IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Augmenting Network Intrusion Detection System using Extreme Gradient Boosting (XGBoost)

R.Vijay, S.Manoj , V.P.Ravikanth, Y.Vikas, Dr.P.Indira Priyadarshini

Student, Student, Student, Student, Associate Professor
Information Technology
Vardhaman College Of Engineering, Hyderabad,Telangana

Abstract: There is a high rise in the design on the utility of Internet technological innovation functioning day by functioning day. This amazing improve ushers in a huge volume of info produced as well as handled. For apparent great factors, undivided focus is thanks for guaranteeing group security. An intrusion detection system plays an important role in the spot of the earlier mentioned protection. The detection of protection associated operates by utilizing Machine Learning (ML) is extensively investigated. Intrusion Detection Systems (IDSs) are safety aids used to determine malicious activity. Network Intrusion Detection Systems (NIDS) are among most well-known contexts of machine learning software program in the security region. IDSs is generally categorized working for lots of criteria. One of these basic- Positive Many Meanings- crucial components is the detection tactic, around terminology of what IDSs (and NIDSs) is signature-based or anomaly-based often.

The former group detects attacks by checking out the info flow below analysis to patterns stashed at bay inside a signature site of recognized attacks. The later detects anomalies dealing with a sort of typical behaviour of monitored phone system and also flagging activity resting outside of the item as anomalous or suspicious. Signature-based IDSs can determine trendy hits with too much precision but do not recognize or perhaps search for new hits, while anomaly based IDSs have that ability. In this specific task we focus on anomaly-based society intrusion detection by employing XGBoost algorithm on KDD CUP 1999 info positioned to get the ideal outcomes. XGBoost is a fairly recently accessible machine learning strategy that is been operating boosting interest. It gotten Kaggle's Higgs Machine Learning Challenge, about different Kaggle competitions, because of the general functionality of its. The match procedure was constructing a method intrusion detector, a predictive model good at distinguishing between "bad" connections, referred to as intrusions or hits, as well as "good" everyday connections. This specific site features a normal variety of info being audited, incorporating an array of intrusions simulated inside a military neighborhood atmosphere. The very first KDD Cup 1999 dataset offered by UCI Machine Learning repository features forty-one qualities (thirty-four constant, and 7 categorical) and also offers 3,925,651 attacks (80.1 %) outside of 4,898,431 papers. The whole goal is studying the integrity of info and in addition have a much better accuracy in the prediction of info. In that manner, the volume of mischievous info drifting in a method might be reduced, making the network a secured area to transfer info. The more secure a channel is, the less instances where data might get hacked or manipulated.

Index Terms **- XGBoost, Intrusion, Anomaly, Signature.**

I. INTRODUCTION

In today's electric age, preserving data over numerous social media sites along with web businesses remain insecure. Many intruders likewise male together with bot, are receiving unauthorized entry to information. Network Intrusion Detection Systems (IDS) display chaotic performance in determining various attacks [1]. Network Intrusion Detection Systems (NIDS) are probably the best-known contexts of machine learning software program in the protection region [2]. IDSs is generally categorized dealing with numerous requirements [3]. One of these basic- Positive Many Meanings- crucial components is the detection tactic, around terminology of what IDSs (and NIDSs) is signature-based or anomaly-based often. The first one detects attacks by checking out the data that flow with the patterns stashed at bay inside a signature repository site of recognized attacks. The later detects anomalies dealing with a sort of typical behavior of monitored Networks and also alerting the activity resting outside of the item as anomalous or suspicious. Signature-based IDS can determine trendy hits with too much precision but do not recognize or perhaps search for new strikes [4, 5], whereas anomaly grounded IDS have that ability. In this we focus on using ML classifier for detection.

An advanced assailant is able to avoid these strategies, therefore the demand for much more clever intrusion detection is growing by the morning. Scientists are trying to generate ML strategies to this part of cybersecurity. Supervised machine Learning algorithms, when used to historic alert information, could substantially boost classification accuracy and also reduce investigate period for analysts. It is able to augment analysts with extra insights as well as details making much better judgment calls. Although prediction designs based on historical details are able to boost analyst productivity, they won't ever change security analysts entirely.

This particular document acts as instruction to help you shape the improvement of a NIDS methods machine mastering model management strategy. Through appropriate maintenance, tuning, and placement of styles, an unwanted effect to community visitors could be held to a minimum while simultaneously attaining an optimum balance of protection as well as community performance

XGBoost (EXtreme Gradient Boosting) [6] is a recently available decision-tree-based ensemble printer mastering algorithm. XGBoost received Kaggle's Higgs Machine Learning Challenge found 2014. XGBoost is becoming more popular due to the effectiveness of its as well as scalability [2,6,7]. XGBoost is proving quicker than some other famous algorithms in one device and also to scale to vast amounts of illustrations in distributed or maybe memory limited configurations as found empirically in Kaggle tournaments [8]. Kaggle is an internet community centered on machine learning, famous for the tournaments it organizes.

The objective of any NIDS is generating alerts when adversaries attempt to penetrate and attack the system.

This particular paper is going to present a unit whereby different parameters associated with the information are estimated, dependent on what an IDS are created to help you secure the system. Area one comes with a short introduction to the IDS as well as information integrity importance; Section two blankets connected work carried out by different researchers concerning the topic; Section three describes additional background and concept of XGBoost as well as the way the algorithm works in overall; next section presents the experiments conducted and major outcomes attained through the XGBoost algorithm on the KDD Cup 1999 dataset; and finally Section 5 gives the conclusion to the paper.

II. RELATED WORK

There are some functions associated with ours in the feeling that they normally use XGBoost for intrusion detection. Nevertheless, they're rather minimal, e.g., since they think about just the detection of DDoS attacks or even concentrate on the particular situation of Software Defined Networks (SDN).

Chen at al. [9] concentrated on using XGBoost to master to determine a DDoS episode within the controller of an SDN cloud. They utilized the existing 1999 ACM KDD Cup dataset. The writers show that XGboost is able to identify DDoS attacks by examining strike traffic patterns. The XGBoost algorithm indicates greater precision minimizing bogus positive price compared to some other algorithms tried, based on the writers. Additionally, XGBoost proved to become faster compared to the additional algorithms tested.

Bansal as well as Kaur [10] utilized the CICIDS 2017 dataset once more to identify DDoS attacks, not alternative sessions [11]. XGBoost proved again to provide better accuracy values than KNN, AdaBoost, MLP and Naïve-Bayes.

In a brief research paper, Amaral et al. [22] utilized XGBoost to classify malicious visitors of SDNs. The writers demonstrate a structure deployed in an enterprise system which gains traffic information with the OpenFlow process and demonstrate exactly how a few machine learning methods may be used for traffic category, like XGBoost. Visitors of a selection of uses (e.g., BitTorrent, YouTube) was labeled as well as the relative outcomes of XGboost found face of various other classifiers (Stochastic Gradient and random Forests Boosting) proved better or comparable.

III. PROPOSED METHODOLOGY

In this Project we are going to use Extreme Gradient Boosting (XGBoost) Algorithm on KDD CUP 1999 data set. XGBoost is an ensemble method that looks for to form a solid classifier (demonstrate) based on “weak” classifiers. In this setting, powerless and solid allude to a degree of how related are the learners to the genuine target variable. By including models on best of each other iteratively, the blunders of the past show are rectified by the following indicator, until the preparing information is precisely anticipated or replicated by the model. The dataset is at that point separated into preparing and testing sets; the previous to prepare the classification show, and last mentioned to assess the exactness against obscure occurrences.

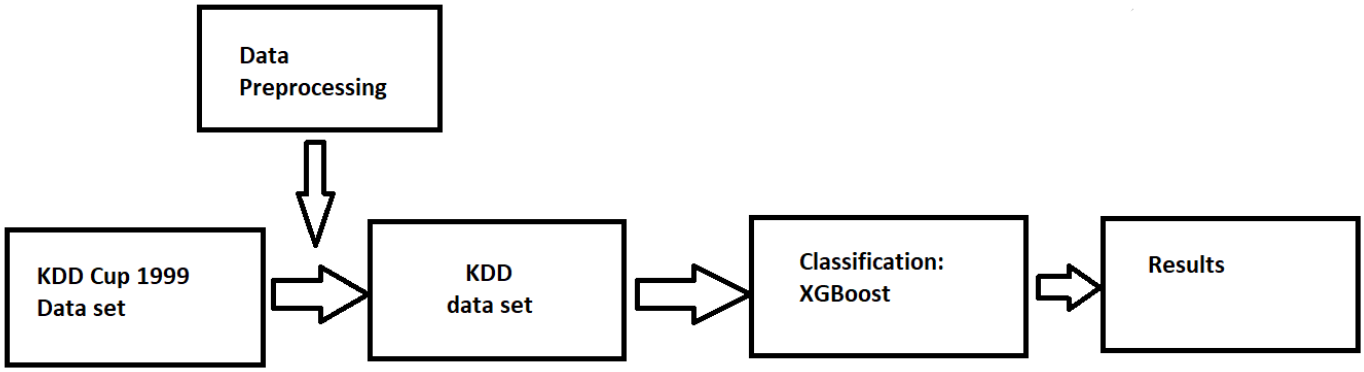

Figure 1. Proposed Flow of work diagram

Proposed XGBoost Algorithm

The proposed XGBoost Algorithm is given below in figure 2

Algorithm: XGBoost

Input: KDD Cup1999 Data set
Step 1: Convert all the non-numerical values present in Data set to numerical values
Step 2: Perform normalization technique on the KDD Cup 1999 data set
Step 3: Divide the Data set into train data and test data
Step 4: Build XGBoost classifier model
Step 5: XGBoost classifier is validated by test data.
Step 6: Classification of NIDS is done based on good and bad connections of data.
Step 7: Finally, the performance of the classifier is evaluated.

Output: Evaluating the results obtained using Accuracy, Precision, Recall, F-score,ROC Curve.

Above Figure depicts suggested XGBoost algorithm. First 2 steps are going to perform preprocessing on information, that is provided as input. Normalization is customarily put on on the dataset. In addition, duplicate data present in data set is eliminated. In level three and four, KDD Cup 1999 Data set is split into a training and testing data. Then a XGBoost design is made. In level five, the model is validated by passing the test data. Next step is utilized for classifying attacks and non-attacks. Lastly, step seven is employed to assess the effectiveness of the classifier.

IV. EXPERIMENTS CONDUCTED & RESULTS OBTAINED

A portion of KDD Cup 1999 data set is for used for our experiment. It's classes with very same proportions as in KDD Cup 1999. It contains 10,230 instances with 1042 Normal instances& 9188 unusual instances. Every class in the KDD Data set are provided with numeric values. They will receive as "0" for Abnormal in addition to "1" for Normal. The forty-one consecutive Features are named as {F1, F2, F3, F4, F5…. F41} and a class label defines “0” for attack instances and “1” for normal instances.

Data preprocessing phase:

Firstly, duplicates are eliminated. Afterward Feature rescaling is achieved for each characteristic separately. Always normalization was put on the information set. It has 1042 "Normal" situations and also 9188 "attack" situations. The KDD Cup 1999 information set is divided as train as well as tests information. The training data is 67% and tests data is 33% from KDD Cup 1999 data set.

After Preprocessing is completed. The tests on the suggested XGBoost algorithm had been done with Python Language in Google Colab. Colaboratory, and "Colab" for brief, is something from Google Research. Anybody is allowed by Colab to create as well as perform arbitrary python code with the internet browser, and it is particularly well suited to machine learning, information evaluation as well as degree. Much more commercially, Colab is a hosted Jupyter notebook assistance that will require zero installation to utilize, while offering access that is free to computing information including GPUs.

The KDD Cup 1999 Data set with forty-two functions (41th characteristic is category label) is given to XGBoost Classifier to establish the model. We conducted the XGboost classifier for network intrusion detection system through the use of all of the default parameters that are identified by the XGBoost designers. Next to look at the functionality of the suggested methodology with XGBoost these 2 distinct datasets have been used: Testing. and Training.

(1) Training dataset: The intent behind the training data set is availed to instruct the cases and also create a model.
(2) Testing dataset: Testing data set aims to assess the performance of ensemble classifier.
Following the model building, the outcomes are captured. Many tests are performed on KDD CUP 1999 information set: 1) Ada boost 2) Gradient Boost 3) SVM 4) KNN 5) MLP 6) Naïve Bayes 7) Proposed XGBoost. The Accuracy rate of six classifiers as well as the suggested XGBoost is compared plus it's discovered that the proposed XGBoost has got the greatest with many with 97.8 %. The delivery of the suggested ensemble classifier is evaluated by using Precision, F1 Score, Recall, Accuracy and Receiver Operating Characteristic (ROC) curve. AUC for XGBoost algorithm is 0.91, and that is identical for Gradient Boost Algorithm. The comparison of Ada boost, Gradient boost, SVM, KNN, MLP, Naïve Bayes classifiers with suggested XGBoost Algorithm are provided with table 1. The Receiver Operating Characteristic (ROC) curve for Ada boost, Gradient boost, SVM, KNN, MLP, Naïve Bayes classifiers with suggested XGBoost Algorithm is depicted in figures 5,6,7,8 respectively below. The Accuracy is 97.8 % and Recall is 0.84 for the suggested algorithm. The various evaluation metrics employed for evaluating are illustrated as.
1) The F1 Score is defined as 2 * (Precision*Recall)/ (Precision + Recall).
2) the accuracy of a classifier is denoted as the fraction of the number of corrected predictions to the complete number of input samples
3) The Area under the Curve is abbreviated as AUC. It maintains on the quality of the classification methodology. It is evaluated as the excellence of the model's estimates irrespective of what classification threshold is taken.
4) Precision is expressed as the success likelihood of generating a correct positive-class classification.
5) A Recall is designated as the model's ability in detection of all the data points of interest in a data set.

Table 1: The Results obtained for proposed XGBoost compared with other classifiers.

S.No.	Classifier Used	Precision	Recall	F1 score	Accuracy(in %)	AUC
1	SVM	0.83	0.71	0.76	95.58	0.84
2	KNN	0.84	0.71	0.77	95.73	0.85
3	Naive Bayes	0.7	0.84	0.76	94.78	0.9
4	MLP	0.85	0.72	0.78	94.97	0.85
5	Gradient Boost	0.93	0.83	0.88	97.71	0.91
6	AdaBoost	0.88	0.81	0.84	97.03	0.9
7	Proposed XGBoost	0.93	0.84	0.88	97.8	0.91

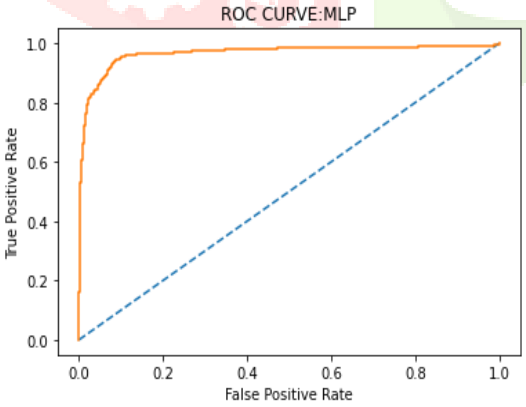

Figure 2: The ROC curve for the MLP classifier

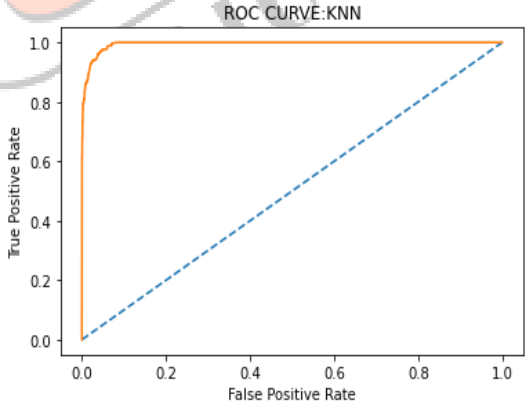

Figure 3: The ROC curve for KNN classifier

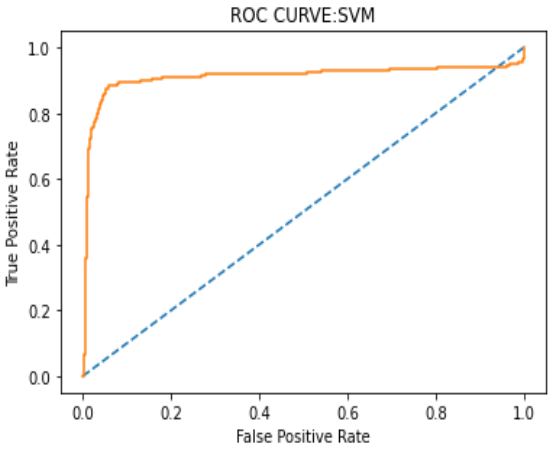

Figure 4: The ROC curve for SVM classifier

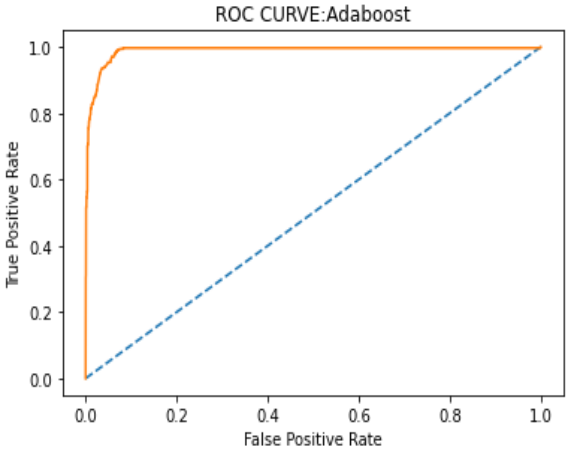

Figure 5: The ROC curve for Adaboost classifier

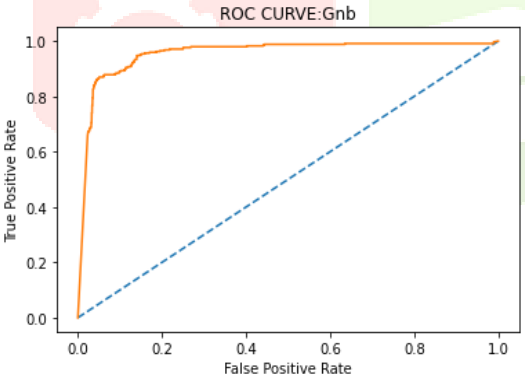

Figure 6: The ROC curve for Naïve Bayes classifier

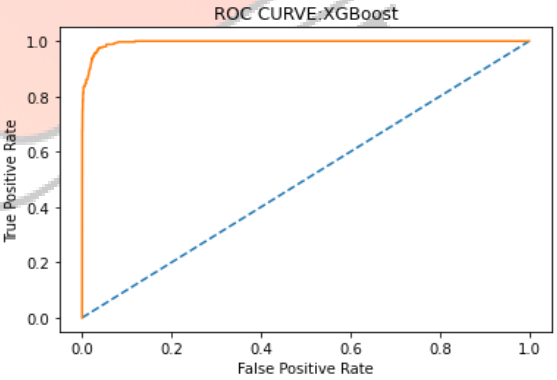

Figure 7: The ROC curve for XGBoost classifier

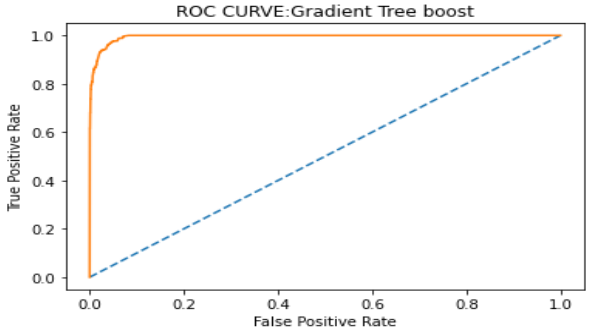

Figure 8: The ROC curve for Gradient Tree boost classifier

```
XGBoost accuracy(in %): 97.80805687203792
Precision: 0.938710
Recall: 0.841040
F1 score: 0.887195
Confusion Matrix:
 [[3011   19]
 [  55  291]]
AUC Score  0.9173849176825197
CV Score   [0.89794721 0.90557185 0.33753666]
```

Figure 9: Accuracy,Precision,Recall,F1 Score of the Proposed XGBoost

V. CONCLUSION

We describe the approach of ours to NIDS utilizing XGBoost. The results indicate that the strategies suggested in the work fulfilling the objectives:

• We show as well as illustrated a highly effective technique of education for XGBoost designs being dependent on just a small portion of the entire parameter area, following very good performance metrics.

• We proved the possibility that XGBoost can be employed in case of Network intrusion detection, in using XGBoost to KDD Cup datasets usually utilized in machine learning analysis.

Network Intrusion Detection System (NIDS) is among the key defense systems to a system against intrusions. To this conclusion, the functionality of IDS could be raised with a latest, systematic, as well as healthy dataset. You will discover not many efforts by the scientists to suggest a complete and efficient NIDS framework for a system (more particularly for contemporary networks including an IoT). Research may be toted in this specific course to propose an effective NIDS framework which can offer total protection against intrusions. The IDS framework must add a mechanism to often upgrade the strike definitions inside a dataset as well as continue on instruction the product together with the updated definitions to help make the unit discover brand new capabilities. This can ultimately enhance the IDS type in detecting zero day strikes and also lessen false alarms. The Learning stage of a Machine Learning primarily based IDS design usually takes a rather long time and could be carried out offline. The secret to strike precision and detection of a unit is based on the continual practice of dataset updating and education for AI based IDS methods.

REFERENCES

[1] R. Sommer, V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in IEEE Symposium on Security and Privacy, pp. 305-316, 2010. DOI: 10.1109/SP.2010.25

[2] García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., Vázquez, E.: Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security 28(1-2), 18–28 (Feb 2009)

[3] Debar, H., Dacier, M., Wespi, A.: A revised taxonomy of intrusion detection systems. Annales des Telecommunications 55(7), 361–378 (2000)

[4] Mitchell, R., Chen, I.R.: A survey of intrusion detection techniques for cyber-physical systems. ACM Computing Surveys 46(4), 55:1–55:29 (Mar 2014)

[5] Wheeler, P., Fulp, E.: A taxonomy of parallel techniques for intrusion detection. In: Proceedings of the 45th Annual Southeast Regional Conference. pp. 278–282 (2007)

[6] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016)

[7] Adam-Bourdarios, C., Cowan, G., Germain, C., Guyon, I., Kégl, B., Rousseau, D.: The Higgs boson machine learning challenge. In: NIPS 2014 Workshop on High-energy Physics and Machine Learning. pp. 19–55 (2015)

[8] Nielsen, D.: Tree Boosting withXGBoost – Why Does XGBoost Win "Every" Machine Learning Competition? Master's thesis, NTNU (2016)

[9] Chen, Z., Jiang, F., Cheng, Y., Gu, X., Liu, W., Peng, J.: XGBoost classifier for DDoS attack detection and analysis in sdn-based cloud. In: BigComp. pp. 251–256. IEEE Computer Society (2018)

[10] Bansal, A., Kaur, S.: Extreme gradient boosting based tuning for classification in intrusion detection systems. In: Advances in Computing and Data Sciences. pp. 372–380. Springer (2018)

[11] Sharafaldin., I., Lashkari., A.H., Ghorbani., A.A.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy. pp. 108– 116. INSTICC (2018)

[12] Petrussenko, D. Incrementally Learning Rules for Anomaly Detection. 2009. Available online: http://cs.fit.edu/pkc/theses/petrusenko09.pdf(accessed on 26 March 2018).

[13] Mahoney, M.V.; Chan, P.K. Learning Rules for Anomaly Detection of Hostile Network Traffic. In Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, FL, USA, 19–22 November 2003.

[14] Mahoney, M.V.; Chan, P.K. Packet Header Anomaly Detection for Identifying Hostile Network Traffic. 2001. Available online: https://cs.fit.edu/mmahoney/paper3.pdf (accessed on 26 March 2018)

[15] Yadav, A.; Ingre, B. Performance Analysis of NSL-KDD Dataset Using ANN. In Proceedings of the 2015 International Conference on Signal Processing and Communication Engineering Systems, Guntur, India, 2–3 January 2015.

[16] M. Rajasekaran and A. Ayyasamy, "A novel ensemble approach for effective intrusion detection system," Second International Conference on Recent Trends and Challenges in Computational Models, 2017. DOI: 10.1109/ICRTCCM.2017.27.

[17] A. Borji, "Combining heterogeneous classifiers for network intrusion detection," in Proceedings of the Annual Asian Computing Science Conference, Springer, pp. 254-260, 2007. DOI: https://doi.org/10.1007/978-3-540-76929-3_24

[18] V. Bukhtoyarov, V. Zhukov, "Ensemble-distributed approach in classification problem solution for intrusion detection systems," Intelligent Data Engineering and Automated Learning-IDEAL, Springer, pp. 255-265, 2014. DOI: https://doi.org/10.1007/978-3-319-10840-7_32

[19] Validated, C. Gradient Boosting Tree vs. Random Forest. Available online:https://stats.stackexchange.com/questions/173390/gradient-boosting-tree-vs-random-forest (accessed on 10 May 2018)

[20] Zhang, F. Multifaceted Defense against Distributed Denial of Service Attacks: Prevention, Detection and Mitigation. 2012. Available online:https://pdfs.semanticscholar.org/eb7d/4c742f6cd110d9c96a08e398cc415c3a8518.pdf(accessed on 26 March 2018).

[21] Fu, Z.; Papatriantafilou, M.; Tsigas, P. CluB: A Cluster-Based Framework for Mitigating Distributed Denial of Service Attacks. In Proceedings of the 2011 ACM Symposium on Applied Computing, Taichung, Taiwan, 21–24 March 2011.

[22] Amaral, P., Dinis, J., Pinto, P., Bernardo, L., Tavares, J., Mamede, H.S.: Machine learning in software defined networks: Data collection and traffic classification. In: IEEE 24th International Conference on Network Protocols (ICNP). pp. 1–5 (Nov 2016)

[23]Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A., 2009. A detailed analysis of the kdd cup 99 data set, in: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, IEEE. pp. 1–6