Machine Learning Based Hybrid Approach to Detect Intrusion in Cyber Communication

Neha Pathak[1] Bobby Sharma[2]
Department of CSE
School of Technology
Assam Don Bosco University
Guwahati, India

Abstract
By looking the importance of communication, data delivery and access in various sectors including governmental, business and individual for any kind of data, it becomes mandatory to identify faults and flaws during cyber communication. To protect personal, governmental and business data from being misused from numerous advanced attacks, there is the need of cyber security. The information security provides massive protection to both the host machine as well as network. The learning methods are used for analyzing as well as preventing various attacks. Machine learning is one of the branch of Artificial Intelligence that plays a potential learning techniques to detect the cyber-attacks. In the proposed methodology, the Decision Tree (DT) which is also a kind of supervised learning model, is combined with the different cross-validation method to determine the accuracy and the execution time to identify the cyber-attacks from a very recent dataset of different network attack activities of network traffic in the UNSW-NB15 dataset. It is a hybrid method in which different types of attributes including Gini Index and Entropy of DT model has been implemented separately to identify the most accurate procedure to detect intrusion with respect to the execution time. The different DT methodologies including DT using Gini Index, DT using train-split method and DT using information entropy along with their respective subdivision such as using K-Fold validation, using Stratified K-Fold validation are implemented.

Keywords:
Cyber security, UNSW-NB 15 Dataset, Decision Tree, Gini Index, Information Entropy, Cross-validation, Accuracy, Execution Time.

1. INTRODUCTION

Cyber security is the massive evolution in the protection from numerous advanced attacks that has spread over in the era of internet. It is a set of processes that are designed to protect personal, governmental and business data from being misused [5]. The great demand of cyber security deals a great attention on exceptional progress in safeguarding the programs and network data. The rapid evolution of social network, cloud and web technologies [2] has faced to many security measures. Cyber security techniques are composed of computer and network security system [7] which has enhanced the integrity to protect the data from the susceptible cyber-attacks and threats [3].

The Intrusion Detection System (IDS) is a part of cyber security integral, as it has the ability to detect new misuse attacks [9] in the network and the system administrator from malicious activities and protect from it. IDS works in two levels, one in the individual machine as a host known as a Host Based Intrusion Detection System (HIDS) and another level monitors the malicious attacks can conduct the confidentiality, integrity or violation [6] in the network named as Network Based Intrusion Detection System (NIDS) [20].

The new concept of Data Science application has advanced in analytic techniques and methods to prevent the emerging sophistication of cyber security attacks that can be adopted in real time data. The application comprises of Artificial Intelligence (AI) with its offshoot such as machine learning (ML), deep learning (DL), etc., [21]. ML is the connecting bridge between AI and computational statistics that gives advantage for the computer to learn more from the given datasets. [2] ML application challenges the methodological and theoretical way of handling [4] to resolves the unknown cyber security situations that are becoming increasingly significant in a potential manner [6]. The new automated ML techniques resolve the problem of solving by the traditional human analyst [8] or conventional signature-based system [3] as it has the capability to learn from the past experience for better future performance [18]. ML models can be broadly classified into supervised, unsupervised and semi-supervised learning. In supervised learning perspective, any labeled data can be used for training and map the new output whereas in the unsupervised, the learning is done from unlabeled dataset where there is exploration, clustering of data. But the semi-supervised is the act of two learning technique for high achievement. The methods undergoes through three

Manuscript received November 5, 2023
Manuscript revised November 20, 2023
https://doi.org/10.22937/IJCSNS.2023.23.11.24

main stages: training, validation and testing [7]. The collection of raw security data sets can be effective in analyzing, with the use of tools for pre-processing classification, regression, clustering, association rules, and visualization [17] the feature which helps in building learning based security model that provides productive services [1].

2. LITERATURE REVIEW

In this section, there is a brief discussion of the related works on some of the NIDS datasets and the learning algorithms:

Alqahtani et al. [10] uses of different machine learning algorithms for intrusion detection from KDD'99 cup datasets. The datasets have classified four main cyber-attacks. The effectiveness is measured by evaluating the performance metrics, precision, recall, f1-score, and accuracy. The experiment shows a comparison among the machine learning algorithms. With numerous kinds of attacks the machine learning algorithms for intrusion detection system (IDS) may vary in the accuracy or sometimes in other prediction area, which may show the performing ability for some is less compare to other classifiers.

Sabar et al. [11] discussed the idea of using the new proposed framework Hyper-Heuristic with the traditional Support Vector Machine (SVM) algorithm. The traditional Kernel SVM model has bi-objection formulation showcasing the accuracy and complexity of the model as conflicting objectives. The new framework consist two levels namely, high-level strategy and low level heuristics(LLHs).The Hyper-Heuristic Support Vector Machine (HH-SVM) model aims to establish the benefits with different effects of using multiple LLHs on the search performance. The HH-SVM even compared with other ML algorithms to check the accuracy performance.

Moustafa et al. [12] proposed the use of new network intrusion detection system (NIDS) dataset, UNSW-NB 15[22] that challenges the existing threats in the cyber security field. The existing benchmarks datasets has limited amount of attacks and information packets that was produced a decade ago which can be an obstacle in the platform of research. To overcome such limitation of previous datasets, UNSW-NB15 is developed IXIA Perfect Storm tool, by the cyber security research lab of the Australian Centre for Cyber Security (ACCS) at UNSW in Canberra. NIDS dataset quality points out two major features that are a comprehensive demonstration of modern threats and the new range of traffics. KDDCups99 dataset lacks the first feature of NIDS dataset. Though NSLKDD is the improve version but it lacks in pointing out the zero-footprint attacks. UNSW-NB15 dataset has 9 major families and 49 features, setting up as a new benchmark in the dataset community.

Bagui et al. [13] generates the comparison of two different classifiers of Machine learning using the hybrid feature selection process. The UNSW-NB15 dataset undergo through the feature selection process that are k-means clustering and a Correlation Based Subset evaluation (CFS). After the process of feature techniques, the data undergoes through the two classifiers. The probabilistic Naïve Bayes (NB) classifier gives a higher classification accuracy rate i.e. 99% and lowers FAR i.e. 1% for most attack families using the selection of feature process. . However, when it comes to J48 decision tree, the classification rate was not impacted by the hybrid process, though it perform almost same with or without the selection factors.

Sarraf et al. [14] detects and classify the kinds of intrusions. To detect the intrusions, binary classification was performed in the network traffic. After the classification, machine learning algorithms are performed to determine the accuracy and other performance metrics. Among them, XGBoost performed quite well then the rest with 83.92% accuracy.

Singh et al. [15] compares between SVM AND iSVM. SVM algorithm is dependent on kernel and each Kernel function has capability to map the non-linearly separable data points into different dimensions. By modifying the traditional Gaussian Kernel, the paper proposed the idea of modified Gaussian Kernel making the SVM into Improved SVM (iSVM). The traditional SVM shows less execution in the detection of cyber-attack compare to iSVM in KDDCUP2009 dataset. The comparison of iSVM is not limited; it is even compared with Bayesian Network (BN) and Classification and Regression Tree (CART) but the performance doesn't show satisfying result for R2L and U2r respectively. But still the iSVM accuracy has remarkably improved for the attacks class like normal and DOS than the traditional SVM algorithm.

Sasan et al. [16] analyze to build a hybrid model with the feature selection method to observe the behavior of the attacks of network packets in the standard version of NSL-KDD data sets. The hybrid model is made up of J48 and CART algorithm. As the proposed model is used for achieving high accuracy in the intrusion detection system with 29 features is about 88.23%. The author further talks about working on more parameters to improve the intrusion detection.

3. PROPOSED METHODOLOGY

In the Figure 1, the raw data of network packet is collected and undergo through the processes of data-preprocessing and normalization as the feature network packets have different ranges, which is converted into numeric values that can be useful and understandable for the machine learning model. After the data is refined, now it is ready to process in the DT algorithm. The hybrid approach in which different types of attributes including Gini Index and Entropy of DT model has been implemented separately to identify the most accurate procedure to detect intrusion with respect to the execution time. The different DT methodologies including DT using Gini Index, DT using train-split method and DT using information entropy along with their respective subdivision such as using K-Fold validation, using Stratified K-Fold validation are implemented

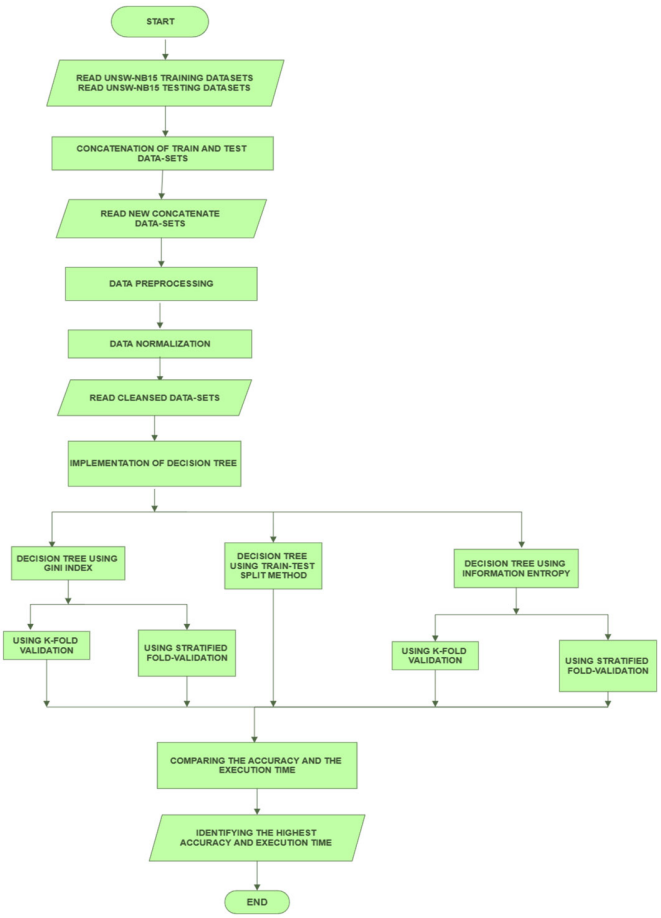

Figure 1. Experimental flowchart of the setup

3.1. Spliting of the dataset

The UNSW-NB15 dataset is encounter through the process of splitting into train and test dataset. The splitting is categorize using different types of cross-validation methods and the simple train and test split methods. Each type of cross validator is go through the process of decision tree-selection attribute methods. The statistical procedure that estimate the potential of the machine learning model in a dataset. Here, K-fold and Stratified K-fold cross validation is used. K-fold cross validator is used to divide the dataset into k consecutive folds but in stratified, each fold is ensure that it has a given the same proportion in both the training and testing of the dataset. Both the cross validation is compare as it is run through the

Here the value of k= 10 folds.

3.2. Decision Tree

Decision tree, a supervised learning method has its internal nodes representing the test of an attribute, the path which is the outcome and lastly the leaf nodes or terminal node showcasing the class label [19]. It is a recursive flowchart-like tree, maps the object and its attribute [2]. The simplicity of this technique is easy for classification rules for solving possible solution in a certain constraint [18]. Some of the famous decision tree models are ID3, C4.5 and CART. The tree uses different selection attribute for better performance than the other learning algorithms. The tree uses its attributes namely, Entropy and Gini index to measure the impurity criteria.

3.2.1. Using Entropy

In Decision Tree, the attribute entropy is select to measure the disorder or amount missing information in a given datasets. The use of entropy is in the ID3 algorithm to split the tree. Let's, consider a dataset with N classes, the entropy is calculated:

$$E = \sum_{i=1}^{N} p_i * \log_2(p_i)$$

Where, p is the discrete set probabilities of random selection of an element i in a data. The Entropy is denoted by E.

As the calculation of uncertainty is done, the goal is now to reduce the missing class in the target features of the dataset for training of the decision tree. The gain of

information is the calculation of comparison of entropy is between the root nodes to leaf nodes. Lesser the entropy, higher the information gain. The information gain is calculated: Information Gain =Entropy (root nodes) −Entropy (leaf nodes)

3.2.2. Using of Gini Index

Table 1. Comparison between Accuracy Gain

Algorithm	DT-SKFOLD	DT-KFOLD	DT
Execution Time(Gini)	17.341461	17.9505	23.42 6391
Execution Time(Entropy)	19.460344	20.509309	24.62 696

Decision tree can be used as both classification and regression techniques. Gini index is the measurement of the probability that a variable will not be classified correctly if it is randomly chosen. The degree varies from 0 to 1. The approach is used by CART (Classification and regression tree) algorithm. The mathematical formula is:

$$Gini(P) = 1 - \sum_{i=1}^{n} (p_i)^2$$

$$P = (p_1, p_2, \ldots . p_n) and\ p_i$$

Where P is the probability of an object that is being classified to a particular class.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, the result is presented of the experiment that has been conducted to evaluate the best combination of attribute and cross validation method that execute the best accuracy in small span of time. The result is shown in two tables along with the graphical representations, the Table 1 and Table 2 shows the comparison of accuracy gain and the least execution time executed between the Gini Index and Entropy on the following:

- Decision Tree-using K-fold Validation(DT-KFOLD)
- Decision Tree-using SK-fold Validation(DT-SKFOLD)
- Decision Tree-using Train and Test split method(DT)

Table 2: Comparison between Execution Time

Algorithm		DT-SKFOLD	DT-KFOLD	DT
Accuracy(Gini)		89.14	89.14	86.75
Accuracy(Entropy)		87.27	87.27	87.18

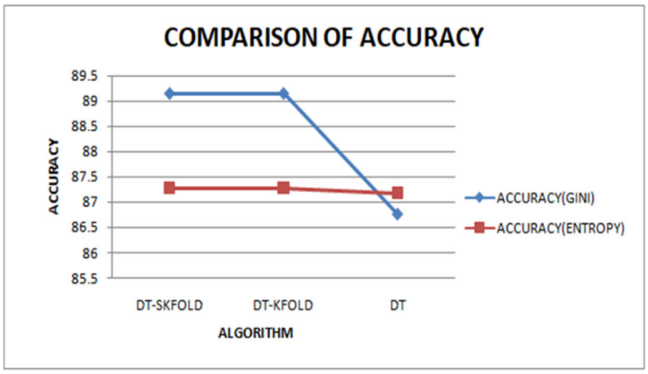

Figure 2: (**a**) Graphical representation of the accuracy compared between the algorithms

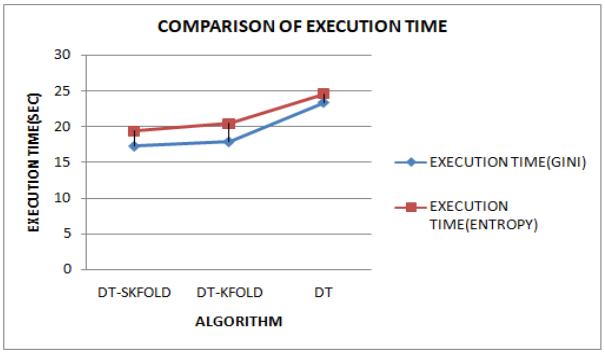

Figure 2: (**b**) Graphical representation of the Execution Time compared between the algorithms.

4. Discussion

From the Table 1 and Table 2, Decision tree using Stratified K-fold validation with the Gini Index impurity measurement attribute gives the best accuracy of 89.14% that is executed in 17.341 sec. It shows that, the assigning of Gini Index in the Decision with the improved K-fold Cross Validation method i.e. Stratified K-fold Cross Validation method works better in validating the splitting method in the tree algorithm to evaluate the highest accuracy in short span of time.

5. Conclusions

Today with the hit of cyber security in the information based world, the more attacks can be seen in the recent days. It makes the network more vulnerable. To prevent or detect, there is need for the computer to learn in short time. So any learning models need to modify its original techniques to classify the attacks at very less amount of time. The proposed hybrid method used to detect the network attack in the mentioned dataset, using different types of Decision Tree technique with Cross Validation has evaluate a performance of accuracy and execution time. Decision Tree using Stratified K-Fold Cross Validation execute the best result do far giving the success rate 89.14% in 17.341 sec. The Future work is to provide to prevent the network attacks in a real time data with the help of modified learning algorithm.

References

1. Sarker IH, Kayes AS, Badsha S, Alqahtani H, Watters P, Ng A. Cybersecurity data science: an overview from machine learning perspective. *Journal of Big data*. **2020 Dec;***7(1):*1-29.
2. Xin Y, Kong L, Liu Z, Chen Y, Li Y, Zhu H, Gao M, Hou H, Wang C. Machine learning and deep learning methods for cybersecurity. *Ieee access*. **2018 May 15**;*6:*35365-81.
3. Shaukat K, Luo S, Varadharajan V, Hameed IA, Xu M. A survey on machine learning techniques for cyber security in the last decade. *IEEE Access.* **2020 Dec 2;***8*:222310-54.
4. Amit I, Matherly J, Hewlett W, Xu Z, Meshi Y, Weinberger Y. Machine learning in cyber-security-problems, challenges and data sets. *arXiv preprint arXiv:1812.07858*. **2018 Dec 19**.
5. Torres JM, Comesaña CI, Garcia-Nieto PJ. Machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*. **2019 Oct;10(10):**2823-36.
6. Ford V, Siraj A. Applications of machine learning in cyber security. InProceedings of the 27th international conference on computer applications in industry and engineering 2014 Oct 13 (Vol. 118). Kota Kinabalu: IEEE Xplore.
7. Buczak AL, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials*. **2015 Oct 26;**18(2):1153-76.
8. Sinclair C, Pierce L, Matzner S. An application of machine learning to network intrusion detection. InProceedings 15th Annual Computer Security Applications Conference (ACSAC'99) 1999 Dec 6 (pp. 371-377). IEEE.
9. Abdulraheem MH, Ibraheem NB. A detailed analysis of new intrusion detection dataset. *Journal of Theoretical and Applied Information Technology*. **2019 Sep 15**;97(17):4519-37.
10. Alqahtani H, Sarker IH, Kalim A, Hossain SM, Ikhlaq S, Hossain S. Cyber intrusion detection using machine learning classification techniques. InInternational Conference on Computing Science, Communication and Security 2020 Mar 26 (pp. 121-131). Springer, Singapore
11. Sabar NR, Yi X, Song A. A bi-objective hyper-heuristic support vector machines for big data cyber-security. *Ieee Access*. **2018 Mar 6**;6:10421-31.
12. Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In2015 military communications and information systems conference (MilCIS) 2015 Nov 10 (pp. 1-6). IEEE.
13. Bagui S, Kalaimannan E, Bagui S, Nandi D, Pinto A. Using machine learning techniques to identify rare cyber-attacks on the UNSW-NB15 dataset. *Security and Privacy*. **2019 Nov;**2(6):e91.
14. Sarraf J, Chakraborty S, Pattnaik PK. Detection of Network Intrusion and Classification of Cyberattack Using Machine Learning Algorithms: A Multistage Classifier Approach. InInternational conference on smart computing and cyber security: strategic foresight, security challenges and innovation 2020 Apr 23 (pp. 285-295). Springer, Singapore.
15. Singh S, Agrawal S, Rizvi MA, Thakur RS. Improved Support Vector Machine for Cyber Attack Detection. InProceedings of the World Congress on Engineering and Computer Science 2011 Oct (Vol. 1).
16. Sasan HP, Sharma M. Intrusion detection using feature selection and machine learning algorithm with misuse detection. *International Journal of Computer Science and Information Technology*. **2016 Feb**;8(1):17-25.
17. Revathi S, Malathi A. A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection. *International Journal of Engineering Research & Technology (IJERT).***2013 Dec;**2(12):1848-53.
18. Das K, Behera RN. A survey on machine learning: concept, algorithms and applications. *International Journal of Innovative Research in Computer and Communication Engineering*. **2017 Feb**;5(2):1301-9.
19. Sharma H, Kumar S. A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR*). **2016 Apr 5;**5(4):2094-7.
20. Sonule AR, Kalla M, Jain A, Chouhan DS. UNSWNB15 Dataset and Machine Learning Based Intrusion Detection Systems. *International Journal of Engineering and Advanced Technology*. **2020;**9:2638-48.
21. Khan A. Data Science in Action: Key to Cyber security.