Online: ISSN 2645-2960; Print ISSN: 2141-3959

Random Forest Based Hypertext Transfer Protocol Distributed Denial of Service Attack Detection System for Cloud Computing Environment

[1]Morufu Olalere, [2]Rukayya Umar, [3]Juliana Ndunagu, [4]Ismaila Idris, [5]Raji Abdullahi Egigogo, [6]Suleiman Muhammad Nasir

[1,2,4,5]Department of Cyber Security Science, School of Information and Communication Technology,
Federal University of Technology, Minna, Nigeria.

[3]Department of Computer Science, National Open University of Nigeria, Abuja, Nigeria.

[6]Computer Science Department, Federal polytechnic, Nasarawa

[1]lerejide@futminna.edu.ng, [2]umarrukayya1@gmail.com, [3]jndunagu@noun.edu.ng,
[4]ismi.idris@futminna.edu.ng, [5]raji.pg610868@st.futminna.edu.ng.

Abstract

There is a need to secure data in the cloud from any form of attack. One among the many feared attacks in the cloud is the Hypertext Transfer Protocol Distributed Denial of Service (HTTP-DDoS) attack. HTTP-DDoS is the most devastating attack which stops the normal functionality of critical services provided by the various sectors in the cloud computing environment. Consequently, detection of HTTP-DDoS attack has attracted attention of many researchers, thereby leading to proposition of different approaches for detection of HTTP-DDoS attack in cloud computing environment. Meanwhile, machine learning approach is the most common approach previous researchers have used in addressing DDoS attack detection. However, achieving high detection accuracy with minimum false positive rate remains issue that still need to be addressed. Consequently, this study proposed solution to address the problem highlighted above by proposing machine learning based HTTP-DDoS attack detection system in cloud computing environment. To achieve this, the study designed a Random Forest based framework for HTTP-DDoS attack detection system. Thereafter, a Random Forest based model was formulated. The validation and testing of the model were carried out by experimentation with the application of data mining tool. Also, experimentation with other machine learning algorithms was carried out. Performance evaluation revealed that the Random Forest based model has an accuracy of 99.94% and minimum false positive rate of 0.001%. Also, when compared with existing detection models, this study detection model performed best in respect to accuracy and false positive rate.

Keywords: *Random Forest, HTTP-DDoS, Detection, Cloud, Accuracy, False Positive Rate (FPR)***.**

1. Introduction

Amongst techniques used in service deliveries by organizations, Cloud Computing have of recent, proven to be most effective and prevalent technique. This is mostly because it gives a medium to the huge advances required towards the development and distribution of an increasing amount of disseminated applications

Online: ISSN 2645-2960; Print ISSN: 2141-3959

(Marinescu, 2012). The principal objective of the cloud computing is that the clients can utilize and pay for just what they need. However, as subsequent information of people and firms are collated in the cloud data servers, questions related to the security and safety of cloud environment arise. Cloud computing can be a very easy target to attackers (Modi, Patel, and Muttukrishnan, 2012). According to (Sun, Guiran, Lina, and Xingwei, 2011), there are various number of securities, privacy and trust issues related with cloud computing. These issues have a great impact on the integrity of a client's information stored in the cloud. As such, despite the flexibility and proficiency provided by cloud computing, most clients seem to be hesitant on discrete data such as Personally Identifiable Information (PII) in the cloud.

In cloud computing, to tackle vital issues such as anonymity, liability, reliability and security when delivering important services on the internet through a pool of disseminated resources, security is of paramount significance, and policies guiding them must exist. In a link of computing mechanisms, three forms intrusion are likely to occur such as Denial of Service (DoS), Scanning and Penetration (Rup, Kausthav, Dhruba, Bhattacharya and Jugal, 2015). In addition to hacking, the cloud is continually under security attacks such as Structured Query Language (SQL) injection, Cross Site Scripting (XSS), DoS and Distributed Denial of Service (DDoS).

Frequently occurring network attacks which affect cloud security at the network layer includes IP spoofing, man-in-middle attack, Address Resolution Protocol (ARP) spoofing, Denial of Service (DoS) port scanning, Routing Information Protocol (RIP) attack and Distributed Denial of Service (DDoS) (Modi et al, 2013). As such, service providers are tasked with securing the systems against both internal and external attacks. The traditional network security can be employed in combating numerous external threat; but threats emerging from the network environment and complex external attacks such as DoS and DDoS threat cannot be easily regulated using such tool (Modi et al, 2012).

This paper presents a system which detects HTTP DDoS attacks in a Cloud environment. Subsequently, experiment was undertaken to choose appropriate classifier for HTTP DDoS detection with considerations on precision, FPR, TPR, and F-measure metrics. The results obtained from the experiments shows that Random Forest classifier exhibit the highest detection performance.

The rest of the paper is sectioned thus; Section 2 comprises of explanation of machine learning and details of recent related literatures. Section 3 present the methodology strategy used in evaluating the performance of the system, while section 4 reveals the result obtained as well as conclusion and analysis.

2. Review of Literature

2.1 Machine Learning Algorithms

Machine learning, as shown in figure 1, uses two types of methods: the supervised and unsupervised learning. The former trains a model on known input and output data so that it can predict future outputs, while the latter finds hidden pattern or intrinsic structures in inpur data ("MathWorks," 2018).

There are numerous machine learning algorithms. But for the purpose of this work, about eleven machine learning algorithms; namely; J48, Naive Bayes, IBK, Kstar, SMO, simpleLogistics, Multilayeperception, Decision Table, PART, NaivebayesSimple, BayesNet were also reviewed and tested for accuracy and False positive rate.

Online: ISSN 2645-2960; Print ISSN: 2141-3959

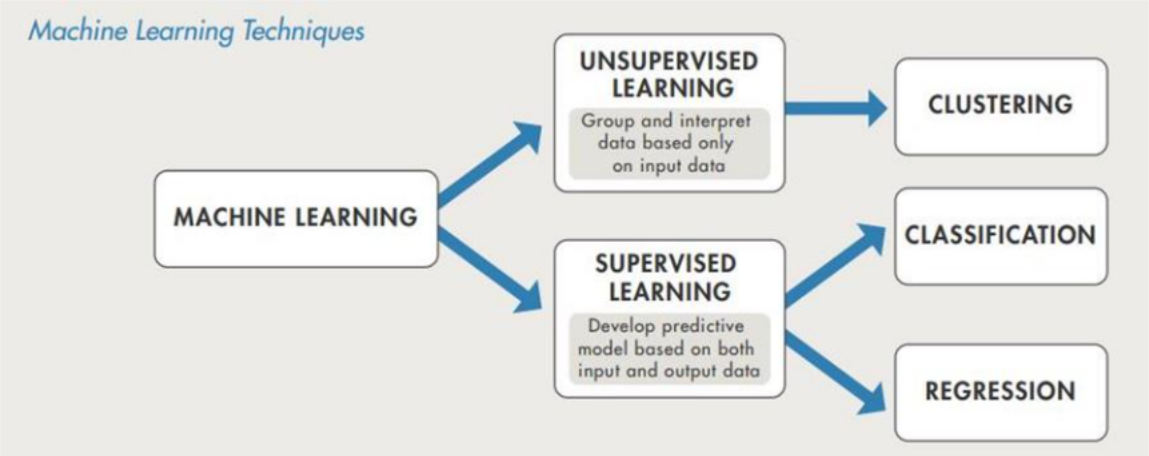
Machine Learning Techniques
MACHINE LEARNING
UNSUPERVISED LEARNING
Group and interpret data based only on input data
SUPERVISED LEARNING
Develop predictive model based on both input and output data
CLUSTERING
CLASSIFICATION
REGRESSION

Figure 1. Machine learning techniques categorization ("Mathworks," 2018)

2.2 Random Forest

Random forest classifiers were developed by LEO Breiman and Adele Cutler. They combine tree classifiers to predict new unlabeled data, the predictor depends on the number of as that are represented by the number of trees in the forest, the attributes are selected randomly, each number of trees represents a single forest and each forest represents a predation class for new unlabeled data (Apale, Kamble, Ghodekar, Nemade, and Waghmode, R, 2015). In this algorithm, random features selection will be selected for each individual tree. A random forest classifier ensemble learning algorithm is used for classification and prediction of the outputs based on an individual number of trees (Araar and Bouslama, 2014). Using random forest classifiers, many classification trees will be generated, and each individual tree is constructed by a different part of the general dataset. After each tree is classified in an unlabeled class, a new object will be implemented under each tree vote for decision. The forest chosen as the winner is based on the highest number of votes recorded.

Hence, for the purpose of this paper, Random Forest was adopted as the classifier in order to effectively classify the datasets as containing HTTP-DDoS attack packet or as normal packets and reduce the rate of misclassification. By so doing, it would be an improvement on the existing system.

2.3 Theoretical Review

Mouhammd *et al*. (2016) collected a new dataset that includes modern types of attack, which has not been used in previous research. The dataset contains 27 features and five classes. A network simulator (NS2) was used in the work. Three machine learning algorithms, namely; Multilayer Perception (MLP), Random Forest, and Naïve Bayes, were applied on the collected dataset to classify the DDoS types of attack namely: Smurf, User

Datagram Protocol Flood (UDP-Flood), HTTP-Flood and SQl Injection DDoS (SIDDOS). The MLP classifier achieved the highest accuracy rate with (98.63%).

Also, in line with Mouhammd *et al.* (2016), four classifiers (Naïve Bayes, Decision Trees, MLP, and SVM) were applied on the collected dataset to classify the DDoS types of attack namely: Smurf, UDP-Flood, HTTP-Flood and SIDDOS in (Irfan, Amit, and Vibhakar, 2017). The Multilayeperception (MLP) classifier achieved the highest accuracy rate with 98.91% which is a bit higher than the former. Examining the different features for feature selection technique and more sorts of future attacks in different OSI layers, such as the transport layer was recommended for future work.

Another experiment was done using the benchmarking dataset in (Indraneel and Venkata, 2017). The bat algorithm was adopted in the work. First, feature metrics was defined to identify if the request stream behavior is of attack or normal, Secondly, the bat algorithm was customized to train and test. Even though the devised bat algorithm amplified detection accuracy, it had maximal process complexity. The experiment achieved an accuracy of 98.4% using the CAIDA dataset.

Sharmila and Roshan (2018) performed series of experiment using the CAIDA UCSD DDoS Attack2007 Dataset and DARPA 2000 and proposed a system that detects DDoS attacks using the clustering technique followed by classification. Based on some network attributes of the data packet, Heuristics Clustering Algorithm (HCA) was adopted to cluster the available data and Naïve Bayes (NB) classification was also adopted to classify the data. Since clustering algorithm is based in unsupervised learning technique and is sometimes unable to detect some of the attack instances and few normal instances, therefore classification techniques are also used along with clustering to overcome this classification problem and to enhance the accuracy. The system's efficiency was tested using the following parameters; accuracy with 99.45% and false positive rate with 0.54%. Though the number of misclassifications need to be reduced,

More recent, (Mohamed, Karim and Mustapha, 2018) proposed a detection system of HTTP DDoS attacks in a Cloud environment. The system which is based on Information Theoretic Entropy and data learning classifier consists of three main steps: entropy estimation, preprocessing, and classification. A time-based sliding window algorithm was used in estimating the entropy of the network heeder features of the incoming network traffic and then classify the data into normal and HTTP DDoS traffic. Performance metrics based on accuracy, FPR, Area Under Curve (AUC), and running time metrics were used for the evaluation of the proposed detection system achieving an accuracy rate of 99.54% with 0.4 FPR.

3. Methodology

3.1 Proposition of HTTP-DDoS Attack Detection Framework

This research detection system for cloud environment is based on a Random Forest approach. In the designed framework, network traffic is classified as either attack or normal. The normal traffic is that which is anticipated between the client and the server, and the attacked traffic is that which is contrary to the anticipated one. This framework is designed to enhance real time detection with high detection accuracy, and low false positive. The detection system operates in a cooperative way with the classification algorithm for detection of the HTTP-DDoS attack on the go. In this way, any process or abnormality that can hinder network performance, availability and security will be analyzed and managed first while the random forest algorithm classifies the traffic as either normal or containing the attack type HTTP-DDoS. The designed HTTP-DDoS attack detection system is presented in Figure 2.

Online: ISSN 2645-2960; Print ISSN: 2141-3959

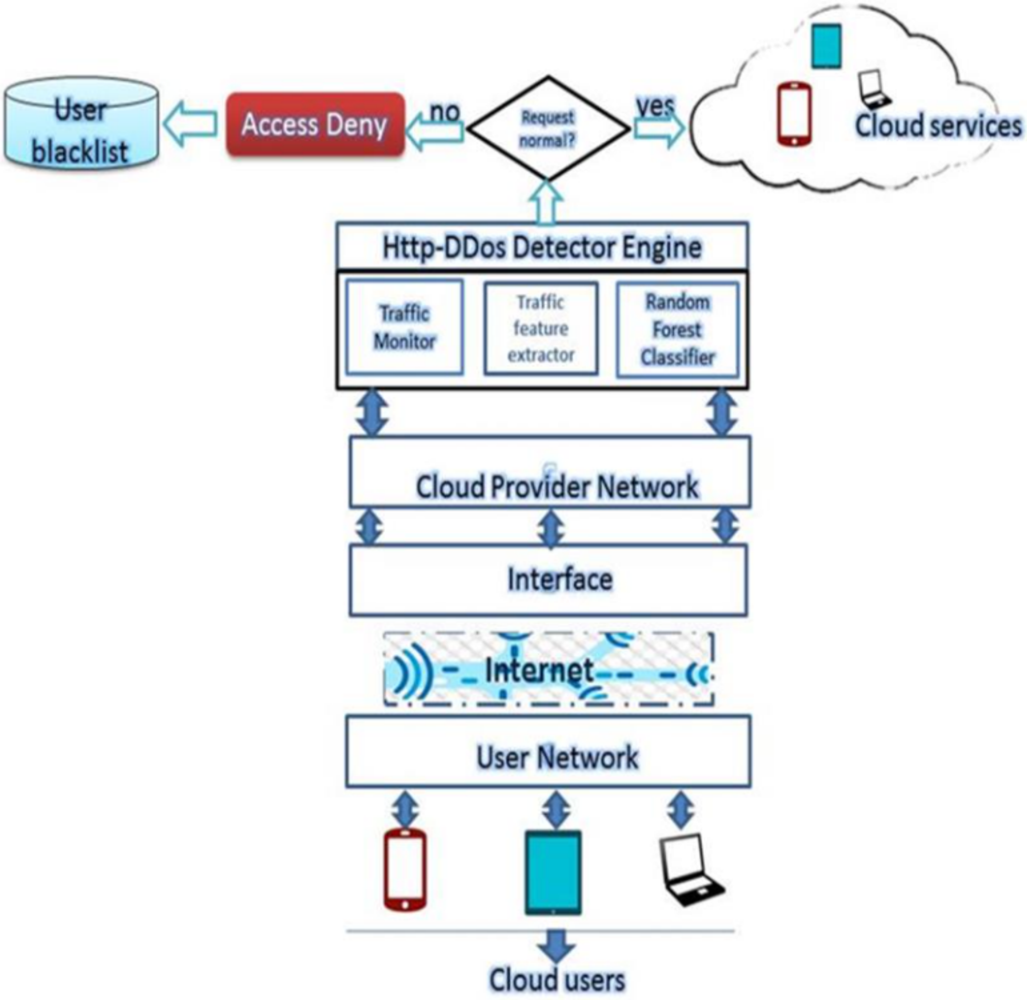

Figure 2. Random forest-based HTTP-DDoS attack detection system framework

Meanwhile, the sub-sections described how each of the components of the designed detection framework works.

I. **Traffic Monitor** - The role of the traffic monitor is to incorporate network sniffing and packet capturing in a network to ensure availability and swift operation. The traffic monitor generally reviews each outgoing and incoming packet for any process or abnormality that can hinder network performance, availability and security before forwarding it to the feature extractor.

II. **Traffic Feature Extractor** - This transform the input data into set of features found on the network packets based on the feature set stored to build derived value to carry out the desired task

III. **Random Forest based Classifier** - Random forest classifier plays the role of analyzing and classifying the received traffic from traffic feature extractor to figure out intrusion before granting access to the cloud information. or forward them to user blacklist database. If the traffic has no feature of HTTP-DDoS attack, then access to the cloud services will be granted, otherwise, there will be a signature database (user blacklist) for future pattern matching.

IV. **User blacklist** - The User blacklist database stores the data that have been classified as malicious by the random forest-based model. Subsequently, incoming traffic will be matched with those in the blacklist database. In doing so, known attacks will be dropped while unknown attack will be filtered by the random forest-based model.

3.2 Proposed Model Formulation

Note that,

$$g(M|\Theta_1),\ g(M|\Theta_k) \qquad (1)$$

is the family classifier with chosen random parameter Θ_k from a random vector Θ.
Given that the training dataset is;

$$D= \{(M_1, N_1),..(M_n, N_n\} \qquad (2)$$

drawn randomly from a possibly unknown distribution $(M_i,N_i) \sim (M,N)$.
and, given a set of possibly features;

$$F= \{f_1\{(M_1), \ldots, fk(M)\} \qquad (3)$$

Using Gini Criterion, we define:
h =attack and n = normal data

If each $C_k(D)$ is a decision tree, then the ensemble is a random forest. We define the parameters of the decision parameter as;

$$\Theta_k = (\Theta_{k1}, \Theta_{k2}, \ldots \Theta_{kp}) \qquad (4)$$

Thus, decision tree k leads to a classifier, $C_k(D) = C(D|\Theta_k)$ (5)

For the final classification $\{ C_k(D|h,n) \}$, each of the instances in the dataset is been classified as either containing an attack or normal.

If,

$D =\{(h_i, n_i)\}$, *for* $i= 1$, we train an ensemble of classifiers $C_k(D)$

Therefore $C_k(D)$ is a predictor of either attack $h= 1$, or non-attack $n= -1$

Hence, we have $Y= \pm 1$ associated with input dataset D. (6)

3.3 Validation and Testing via Experimentation

To effectively evaluate and compare the efficiency of the proposed random forest-based model for detection of HTTP-DDoS attack in cloud computing environment, 7256 (42%) instances of HTTP-DDoS attacks and 10256(58%) of normal traffic were extracted from the actual dataset.

The features were used to feed the Random Forest in Waikato Environment for Knowledge Analysis (WEKA). Six different performance metrics which are; accuracy, precision, recall, f-measure, false positive rate and true positive rate were used in evaluating the results obtained. Figure 3 present the flowchart of the proposed model.

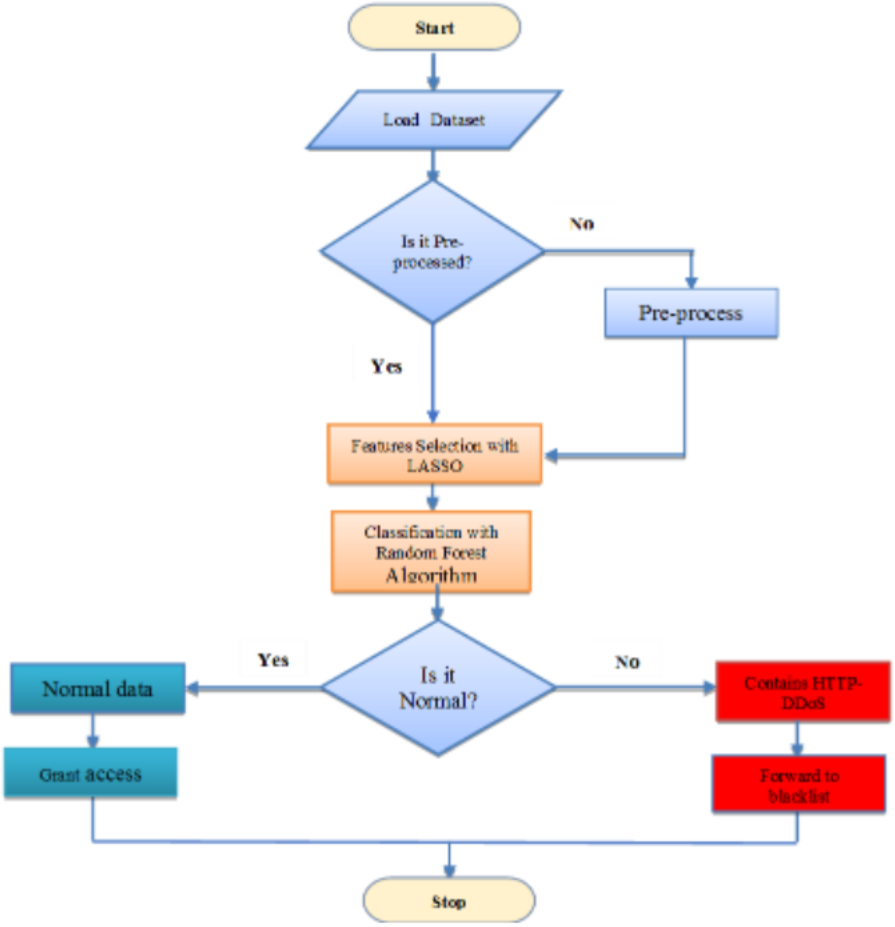

Figure 3. Flowchart of the experimental procedure

3.4 Performance Evaluation

The proposed HTTP-DDoS detection system performance largely depends on the effectiveness of the model formulated. The formulated Random Forest based model performance was evaluated based on accuracy, precision, recall, f-measure, false positive rate and true positive rate.

3.4.1 Accuracy: This is computed as the percentage of the dataset that are correctly classified by the algorithm. Accuracy looks at negatives or positives dependently, other measures for performance evaluation was therefore used.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \tag{7}$$

3.4.2 Precision: This indicates number of instances which are relevantly classified positives. Precision shows great relevance in detection of positives.

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

3.4.3 Recall: indicates the extent in which a system can detect positives**.**

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

3.4.4 F-Measure

$$F - Measure = 2 * \frac{precision*Recall}{Precision+Recall} \tag{10}$$

3.4.5 False Positive Rate: true and false signifies the expectation of the classifier, while positive and negative represents the prediction of the classifier.

$$TN = \frac{TN}{TN+FP} \tag{11}$$

4. Result and Discussion

Random forest classifier was adopted as it has the best performance in all the parameters. From the results analysis in Table 1, Figure 4 and Figure 5, it can be inferred that the random forest-based model outperformed the MLP of (Mouhammd *et al.,* 2016). Also, in line with (Mouhammd *et al*., 2016), it outperformed the MLP of (Irfan, Amit & Vibhakar, 2017). This result also obtains higher accuracy as compared to Random Forest of (Mohammed, Karim & Mustapha, 2018), SVM of (Indraneel & Venkata, 2017) and Naïve Bayes of (Sharmila & Roshan, 2018) with an accuracy and false positive rate of 99.94%. and 0.001 respectively. (Mouhammd *et al*., 2016) compared with other three machine learning algorithms, (Irfan, Amit & Vihakar, 2017) compared with four, (Indraneel & Venkata, 2017) compared with two, (Sharmila & Roshan, 2018) also compared with two while this research work also evaluated with other twelve algorithms. Random Forest based Model for detection of HTTP DDoS attack in Cloud Computing environment performed best.

To further evaluate the performance of this study, comparative analysis of other eleven machine learning algorithms was carried out as shown in Table 2. While Figure 6 and Figure 7 shows the accuracy and FPR performance of the several algorithms compared with Random Forest been adopted in this study.

Table 1. Performance Comparison of this Model with existing Models

SN	Author(s) & year	Machine learning	Attack Type	F-measure	TPR	FPR	Precision	Recall	Accuracy (%)
1	Mohamed, Karim& Mustapha (2018)	Random Forest	HTTP-DDoS	-	-	0.04	-	-	99.54%
2	Mouhammd *et al.* (2016)	MLP	DDoS	-	-	-	0.48%	0.93%	98. 63%
3	Irfan, Amit & Vihakar (2017)	MLP	HTTP DDoS	-	-	-	0.92	0.96	98. 91%
4	Indraneel & Venkata (2017)	SVM &BA	HTTP DDoS	0.9457	0.96	-	0.945	0.94	94. 8%
5	Sharmila & Roshan (2018)	HCA &NB	DDoS	-	-	0.54	-	-	99. 45%
6	Proposed model	Random Forest	HTTP-DDoS	0.999	0.999	0.001	0.999	0.999	99. 94%

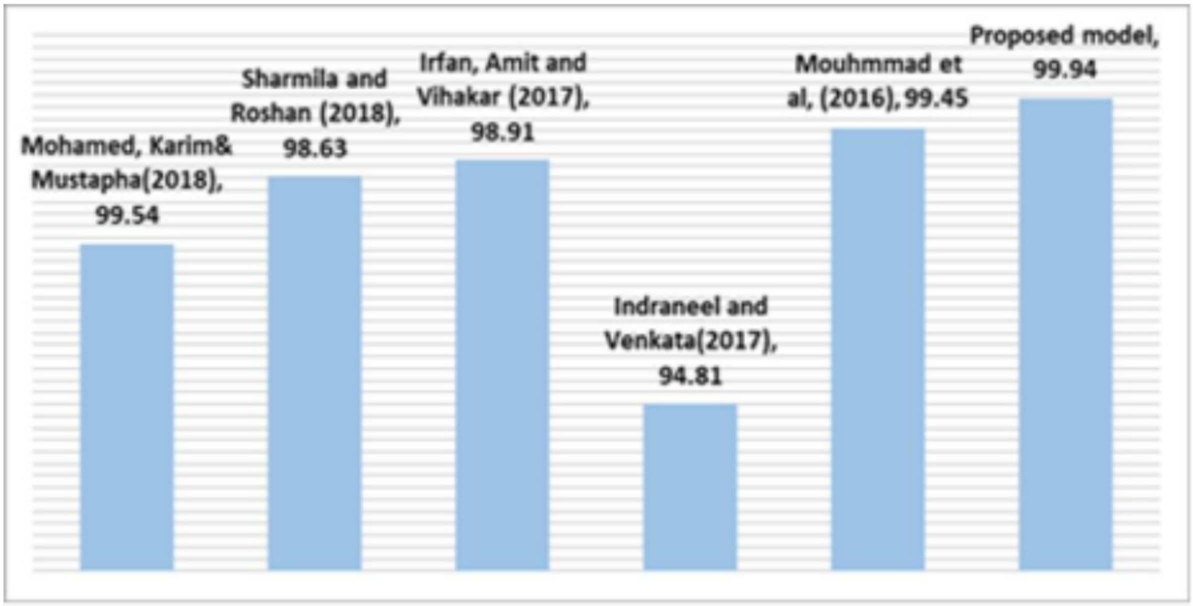

Figure 4. Comparison of accuracy with existing models

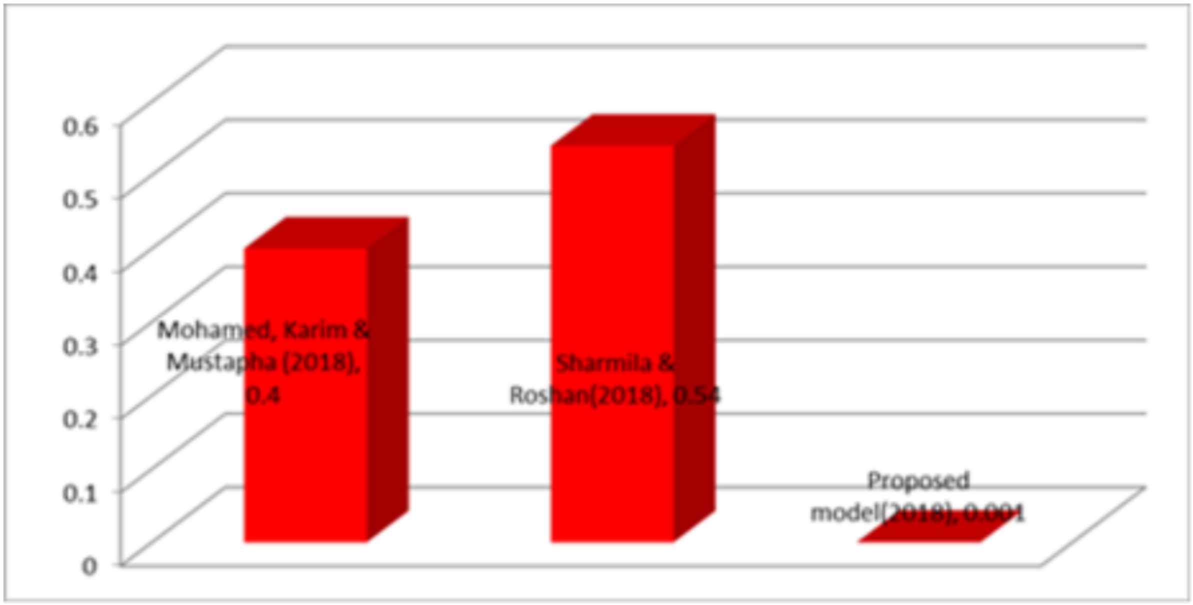

Figure 5. Comparison of FPR with existing models

Table 2. Results of performance evaluation of different machine learning algorithms and Random Forest

Classifier	TPR	FPR	Precision	Recall	F-Measure	Accuracy
Random Forest	0.999	0.001	0.999	0.999	0.999	99.9371
J48	0.994	0.006	0.994	0.994	0.994	99.3713
Naivebayes	0.935	0.056	0.942	0.935	0.935	93.524
IBK	0.999	0.001	0.999	0.999	0.999	99.9057
Kstar	0.991	0.008	0.991	0.991	0.991	99.0883
SMO	0.984	0.015	0.984	0.984	0.984	98.3967
simpleLogistics	0.994	0.006	0.994	0.994	0.994	99.4027
Multilayeperseption	0.995	0.005	0.995	0.995	0.995	99.497
Decision Table	0.995	0.005	0.995	0.995	0.995	99.5285
PART	0.997	0.003	0.997	0.997	0.997	99.7485
NaivebayesSimple	0.946	0.045	0.952	0.946	0.946	94.6226
BayesNet	0.995	0.005	0.995	0.995	0.995	99.4656

Online: ISSN 2645-2960; Print ISSN: 2141-3959

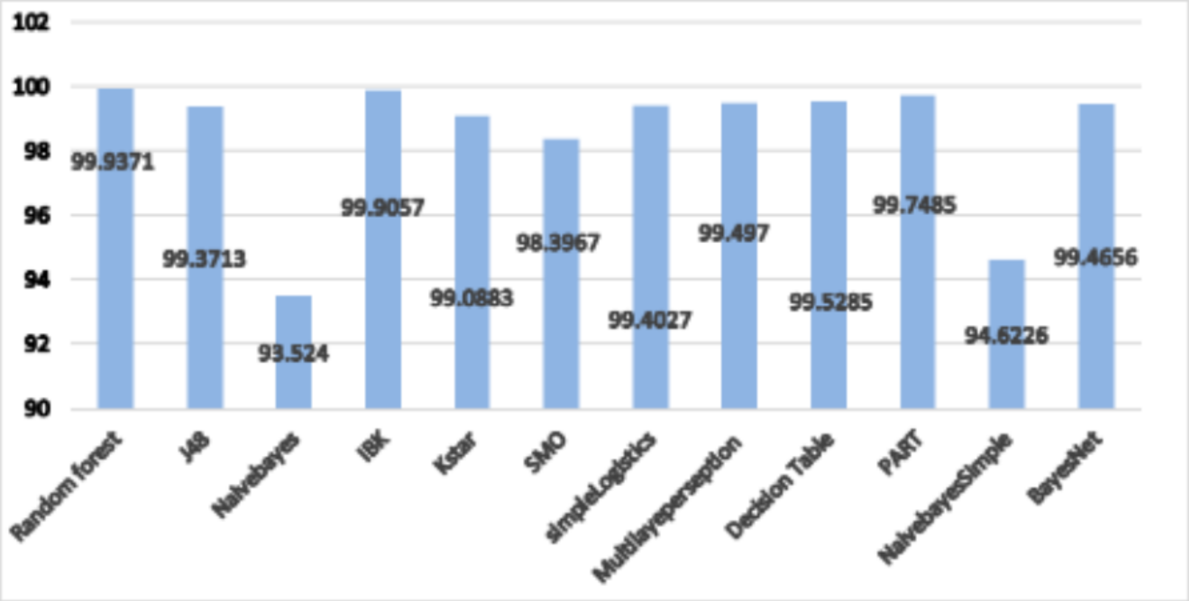

Figure 6. Accuracy results of random forest with other algorithms

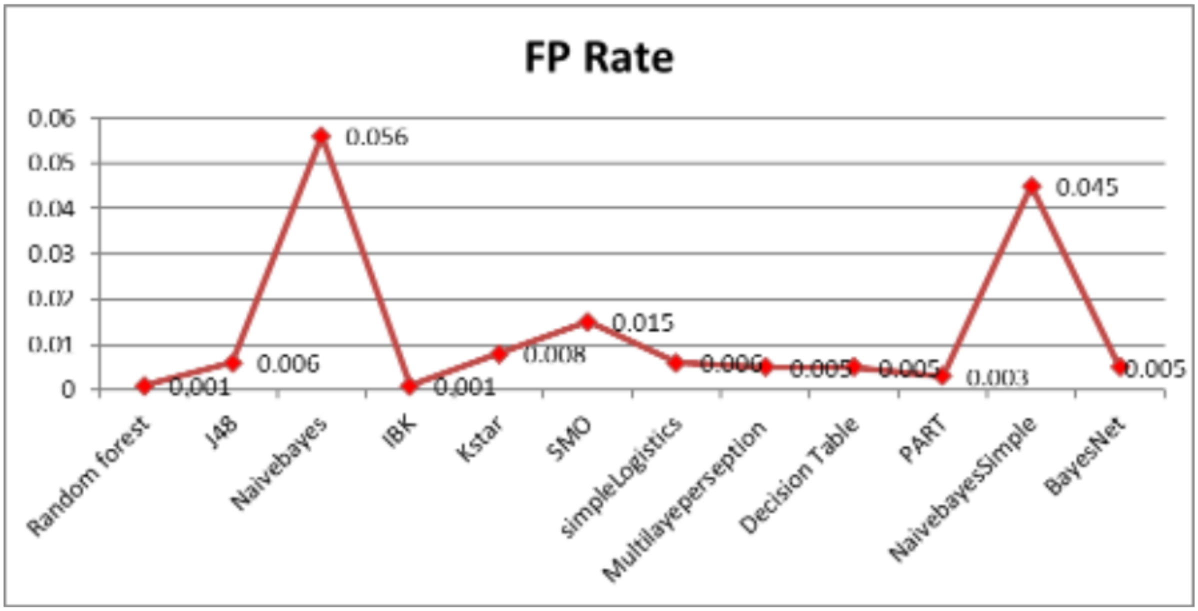

Figure 7. FPR performance of random forest with other algorithms

4. Conclusion and Contribution to knowledge

This paper was able to design a framework for the detection of HTTP DDoS attack in cloud computing environment and adopted random forest in the framework. It also provides a model that reduces the rate of HTTP DDOS attacks success, thereby improving accuracy and reducing false positive rate.

Online: ISSN 2645-2960; Print ISSN: 2141-3959

To further evaluate the performance of the model, eleven machine learning algorithms (J48, Naïve Bayes, IBK, Kstar, SMO, Simple Logistics, Multilayer Perception, Decision Trees, PART, and NaivebayesSimple) were also selected based in literature and applied on the extracted dataset to classify the data as either Normal or HTTP-DDoS. The Random Forest Model achieved the highest accuracy rate with 99.94%, outperforming some of the most recent existing models proposed by Mohamed, Karim & Mustapha, (2018) with 97.5% , Indraneel & Venkata, (2017) with 94.8%, Irfan, Amit, & Vibhakar, (2017) with 98.91%, and Mouhammd *et al*., (2016) with 96%.

Reference

Apale, S., Kamble, R., Ghodekar, M., Nemade, H. and Waghmode, R., (2015). Defense mechanism for ddos attack approaches, methods and techniques. *Journal of Network and Computer Applications*, 57, 71–84.

Araar, A. & Bouslama, R. (2014). A comparative study of classification models for detection in IP networks intrusions. *Journal of Theoretical & Applied Information Technology,* vol. 64, no. 1.

Ardjani, F., Sadouni K., and Mohmed, B. (2010). Optimization of SVM multiclass by particle swarm (PSO-SVM). *Journal of Mod Education Computer Science, 2,32–38.*

Bace, R. G., and Mell, P. (2001). Intrusion detection systems. Gaithersburg, MD: U.S. Dept. of Commerce, Technology Administration, National Institute of Standards and Technology.

Indraneel, S., Venkata, P., & Kumar, V. (2017). HTTP flood attack detection in application layer using machine learning metrics and bio inspired bat algorithm. *Applied computing and informatics* (2017). https://doi.org/10.1016/j.aci.2017.10.003

Irfan S., Amit, M., and Vibhakar M. (2017). Machine learning techniques used for the detection and analysis of modern types of DDoS attacks. *International Research Journal of Engineering and Technology,* 4(6), 1-8.

Liu, H., Hiroshi M., Rudy, S., and Zheng, Z.(2010). Feature selection: An ever-evolving frontier in data mining. *Journal of machine learning research* 10, 4-13.

Mathworks. (2018, June 26). Retrieved from https://www.mathworks.com/discovery/machine-learning.html.

Marinescu, D. (2017). Cloud computing theory and practice. Retrieved from https://www.amazon.com/Cloud-Computing-Practice-Dan-Marinescu/dp/0124046274

Modi, C. N., Patel, D. R., Patel A. and Muttukrishnan, R., 2012. Bayesian classifier and snort based network intrusion detection system in cloud Computing. *In: The third IEEE international conference on computing communication & networking technologies, ICCCNT, Coimbatore, India; 1-7.*

Mohamed, I., Karim, A., and Mustapha, B. (2018). Detection system of HTTP DDoS attacks in a cloud environment based on information theoretic entropy and random Forest. *Security and Communication Networks,* 2018(2), 1-13.

Mouhammd, A., Ghazi A., Ahmad, B.A., and Hassanat, M. A. (2016). Detecting distributed denial of service attacks using data mining techniques. *In International Journal of Advanced Computer Science and Applications,* 7(1), 436-445.

Rup, D., Kausthav, P., Dhruba, K., Bhattacharya, D., and Jugal, K. (2015). Network defense. *Journal of network and computer applications*, 57,71-84.

Sharmila, B., and Roshan, C. (2018). DDoS attack detection using heuristics clustering algorithm and naïve bayes classification. *Journal of Information Security*, 9(1), 33-44.

Sun, D., Guiran. C., Lina, S., and Xingwei, W. (2011). Surveying and analyzing security, privacy and trust issues in cloud computing environments. *Procedia Engineering*, 15 (2011) 2852 – 2856.

Valeria, F., and Eduard, B. (2017, May 21). Feature selection with LASSO. Retrieved from https://beta.vu.nl/nl/Images/werkstuk-fonti_tcm235-836234.pdf