Decision Tree Based Classifications on CICIDS 2017 Dataset for the Identification of DDoS, Botnet, and Web Attack

Reshamlal Pradhan
Department of Computer Science, PSSOU Chhattisgarh, Bilaspur, India
reshamlalpradhan6602@gmail.com

***Abstract*—**
Today's workforce is reliant on the internet. Internet use is growing rapidly with popularity, and so are malicious online activities. A group of researchers are focusing their efforts on intrusion detection systems (IDS). The security for a network is provided by IDS. In a computer network, it is the process of identifying various security violations. Intrusion detection systems (IDSs) are surveillance tools that guard against hostile behavior on computer networks. In this study decision tree-based classification is used for identifying intrusions on the CICIDS2017 dataset, which includes a range of attacks. For the experimental results Weka machine learning tool is employed. The findings of empirical studies demonstrate that excellentclassification accuracy based on decision tree approaches is achieved in the detection of DDoS, Botnet, and Web attack.

***Index Terms* - Intrusions, IDS, CICIDS2017, Classification, Decision Tree Techniques, J48, REP Tree, Random Tree.**
DOI Number: 10.48047/NQ.2022.20.12.NQ77771 NeuroQuantology2022;20(12): 4468-4475

I. INTRODUCTION

Information or network security has recently been a crucial concern for every organization to safeguard data and information on their computer system or internet against different attack types. It's not so easy for the world's top specialists to completely comprehend the innermost workings of cyberspace due to the internet's increased complexity. Every year, personal computers and networking technology are becoming more rapid. It's easy to see how a malevolent user may carry out adverse behaviors in these circumstances without being confronted or even discovered. A large number of consumers' favorite security solutions, such as firewalls and anti-virus software, were created by security professionals to make up for the Internet protocols' security issues. Intrusion detection system (IDS), a topic of interest for researchers, is becoming more and more popular [17]. By identifying an intruder's behavior, an intrusion detection system can inform users in advance of attacks or infiltration attempts. Monitoring tools called intrusion detection systems (IDSs) protect against malicious activities on a system. The term "intrusions" refers to attempts to compromise the confidentiality, integrity, accessibility, or circumvent security mechanisms of a computer system or network. Intrusion detection systems (IDSs) automate the procedure of tracking the events occurring in a computersystem or network and analyzing them for symptoms of security issues [16]. For intrusion detection, a variety of machine learning approaches are utilized, including

classification, prediction, etc. One of the most popular uses of data mining is classification, which involves supervised grouping of samples of a similar kind.

In the current technological era, many research institutes and organizations deal with an extensive amount of organized and unstructured data. These organizations and researchers have difficulties in more precisely and effectively detecting intrusions (or assaults). When dealing with various kinds of data, data mining techniques are crucial. Machine learning security measures are identified and ensured through classification. The effectiveness of intrusion detection and prevention systems is required due to a number of computer security risks [10].

In this study, for the experimental and research purpose, Decision tree based classification techniques are used on CICIDS2017 dataset. CICIDS2017 dataset is very recent benchmark dataset consisting variety of novel attack types (intrusions) [5]. Hence the study on intrusions of CICIDS2017 dataset is of prime concern for theresearchers.

In the further sections of the study, Classification Techniques, Related works, Methodology, CICIDS 2017 dataset and Results are discussed.

II. CASSICATION

One of the most popular uses of data mining is classification, which groups together samples of a similar kind in a supervised way. The process of classifying data has two steps: learning, in which a classification model is created, and classifying, in which the model is applied to forecast class labels for supplied data. Through a series of decisions, classification is accomplished. This knowledgecan be seen as being represented by the tree structure's decisions. Multiple branches are used to transport the data samples from the root to the leaf node. CART, J48, Random Tree, REP Tree, and more decision tree approach types are available [11].

CART is one of the methods for building decision trees that is most frequently utilized in the machine learning community. CART constructs a binary decision tree by splitting the record at each node according to the role of a single attribute. CART calculates the ideal split using the Gini index. It tries to split apart each node similarly to how the root node was created by the initial split. We once more search all input fields for candidate splitters. If no split can be found and doing so severely diminishes the variety of that node, we label a node as a leaf node. The decision tree is finally developed to the point where only leaf nodes are left [12].

One quick decision tree learner is REP Tree. It builds decision trees using information gain as the division criterion and prunes them using approaches that minimize error. Even with large test and training data sets, the error reduction produces a classification tree that is more accurate [13] [18].

The J48 Decision Tree Classifier classifies a newobject using the fundamental approach. The provided training data's attribute values must first be used to build a decision tree. As a result, it searches for the characteristic that best distinguishes the various records whenever it encounters a group of objects (training set). This trait is thought to have the most information gain. Since it can indicate how close data examples are to one another, it allows us to classify them as accurately as possible [10] [12] [15].

A tree-building method called random tree considers K random properties at each node. This approach doesn't involve any pruning. Every tree has a probability of getting sampled in Random Tree due to the uniform distribution of trees [18].

III. RELATED WORKS

Since 1980, there has been ongoing research in the fields of machine learning and data mining. As networking services get more advanced technologically, new security issues arise every year. Researchers areattempting to pinpoint security issues and develop the right models and approaches to address them. Data mining is one of the main fields of research for computer scientists in the current environment.

Katkar & Kulkarni (2013) presented a method to detect DOS attacks using Ensemble of Classifiers. Naive Bayesian (NB), Bayesian Network (BN), Sequential Minimal Optimization (SMO), J48 (C4.5), and Reduced Error Pruning Tree (REP Tree) are the classifiers that were employed. The end result demonstrates that the ensemble of REP Tree, Bayesian Network, and J48 classifiers, without any pre-processing of the data, significantly increased accuracy with the least amount of resources. It has also been demonstrated that using an ensemble of these classifiers can produce extraordinarily high accuracy compared to building a brand-new classifier [1].

Bolón-Canedo et al (2011) proposed a method which uses a lesser collection of features while maintaining the classifiers' performance outcomes. Its foundation is an algorithmic blend of discretization, filtering, and classification. It has been especially applied to intrusion detection on the benchmark KDD Cup 99 dataset. The suggested approach, which is based on classifiers like C4.5, is applicable to huge databases due to the advantages of these machine learning algorithms, which include being faster and more computationally efficient. The comparative study demonstrates that the proposed technique outperformed the findings of the other authors, particularly in the classes where detection was more difficult [2].

Using the NSL-KDD dataset, Kalyani & Lakshmi (2012) compared classification methods such Naive Bayes, J48, OneR, PART, and RBF Network algorithm. Also mentioned the benefits of the NSL-KDD dataset over the KDDCUP'99 dataset [3].

Masarat et al (2014) introduced a multi-step IDS methodology. Although KDD provides a wide range of attributes, not all of them are applicable to classification tasks. The option to select features based on gain ratio is offered. J48 trees are trained with the best features feasible using a Roulette Wheel based on feature gain ratios, which promote best features in random feature selection. The last stage involves ensemble classifiers with fuzzy weighting. The fuzzy weighted combiner weighs the price and effectiveness of classifiers. Results show that the suggested approach works better than rival methods [4].

To evaluate the performance of suggested intrusion detection and intrusion prevention systems, Sharafaldin et al (2018) used a number of datasets, including DARPA98, KDD99, and ISC2012 etc. Their research of 16 datasets from 1998 shows that many are out-of-date and unreliable. There are a number of problems, including a lack of variety and volume in the traffic, inadequate attack coverage, inaccurate payload and packet information from anonymized packets, and the absence of some feature sets and metadata. The study focused on CICIDS2017, the most recent IDS dataset that is publically available, fits industry standards, and covers benign and seven common attack network patterns. In order to establish the best combination of features for identifying a particular attack type, it also evaluates the effectiveness of a set of network traffic attributes using machine learning techniques [5].

Almseidin et al (2017) evaluated the effectiveness and performance of their investigations using J48, Decision Table, Random Tree, MLP, Naive Bayes, and Random Forest. In all of the studies, the KDD intrusion detection dataset was utilized. The rate of various attacks in the KDD dataset is roughly 2% for other sorts of attacks (R2l, U2R, and PROBE), 79% for DOS attacks, and 19% for routine packets [6].

Choudhury & Bhowal (2015) stated that data mining can be useful in the creation of a network intrusion detection system. Data mining is used for extracting informational value from vast volumes of data. The proposed essay looked at various machine learning methods and classification strategies to classify network traffic. Various decision tree techniques, i.e. J48, REP Tree, Random Forest, and Random Tree were determined to be suitable among the classification algorithms examined. Machine learning techniques such as stacking,

boosting, and bagging have all been researched for accuracy. The algorithms were compared using WEKA. 10-fold cross validation was used to simulate these categorization models. A data set based on NSL-KDD was used with WEKA for this simulation. Additionally, the machine learning algorithms were contrasted, with Boosting emerging as the most successful [7].

The Kyoto 2006+ dataset, a brand-new labelled network dataset, was used by Sahu & Mehtre (2015). Every moment in the Kyoto 2006+ data set is classified as normal (no attack), attack (known attack), or unknown attack. A technique called Decision Tree (J48) is used to categorize network packets that can be used by NIDS. There were 134665 network instances deployed for testing and training. The created rules successfully identify connections between no attack, known attack, and unknown attack with an accuracy rate of 97.2% [8].

The most widely used data mining method frequently used for classification is decision trees. The strategy that was presented concentrated on several currently utilized feature selection strategies to exclude unimportant characteristics from the NSL-KDD data set to construct a reliable classifier that will be both computationally effective and efficient. Info Gain, Correlation, Relief, and Symmetrical Uncertainty are four alternative feature selection strategies that are paired with the C4.5 decision tree technique to create IDS. The results show that C4.5 with Info Gain feature selection technique produced the highest accuracy of 99.68% with17 features, but that Symmetrical Uncertainty with C4.5 also produced promising accuracy of 99.64% with 11 features [9].

The illustrations provided from the literature show how different authors used various independent machine learning and decision tree techniques. They trained their model using a variety of datasets, including KDD Cup 1999 and NSL KDD. One of the consistent findings from all of the aforementioned study is the improvement in terms of various computational parameters, i.e. accuracy,memory and computation time, or reduced features.

IV. METHODOLOGY

In this research, work has been explained in terms of Classification model, CICIDS2017 dataset and K-fold Cross validation method.

A. Classification Model

Data classification involves two steps: learning, during which a classification model is developed, and classifying, during which the model is used to predict class labels for data that has been provided. Model isdepicted in the fig. I.

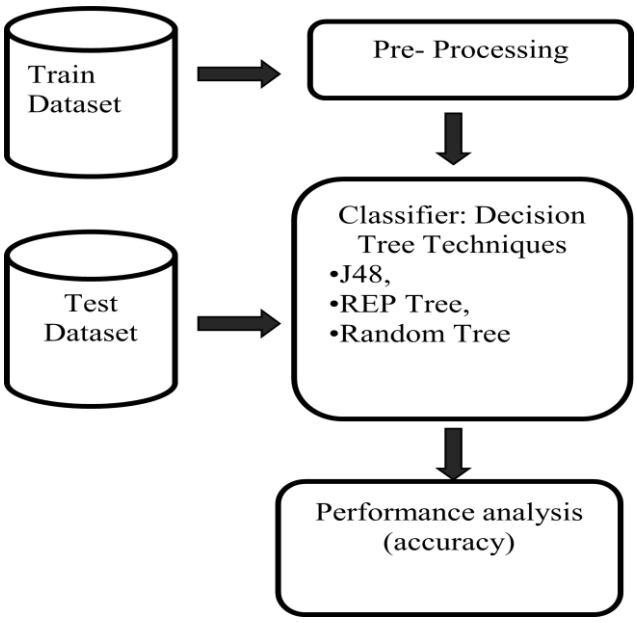

Fig. I Classification Based on Decision Trees

In learning phase a classification model can be created with the help of the training set of data. A training set of data comprises of labeled data with precisely specified data definitions for average and attack-type data. As a result, these data are utilized to develop models. In classification phase models are validated using the test set of data after they have been built using the training setof data. Unlabeled data make up the testing set of data. This test batch of unlabeled information is classified by a trained model. The model is validated with the results of the classification. Here Decision tree techniques J48, REP tree and Random tree

are used in the model evaluation process and performance is measured in terms of accuracy.

B. CICIDS 2017 dataset

The CICIDS2017 dataset was generated by the Canadian Institute of Cyber Security. This dataset closely resembles real-world data because it contains both information on fresh malware attacks and information about benign behaviors. 2830740 occurrences total are in the dataset. It is in CSV format and has about 80 features for each netflow record. About 11 dataset features that have to be present in all datasets are met by CICIDS 2017. According to Sharafaldin et al these include attack variety, labeled dataset, complete interaction, feature set, anonymity, complete capture, complete traffic, available protocols, heterogeneity, complete network configuration, and metadata [5].

The findings of network traffic analysis using CICFlowMeter are included in CICIDS 2017 together with labeled flows that include a time stamp, source and destination IP addresses, source and destination ports, protocols, and attacks. The CICIDS2017 dataset containsdata on attacks as well as traffic statistics over five days. The traffic statistics on Monday are generally benign. The attacks (intrusions) utilized include Web Attack, DoS, Botnet, and DDoS etc. Every assault took place between Tuesday and Friday.

In this study DDoS, Botnet and Web attacks are considered for the model evaluation process with Decision Tree Techniques.

C. K-fold Cross Validation

The decision tree-based models are tested and evaluated using the K-fold cross validation technique. The data set is partitioned into K subgroups for the K- fold cross validation process. The training set is made up of the remaining k-1 subsets, and one of the K subsets is always utilized as the test set. All K trials are combined to obtain performance statistics [14].

V. RESULTS

In this research, Decision tree techniques J48, REP tree and Random tree are used in the model evaluation process and performance is measured in terms of accuracy. Weka machine learning tool is used for the performance measurements. The performance of the classifiers is compared according to the accuracy parameter of confusion metrics. DDoS, Botnet and Web attacks are identified in the CICIDS 2017 dataset.

Decision trees are applied to the dataset to detect DDoS intrusions. Results are demonstrated in the table I given below. It denotes that the decision trees are results in good accuracy with computational time.

Table I. DDOS INTRUSION IDENTIFICATION WITH DECISION TREE

Decision Tree	**DDoS attacktype**	
	accuracy	***Computation time(Sec)***
J48	99.90	17.23
Random Tree	99.92	2.84
REPTree	99.94	7.47

With accuracy REPTree is providing best result, while with computation time Random tree is performing well. A graph plotting is given in the fig.II.

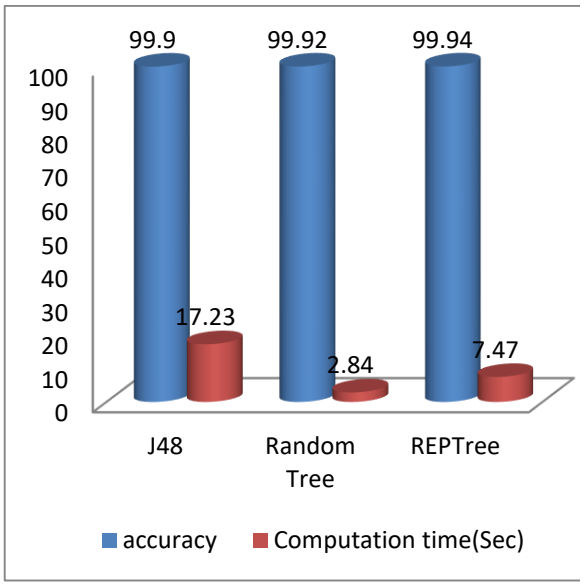

Fig. II DDoS Intrusion Identification with Decision Tree

Web attacks in the dataset are identified with thedecision trees. Results demonstrated in the table II given below. Decision trees are results in good accuracy with computational time.

Table II. WEB ATTACK IDENTIFICATION WITH DECISION TREE

Decisio nTree	**Web attack**	
	accuracy	***Computation time(Sec)***
J48	99.71	8.95
Rando m Tree	99.53	0.66
REP Tree	99.67	5.38

With accuracy J48 decision tree is providing best result, while with computation time Random tree is performing well. A graph plotting is given in the fig.III.

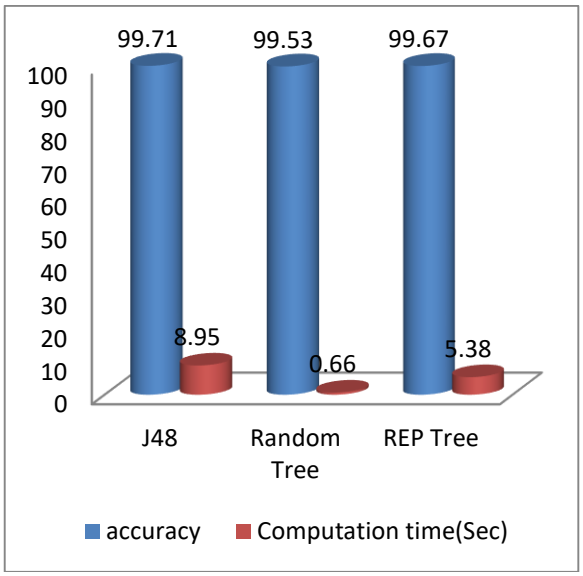

Fig. III Web attack identification with decision tree

Dataset is classified with the decision trees to detect Botnet attacks. Results are demonstrated in the table III. It shows that the decision trees are results ingood accuracy with computational time.

Table III. BOTNET ATTACK IDENTIFICATION WITH DECISION TREE

Decisi on Tree	**Botnet attack type**	
	accuracy	***Computatio n time(Sec)***
J48	99.94	3.98
Rando m Tree	99.91	0.64
REP Tree	99.89	3.14

With accuracy J48 Tree is providing best result, whilewith computation time Random tree is performing well. A graph plotting is given in the fig.IV.

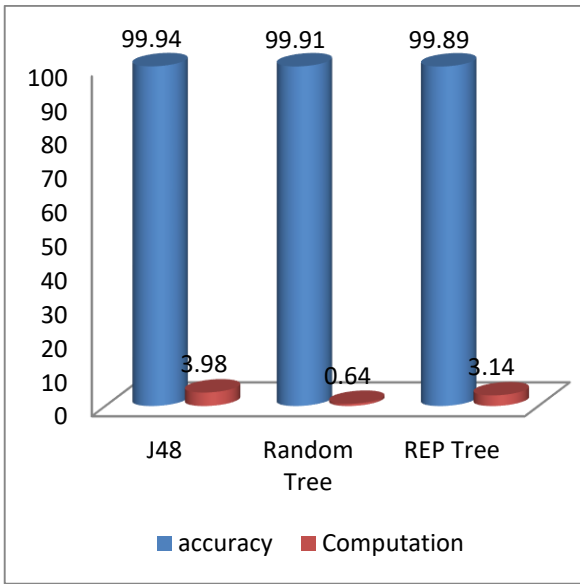

Fig. IV Botnet attack identification with decision tree

VI. CONCLUSION

In this paper, Decision tree based classification model to detect intrusion on CICIDS2017 dataset is presented to explore security issues and data analysis. The performance of different decision tree is tested on the CICIDS 2017 dataset. It is observed that decision tree techniques are maintaining good results. For DDoS attack type, J48 decision tree is providing best results, while for Botnet and Web attack REP tree is maintaininggood results. Related work in the field is also presented. Related work in the field and result analysis on CICIDS 2017 dataset depicts that decision tree perform well. CICIDS 2017 dataset is a high volume dataset consisting of 80 attributes with various attack type. In present studyDDos, Botnet, and Web attacks are explored, other attack types are also needs to be explored, which resembles real type of attacks. Advancement in machine learning techniques is also desirable to achieve highest accuracy.

REFERENCES

[1] Katkar, V.D., & Kulkarni, S.V. (2013). Experiments on Detection of Denial of Service Attacks using Naïve Bayesian Classifier. International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), IEEE(2013).

[2] Bolón-Canedo, V., Marono, N.S., & Betanzos, A.A. (2011). Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. Expert Systems with Applications 38 (2011) 5947–5957, Elsevier.

[3] Kalyani, G., & Lakshmi, A. J. (2012). Performance assessment of different classification techniques for intrusion detection. Learning, 2(1), J48.

[4] Masarat, S., Taheri, H., &Sharifian, S. (2014, October). A novel framework, based on fuzzy ensemble of classifiers for intrusion detection systems. In 2014 4th international conference on computer and knowledge engineering (ICCKE) (pp. 165-170). IEEE.

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

[6] Almseidin, M., Alzubi, M., Kovacs, S., &Alkasassbeh, M. (2017, September). Evaluation of machine learning algorithms for intrusion detection system. In 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY) (pp. 000277- 000282). IEEE.

[7] Choudhury, S., &Bhowal, A. (2015, May). Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection. In 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM) (pp. 89-95). IEEE.

[8] Sahu, S., &Mehtre, B. M. (2015, August). Network intrusion detection system using J48 Decision Tree. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 2023-2026). IEEE.

[9] Hota, H. S., &Shrivas, A. K. (2014). Decision tree techniques applied on NSL-KDD data and its comparison with various feature selection techniques. In Advanced Computing, Networking and Informatics-Volume 1: Advanced Computing and Informatics Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI-2014) (pp. 205-211). Springer International Publishing.

[10] Nejad, A.F., Kharazmi, S., & Bayati, S. (2008). Improving Admission Control Policies in Database Management Systems, Using Data Mining Techniques. International Conference on Computer Science and Software Engineering (ICCSSE) 2008.

[11] Tsai, C., Hsu, Y., Lin, C., & Lin, W. (2009). Intrusion detection by machine learning: A review. Expert Systems with Applications, Elsevier, 36 (2009) 11994–12000, 0957-4174/ 2009.

[12] Pujari, A. K. (2001). Data mining techniques. 4th edition, Universities Press (India) Private Limited.

[13] Belouch, M., Hadaj, S.E. & Idhammad, M. (2017). A Two-Stage Classifier Approach using RepTree Algorithm for Network Intrusion Detection. (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 8, No. 6, 2017.

[14] Mukherjee, S., & Sharma, N. (2012). Intrusion Detection using Naive Bayes Classifier with Feature Reduction. Procedia Technology, 2012.

[15] Revathi, S., & Malathi, A. (2013). A detailed analysis on NSL-KDD Dataset using various machine learning techniques for intrusion detection. International Journal of Engineering &Technology (IJERT),ISSN: 2278-0181,vol. 2 issue 12. December-2013.

[16] Anand, A., & Patel, B. (2012). An Overview on Intrusion Detection System and Types of Attacks It Can Detect Considering Different Protocol. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 2, Issue 8, August 2012.

[17] Ambusaidi, M.A., & Nanda, P. (2016). Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm. IEEE Transactions On Computers, VOL. 65, NO. 10, OCTOBER 2016.

[18] Kalmegh, S. (2015). Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news. International Journal of Innovative Science, Engineering & Technology, 2(2), 438-446.