle trending cyber-attacks:

- Current IDS are non-intelligent systems requiring manual configuration.
- Many advance types of attacks can go undetected suchas day zero viruses.
- There are many cutting-edge algorithms that can be incorporated with the IDS, but still researchers are not fully aware of which is more efficient.
- Different data mining algorithms exist that can fit wellwith IDS.

Figure 1 identifies the typical placement of an IDSwithin a network that monitors a network or system for malicious activities or policy violations. The IDS deployment could be hardware, software or a combination of both [3]. An IDS finds out the unusual behaviour in data traffic by considering the following two steps:

- Checking and evaluating the received traffic.
- Identifying irregular activities in data.

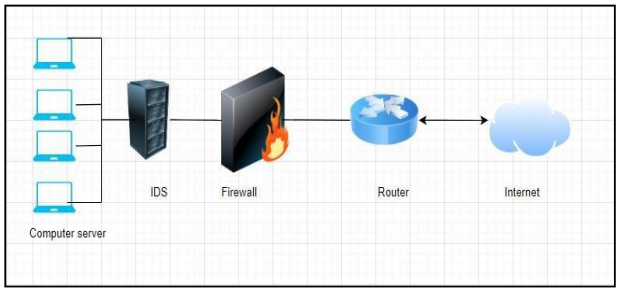

Figure 1. Typical Placement of an IDS within a Network.

IDS can be classified into two categories. Based on the analysis methods, the IDS can be classified as misuse-based IDS and anomaly-based IDS. On the other hand, based on the sources of data, the IDS can be classified as host-based IDS and network-based IDS.

Misuse Detection IDS: Misuse detection technique matches the sequence of actions for known intrusion scenarios with a predefined signature. The limitation of this category is that it is limited to find only known cyber-attacks and fails to find new attacks, such as zero-day attacks [4, 5].

Anomaly Based IDS: This detection technique uses profiles, which present the normal behavior of the system. The disadvantage of this category is sometimes incorrect profile data causes false alarms for normal data traffic, hence being identified as an attack[4, 5].

Host-Based IDS: Host-based IDS aims to collect information and perform analysis on a particular host or system and host agent monitor. It prevents intruding on a particular system but does not support to monitor the whole network. HIDs rely upon heavy audit trails and system logs to identify unusual activity in the system [4].

Network-Based IDS: Network-based IDS are active systems deployed on networks to monitor internal network traffic [4, 5].

Intruders are using enhanced techniques for executing DDoS attacks due to which such attackers are mimicking authentic users while accessing network resources, hence making it difficult for the security mechanisms to block such DDoS attacks. However, Machine Learning (ML) based classification modules integrated with IDS has the potential to detect such cyber-attacks. Such techniques can play a vital role in identifying of attacks leading towards improvement in overall accuracy rate in classifying DDoS attacks from normal traffic. This research aims to study the performance of several machine learning algorithms for their capabilities to classify DDoS attacks from normal network traffic.

Rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 describes the proposed mechanism adopted for building the classification model. Section 4 presents and discusses the obtained results. Finally, Section 5 concludes this work and provides future directions.

2. Literature Survey

In present literature, several works have been performed already related to classifications and surveys on IDS in Cyber System using data mining, but none of them has proposed the authentic DM techniques for IDS. In our research, we will use supervised learning and compare their results along with suggesting the authentic techniques for detection of the intrusion.

In [6], the author discussed the importance of IDS along with the classification process of data mining. Classification is the process to assign data items to a predefined class and it follows two steps which are training and testing. Firstly, classifier is trained to predict the class of labelled and un-labelled data instances, Binary class has one or two classes while for multiclass serval classes are involved, there are many techniques of classification including naïve Bayes, support vector

machine, decision tree, genetic algorithm, fuzzy logic, neural networks etc. However, in this paper authors have used misuse detection to identify intruders [7].

In a decision tree algorithm, a tree-like structure expresses a classification rule. The algorithm uses the divide and conquers method to split data according to the given attribute values [8]. Splitting process considers every child node till all selected attributes. Decision tree converts given data set to tree structure node of the tree presents features and edges which represent an association between features [9]. J48 is an extension of C4.5 and labelled as C4.5 algorithm uses gain ratio in procedure, J48 tree is a non-binary tree, data set split through the value of the root node and the value of the root node depends on the feature whose value is highest [10]. In [10], the paper is divided into six parts, the first part consists of the introduction, In second part data-driven framework is presented in terms of cyber security situational awareness after this author discusses about the data mining base attack detection from data analysis and data pre-processing author briefly define the techniques of the DM which includes classification, clustering, association rule mining, and outlier

Random forest is a supervised ML algorithm established on a group of trees where each tree produces random selection. Naïve Bayes is based on naïve notion, where the notion denotes that the presence of one variable in the problem has no impact on the presence of another variable. Naïve Bayes uses conditional probability to classify the problem by combining prior calculated likelihood and probabilities to make the next probability [11]. In [11] author briefly discussed the concept of the AI with IDS, mentioned review of the related studies for most popular used AI algorithms from the comparisons and results of the studies concludes that IDS based on naïve Bayes and decision tree gives more accurate results in terms of performance and accuracy.

In [12] , the author used rule-based and ML algorithms to improve efficiency of the IDS, used Neural Networks (NN), Random Forest and support vector machine algorithms (SVM) with KDDcup 99 dataset, the accuracy of SVM is better than other algorithms. In [13] author used three algorithms Naïve Bayes (NB), Support Vector Machine (SVM), and K-nearest neighbour (KNN) with KDDcup 99 dataset, detect the accuracy by reducing the processing time compared results of all algorithms, confusion matrix have created for better comparison, with help of the confusion matrix author have suggested that SVM gives better results as compared to NB and KNN. In [14], the author did literature studies using different data sets reviewed all datasets in detail, perform normalization on these datasets. For classification, different algorisms are used which includes support vector machine (SVM), K-Nearest neighbour (KNN) and Decision Tree (DT).

In [15], the authors highlight network security as a central non-functional requirement of the system along with defining types of IDS, necessity of the intrusion detection system, and a genetic block of IDS. Moreover, the concept and the need of misuse and anomaly-based IDS are briefly described. In this paper, two techniques of data mining have been used, which are J48 and Naïve Bayes algorithm. In [16] author has proposed a survey paper in which he covered the most common ML techniques. In [17] author used association rule for detection of anomaly-based IDS, also defines association rule in relation with database.

In [18], the author proposed a comparative study for this purpose compared performance of three classification algorithms which includes Naïve Bayes, J48 and random forest algorithm. KDD NSL data set have been used for comparison of all algorithms after experiment author concludes that random forest gives better results in terms of rate detection and accuracy as compared to other two algorithms and security is compromised when IDS takes place. In [19], the author used DM to build an IDS model to increase the security of the network system. It also discusses the approach to improve efficiency at run time, describes the basic concepts of anomaly and signature-based IDS, used the bi-clustering technique to analyse the network traffic.

In [20], the author mentioned types of IDS, briefly discussed network-based IDS and the spasms which take place at an enterprise level, classification rule has been used for identification of the attack/intruder, author explained different types of network attacks, suggested a classification-based model for the identification of intrusion.

In [21], the author discussed anomaly-based IDS in detail, also mentioned issues of anomaly-based IDS, proposed anomaly-based IDs with network-based for identification of the intrusion.In [22], the author proposed that the attack detection rate increased after integration with DM techniques, for the experiment author has used the NSL-KDD cup dataset, captured results that the detection rate has been increased after integration with data mining techniques.

In [23], the author presents a survey in which he covered types of a network attacks, machine learning, and data mining technique integration for identification of attack in-network, covered complexity of ML/DM. In [24] author performed a survey on the existing work of the researchers and finalize the results that most of the researchers have used anomaly ID, use DARPA1998 and KDDCup1999 datasets mostly. Cyber-security structure is mainly composed on the network security system, IDS is used to find out the duplicates and safety breaks, these breaks could be external or internal. Cyber-security structure is mainly composed on the network security system, IDS is used to find out the duplicates and safety breaks these breaks could be external or internal.

In [25], author has proposed integration of IDS with DM, apply IDS with different DM techniques which are Association Rule, Clustering, Decision tree, SVM furthermore discuss about different types of the data. Detecting cyber-attacks indisputably has become a big data problem. IDS is used to find the activities which compromise privacy. In [25], the author presented a new approach called outlier detection method for detection of

the interruption, author chose network-based IDS, KDD has been dataset which received from real-world, from end results author concluded that the performance of the proposed IDS is better than the existing.

In [26], the author used a decision tree and random forest algorithm, to avoid overfitting problems author obtained the accuracy by using 10-fold-cross validation, from the evaluation concluded that the random forest is the finest DM algorithm for identification of intrusion. In [2 7] researcher introduced a hybrid algorithm for the identification of intrusion in the network for this purpose using the Naïve Bayesian and ID3 algorithms with KDD 99 dataset. Gets 99% accuracy from the suggested method. In [28] author proposed a multi-layer IDS, in which author compared results of MLP, Naïve Bayes and C4.5 algorithms, result shows that C4.5 achieved the highest accurate results as compared to the remaining algorithms.

In [29] author has proposed a comparative study for this purpose author examines the performance of four supervised algorithms to detect attack results, results indicate that C 4.5 algorithm performs outclass prediction accuracy other than three algorithms Naïve Bayes, Based learning, Multilayer perceptron.

In [30] author tried to figure out the best ML classification algorithm for identification of the intrusion, for this purpose author used KDD99 dataset, ML algorithms J48, Bayes Net, OneR, and BN. Compared results of all four procedures and concluded end results that J48 is the finest ML algorithm.

In [31] Jain presented a study about the drawbacks of generic-based IDS and IDS with real-time applications. In [32] author explained DM practices and different types of attacks, suggested a method in which integrated DM techniques with IDS for detection of the attack in the network. From the review, author concluded that mostly used DM techniques are classification, clustering and association rule which most researchers used for the identification of intrusion/intruders. Author implemented DT and GA approach for the detection of the known attacks, for better results author combined both DT and GA together which gives more accuracy [33]. In the research paper [34], author implement machine learning technique C4.5 for identification of the intrusion, for this purpose author used NSL-KDD data set with signature-based IDS, signature-based IDS is used to identify known attacks only which is the major drawback of this system.

Link flooding attack (LFA) is a dangerous type of DDoS attack which targets the modern-day network by blocking important links and eventually bringing the entire network down. In [35] author performed a survey of LFA pattern on all layers of software-defined network (SDN) with comparative analysis of migration techniques. In [35], the authors also discussed LFA different types, techniques, and behaviour over wire and wireless SDNs. A deep analysis of mitigation techniques is performed with their appropriateness for each SDN.

In [36], the authors proposed Machine Learning-based security framework, termed as CyberPulse++. The authors cover gaps in an existing solution of LFA in SDN where a trained machine learning repository is utilized to test captured network statistics in real-time to identify the abnormal path performance on network links. Experiments have been conducted to evaluate the efficiency and effectiveness of CyberPulse++ on a testbed.

Learning-based classifiers are shown to have the potential to ensure the security of online systems [37]. Recent studies suggested that classifiers merged together manage to be stronger than a single classifier against attacks. In [46], the authors show that discrete-valued random forest classifiers can be easily run away from opposed inputs. It was shown that random forest can be more exposed than support vector machine (SVM) either single or in a group to evasion attacks.

In [38], the authors proposed a framework named ExBERT which will predict if vulnerability can be manipulated or not. ExBERT is an improved Bidirectional Encoder Representations from Transformers (BERT) model, and results showed that the proposed framework achieves 91.1 % accuracy and 91.8% precision [38].

3. Proposed Mechanism

This section presents the used methodology in which four different ML algorithms have been utilized, namely Decision Tree, Naïve Bayes, Support Vector Machine and Random Forest. Each of these algorithms is investigated for its performance while considering its features. As depicted in Figure 2, the methodology includes seven steps which mainly encompasses the selection of data, feature selection, transformation, training and testing of the model. Each step is discussed in detail in the following sub-sections.

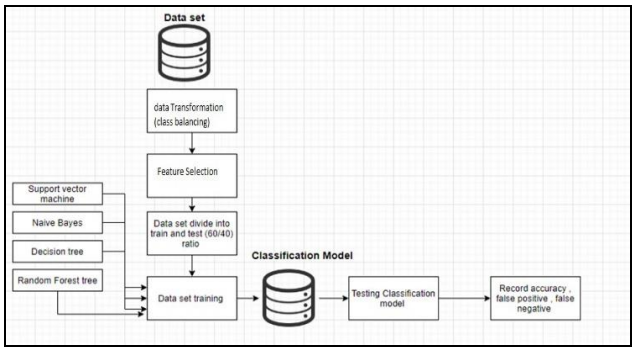

Figure 2. Mechanisms followed in the proposed Methodology.

3.1. Mechanisms followed in the proposed methodology

Dataset Selection

A survey for identification and analysis of the dataset is conducted where it is observed that the most used datasets

are KDD, DARAPA and NSL KDD [30, 39-42]. We have selected a suitable dataset that contains both DDoS attack records and normal traffic records, taken from [43]. This dataset contains recent types of DDoS attacks which are exactly of four types, encompassing HTTP Flood, SIDDOS, UDP Flood and Smurf attack records. Originally, the instances of each class in the given dataset are in imbalanced form, as shown by Table 1.

Table 1. Distribution of DDoS Attack and Normal Instances.

Class	Attack type	No of samples
DDoS	UDP Flood	12,514
DDoS	Smurf	12,777
DDoS	SIDDoS	12,739
DDoS	HTTP	12,602
Normal	Normal	50,224
Total		**100,856**

Transformation

After selection of the dataset, normalization is applied by removing duplicate and redundant records. Further, the dataset is transformed by class balancing process where each class, i.e., DDoS and normal traffic, makes 50% of the total records taken for executing training and testing processes.

Feature Selection

Several attributes present within the adopted dataset would not play any significant role in the classification process. In fact, their presence would become a potential reason for the overfitting phenomenon, which is an undesired outcome to have. Furthermore, greater number of involved attributes would lead towards a larger amount of processing time required by the classification model. We have used Weka (a machine learning tool) to identify the correlations between the set of attributes used in the considered dataset. Table 3 shows the correlation values between each attribute within the dataset. Correlation is mainly the degree of relationship between variables, where the condition is defined w.r.t. the correlation between attributes. We get different datasets having a different number of attributes in it, as shown below along with the defined conditions:

- ***Scenario#01:*** Select all attributes in the original dataset,hence all 27 attributes are selected.
- ***Scenario#02:*** Select all attributes having correlation value ≥ 0.1, hence the top 24 attributes are only selected.
- ***Scenario#03:*** Select all attributes having correlation value ≥ 0.2, hence top 18 attributes are only selected.
- ***Scenario#04:*** Select all attributes having correlation value ≥ 0.3, hence the top 7 attributes are only selected.

Splitting of Dataset

The dataset is split into 60% & 40%. Training data is 60% of the whole dataset, while the created classification models are evaluated against test records which are 40% of the whole dataset. Table 2 and Table 3 show the number of records used for training and testing, along with the detailed breakup of normal and attack traffics.

Table 2. Distribution of DDoS Attack and Normal Instances for Training Set.

Class	Attack type	No of samples
DDoS	UDP Flood	7,504
DDoS	Smurf	7,594
DDoS	SIDDoS	7,598
DDoS	HTTP	7,493
Normal	Normal	30,219
Total		**60,408**

Table 3. Distribution of DDoS Attack and Normal Instances for Test Set.

Class	Attack type	No of samples
DDoS	UDP Flood	5,010
DDoS	Smurf	5,183
DDoS	SIDDoS	5,141
DDoS	HTTP	5,109
Normal	Normal	20,005
Total		**40,448**

Training of Classification Model

In this step, classification algorithm is being used to train classification algorithm on each dataset one by one, we have tested the dataset on all selected algorithms.

Table 4. Time taken to build Training Model.

No of attributes (Features)	Processing Time NB	Processing Time DT	Processing Time SVM	Processing Time RF
27	0.81 sec	27.35 sec	902.8 sec	72.3 sec
24	0.38 sec	21.87 sec	890.1 sec	89.2 sec
18	0.3 sec	17.09 sec	818.9 sec	52.6 sec

7 Attributes	0.22 sec	9.34 sec	784.3 sec	47.5 sec

Testing of Classification Model

Test experiments are conducted in which the data is tested against each trained model of algorithms.

Evaluate Result

In this section we have evaluated the results for each single algorithm against each dataset by calculating the accuracy of the model. In addition, by finding the count of true positive, true negative, false positive and false negative, the time required by each classification model in the testing phase is evaluated since this parameter would depict the delay trends possibly induced by the classification model when deployed in an IDS in classifying traffic as normal and attack traffic.

3.2. Experimentation environment and setup

We have evaluated the performance of proposed classification models in-terms of accuracy rate, confusion matrix and the time needed for classifying the network traffic. We have conducted experiments on the WEKA tool whileconsidering version 3.9.3 that has been installed on Windows 10 operating system. The hardware configuration over which experiments were executed is 16 GB RAM. An experiment was conducted in an ideal system situation, i.e., while executing each experiment, the system was not engaged in any other activity. For each experiment, we have split the dataset into 60: 40 ratios.

Table 5. Balanced Dataset.

	Training	Testing
Normal Traffic	30219	20005
Attack Traffic	30189	20443
Total	60408	40448

4. Results and Discussion

Below, the table shows the accuracy rate of all algorithms of classification including Decision tree, Naïve Bayes, SVM & Random Forest. These were implemented on all scenarios mentioned above. Results of Scenario#1, Scenario#2, Scenario#3 and Scenario#4 are showed in Table 6, Table 7, Table 8 and Table 9.

Table 6. Test Results of 27 attributes.

	NB	SVM	RF	DT
Time taken to test model	3.16 sec	4.0 sec	7.45 sec	2.7 sec
Accuracy	84.2 %	86.7%	98.1%	97.2%
Recall	84%	86%	98%	97.6%
Precision	79%	86%	98%	97.6%

Table 7. Test Results of 24 attributes.

	NB	SVM	RF	DT
Time taken to test model	3.0 sec	3.34sec	6.9 sec	2.1 sec
Accuracy	84%	86.2%	98%	97.2%
Recall	84.1%	86%	98.1	97.2%
Precision	79.2%	86%	98%	97.2%

Table 8. Test Results of 18 attributes.

	NB	SVM	RF	DT
Time taken to test model	2.27 sec	2.98 sec	5.3 sec	1.5 sec
Accuracy	84%	86.7%	97.9%	97.1%
Recall	84.1%	86%	97.9%	97.1%
Precision	79%	86%	97%	97%

Table 9. Test Results of 07 attributes.

	NB	SVM	RF	DT
Time taken to test model	1.88 sec	2.0 sec	4.39 sec	0.9 sec
Accuracy	83.8%	86.63 %	97.2%	96.6%
Recall	83%	86.6%	97%	96.7%
Precision	78.8%	86.3%	97%	96.7%

Figure 3 presents the accuracy rate comparison between all investigated algorithms. While considering all four experimental scenarios, the highest recorded accuracy is 98% at scenario#1 in which the total number of attributes were 27. However, while having attributes as low as 7 only, the RF algorithm still shows a high level of accuracy, i.e., of 97%. In general, RF algorithm is displaying the highest rates of accuracy in all experimental scenarios. Although, RF has the highest accuracy, it also displays the highest amount of time needed in solving the classification, graph for classification needed time shown in Figure 4.

Confusion Matrix: In [2-4], the basic concept of the confusion Matrix is identification of false alarm rate, un-detection rate and new techniques for attribute selection. Confusion Matrix is used to define the performance of

the classification algorithm; it is a technique that summarizes the performance of the algorithm in which each row contains values showing instances of the predicted class while each column contains values representing an instance of the actual class.

False Alarm Rate (FAR): This represents the percentage (%) of the normal traffic classified as an attack by the model:

$$FAR\ \% = FP / FP + TN \times 100. \quad (1)$$

Un-Detection Rate (UND): The division of the attack that are misclassified as normal by the model:

$$UND\ \% = FN / FN + TP \times 100. \quad (2)$$

The below tables depict the confusion matrix of all algorithms with respect to all scenarios which we have mentioned in the methodology section. Table 10, Table 11, Table 12, and Table 13 shows the confusion matrix of Scenario#1, Scenario#2, Scenario#3 and Scenario#4, respectively.

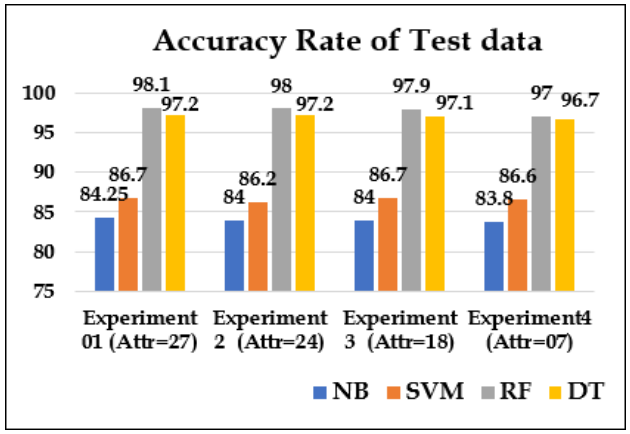

Figure 3. Accuracy Rate of Test Data.

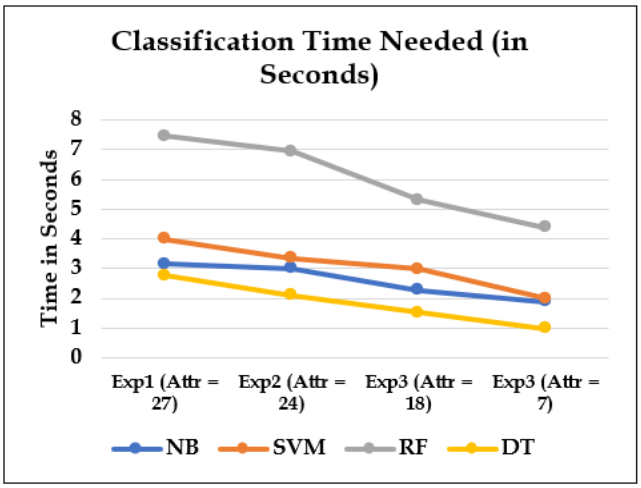

Figure 4. Time Required for Classification in Testing Phase by all Four Models.

Table 10. Experiment with 27 Attributes.

	NB	SVM	RF	DT
True positive	19658	19752	19869	19978
True negative	16539	16428	19942	19817
False positive	347	280	136	27
False negative	3904	3988	501	626

Table 11. Experiment with 24 Attributes.

	NB	SVM	RF	DT
True positive	19661	19764	19861	19989
True negative	16537	16419	19949	19820
False positive	344	278	144	16
False negative	3906	3987	494	623

Table 12. Experiment with 18 Attributes.

	NB	SVM	RF	DT
True positive	19645	19749	19875	19959
True negative	16536	16425	19843	19797
False positive	360	282	130	46
False negative	3907	3992	600	646

Table 13. Experiment with 7 Attributes.

	NB	SVM	RF	DT
True positive	19608	19712	19750	19925
True negative	16527	16431	19889	19716
False positive	397	293	255	80
False negative	3916	4012	554	727

Figure 5 shows rate of false positive for all classification algorithms. It can be seen that NB has the highest rate of false positive while DT has the lowest. In addition, Figure 6 shows rate of the false alarm, which is the percentage of the regular traffic misclassified as an attack by the model (equation 02). Figure 7 shows rate of false negative for all classification algorithms. With respect to all experiments for false negative, we have observed that SVM has the highest rate of false negative and RF has the lowest. Figure 8 shows the rate of un-detection UND (equation 02) which represents a percentage of the traffic as an attack but is classified as normal traffic (opposite of FAR).

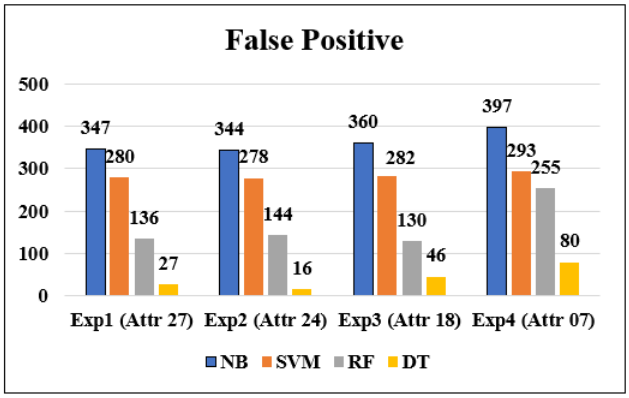

Figure 5. Rate of False Positive.

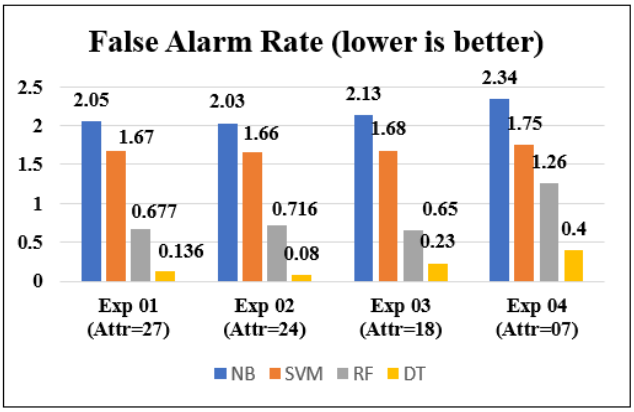

Figure 6. False Alarm Rate.

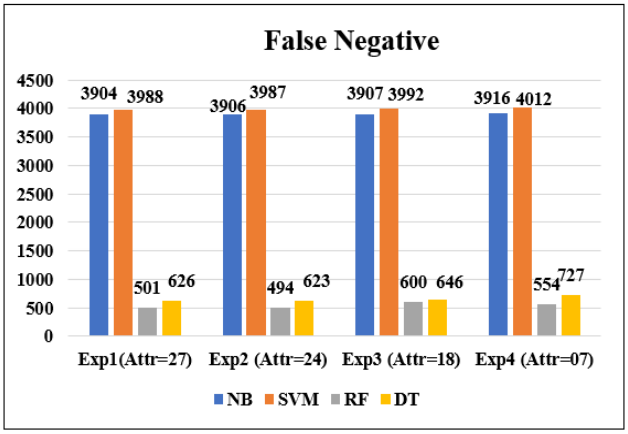

Figure 7. Rate of False Negative.

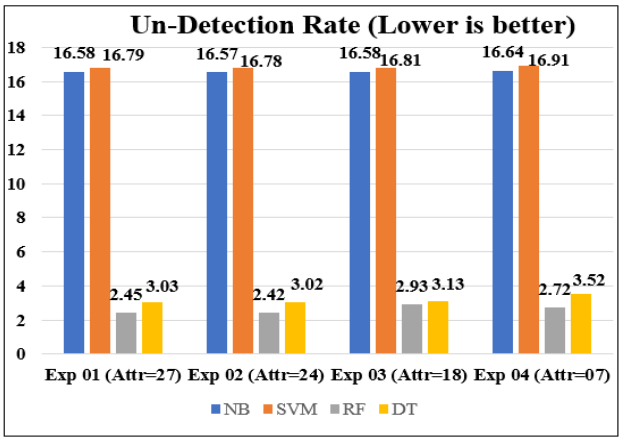

Figure 8. Un-Detection Rate (UND).

4. Conclusion

The objective of our research is to build machine learning based classification model using different classification algorithms. Firstly, an unbalanced dataset initially containing redundant data is selected for pre-processing, followed by balancing and dividing the data into an equal ratio for training and testing purpose. We have used four types of multiclass data of DDoS attack including Smurf, SIDDoS, HTTP-Flood and UDP-Flood on which evaluation is performed using different experiments and each experiment contains a different set of attributes. We have changed the set of attributes with the help of value of the correlation between attributes. Four experiments on the same dataset with different attribute rates have been conducted and from the experimental results we have observed that the random forest has the highest accuracy rate as compared to other classification algorithms including Naïve Bayes, Support Vector Machine and Decision Tree (J48). However, the algorithm Random Forest is not time-efficient when it comes to testing and training. In addition, it has also been observed from confusion matrix that the decision tree has the lowest rate of false positive while for false negative RF has the

lowest rate but the difference between RF and DT for false negative is very slight. From the results we concluded that the decision tree is the best classification algorithm as compared to other algorithms in term of time efficiency, accuracy percentage, and false positive and false negative.

References

[1] M. Alkasassbeh, G. Al-Naymat, A. Hassanat, and M. Almseidin, "Detecting distributed denial of service attacks using data mining techniques," *International Journal of Advanced Computer Science and Applications*, vol. 7, pp. 436-445, 2016.

[2] R. M. George and J. A. Mathew, "Emotion classification using machine learning and data preprocessing approach on Tulu speech data," *Int. J. Comput. Sci. Mobile Comput.*, vol. 5, pp. 589-600, 2016.

[3] B. P. Salmon, W. Kleynhans, C. P. Schwegmann, and J. C. Olivier, "Proper comparison among methods using a confusion matrix," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015, pp. 3057-3060.

[4] M. A. Teixeira, T. Salman, M. Zolanvari, R. Jain, N. Meskin, and M. Samaka, "SCADA system testbed for cybersecurity research using machine learning approach," *Future Internet*, vol. 10, p. 76, 2018.

[5] A. Kumra, W. Jeberson, and K. Jeberson, "Intrusion Detection System Based on Data Mining Techniques," *Oriental Journal of Computer Science and Technology*, vol. 10, pp. 491-496, 2017.

[6] D. K. Denatious and A. John, "Survey on data mining techniques to enhance intrusion detection," in *2012 International Conference on Computer Communication and Informatics*, 2012, pp. 1-5.

[7] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*: Elsevier, 2011.

[8] S. Dua and X. Du, *Data mining and machine learning in cybersecurity*: CRC press, 2016.

[9] G. L. Agrawal and H. Gupta, "Optimization of C4. 5 decision tree algorithm for data mining application," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, pp. 341-345, 2013.

[10] J. Patel and K. Panchal, "Effective intrusion detection system using data mining technique," *Journal of Emerging Technologies and Innovative Research*, vol. 2, pp. 1869-1878, 2015.

[11] M. Stampar and K. Fertalj, "Artificial intelligence in network intrusion detection," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, pp. 1318-1323.

[12] G. S. Sajja, M. Mustafa, R. Ponnusamy, and S. Abdufattokhov, "Machine Learning Algorithms in Intrusion Detection and Classification," *Annals of the Romanian Society for Cell Biology*, vol. 25, pp. 12211-12219, 2021.

[13] A. Agarwal, P. Sharma, M. Alshehri, A. A. Mohamed, and O. Alfarraj, "Classification model for accuracy and intrusion detection using machine learning approach," *PeerJ Computer Science*, vol. 7, p. e437, 2021.

[14] I. F. Kilincer, F. Ertam, and A. Sengur, "Machine learning methods for cyber security intrusion detection: Datasets and comparative study," *Computer Networks*, vol. 188, p. 107840, 2021.

[15] R. Samrin and D. Vasumathi, "Review on anomaly based network intrusion detection system," in *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, 2017, pp. 141-147.

[16] A. Dey, J. Singh, and N. Singh, "Analysis of supervised machine learning algorithms for heart disease prediction with reduced number of attributes using principal component analysis," *International Journal of Computer Applications*, vol. 140, pp. 27-31, 2016.

[17] H. Kong, C. Jong, and U. Ryang, "Rare association rule mining for network intrusion detection," *arXiv preprint arXiv:1610.04306*, 2016.

[18] N. Ashraf, W. Ahmad, and R. Ashraf, "A comparative study of data mining algorithms for high detection rate in intrusion detection system," *Annals of Emerging Technologies in Computing (AETiC), Print ISSN*, pp. 2516-0281, 2018.

[19] G. Nadiammai and M. Hemalatha, "Research Article Handling Intrusion Detection System using Snort Based Statistical Algorithm and Semi-supervised Approach," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 6, pp. 2914-2922, 2013.

[20] N. D Harale and D. B. Meshram, "Data mining techniques for network intrusion detection and prevention systems," *International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN*, pp. 2347-5552, 2016.

[21] V. Jyothsna, R. Prasad, and K. M. Prasad, "A review of anomaly based intrusion detection systems," *International Journal of Computer Applications*, vol. 28, pp. 26-35, 2011.

[22] K. K. Tiwari, S. Tiwari, and S. Yadav, "Intrusion detection using data mining techniques," *International Journal of Advanced Computer Technology*, vol. 2, pp. 21-25, 2013.

[23] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications surveys & tutorials*, vol. 18, pp. 1153-1176, 2015.

[24] H. Tianfield, "Data mining based cyber-attack detection," *System simulation technology*, vol. 13, 2017.

[25] J. Jabez and B. Muthukumar, "Intrusion Detection System (IDS): Anomaly detection using outlier detection approach," *Procedia Computer Science*, vol. 48, pp. 338-346, 2015.

[26] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, p. 272, 2012.

[27] D. M. Farid, N. Harbi, and M. Z. Rahman, "Combining naive bayes and decision tree for adaptive intrusion detection," *arXiv preprint arXiv:1005.4496*, 2010.

[28] H. E. Ibrahim, S. M. Badr, and M. A. Shaheen, "Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems," *arXiv preprint arXiv:1210.7650*, 2012.

[29] G. MeeraGandhi, "Machine learning approach for attack prediction and classification using supervised learning algorithms," *Int. J. Comput. Sci. Commun*, vol. 1, pp. 247-250, 2010.

[30] Y. K. Jain, "Upendra,"An Efficient Intrusion Detection based on Decision Tree Classifier Using Feature Reduction,"" *International Journal of Scientific and Research Publication*, vol. 2, pp. 1-6, 2012.

[31] S. Agrawal and G. Jain, "A review on intrusion detection system based data mining techniques," *Int. Res. J. Eng. Technol (IRJET)*, vol. 4, pp. 402-407, 2017.
[32] J. K. Chahal and A. Kaur, "Use of data mining techniques in intrusion detection–a survey," *Imperial Journal of Interdisciplinary Research*, vol. 2, pp. 452-6, 2016.
[33] K. Kaliyamurthie, D. Parameswari, and R. Suresh, "Intrusion Detection System using Memtic Algorithm Supporting with Genetic and Decision Tree Algorithms," *IJCSI International Journal of Computer Science Issues*, vol. 9, 2012.
[34] P. Gupta, S. Tandan, and R. Miri, "Decision Tree Applied For Detecting Intrusion," *International Journal of Engineering Research & Technology (IJERT) Vol*, vol. 2, pp. 2278-0181, 2013.
[35] R. ur Rasool, H. Wang, U. Ashraf, K. Ahmed, Z. Anwar, and W. Rafique, "A survey of link flooding attacks in software defined network ecosystems," *Journal of Network and Computer Applications*, vol. 172, p. 102803, 2020.
[36] R. U. Rasool, K. Ahmed, Z. Anwar, H. Wang, U. Ashraf, and W. Rafique, "CyberPulse++: A machine learning-based security framework for detecting link flooding attacks in software defined networks," *International Journal of Intelligent Systems*, vol. 36, pp. 3852-3879, 2021.
[37] F. Zhang, Y. Wang, S. Liu, and H. Wang, "Decision-based evasion attacks on tree ensemble classifiers," *World Wide Web*, vol. 23, pp. 2957-2977, 2020.
[38] J. Yin, M. Tang, J. Cao, and H. Wang, "Apply transfer learning to cybersecurity: Predicting exploitability of vulnerabilities by description," *Knowledge-Based Systems*, vol. 210, p. 106529, 2020.
[39] B. Ingre, A. Yadav, and A. K. Soni, "Decision tree based intrusion detection system for NSL-KDD dataset," in *International conference on information and communication technology for intelligent systems*, 2017, pp. 207-218.
[40] L. Dhanabal and S. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," *International journal of advanced research in computer and communication engineering*, vol. 4, pp. 446-452, 2015.
[41] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*, 2009, pp. 1-6.
[42] S. T. Brugger and J. Chow, "An assessment of the DARPA IDS Evaluation Dataset using Snort," *UCDAVIS department of Computer Science*, vol. 1, p. 22, 2007.
[43] W. J. Abhinav Kumra, and Klinsega Jeberson, "Intrusion Detection System Based on Data Mining Techniques" *Orient.J. Comp. Sci. and Tech*, vol. vol. vol. 10, 2017.