

# An Efficient Machine Learning Technique for Network Intrusion Detection System for Cyber Security Application

Bibhu Baibhav<sup>1</sup>, Prof. Sarwesh Site<sup>2</sup>

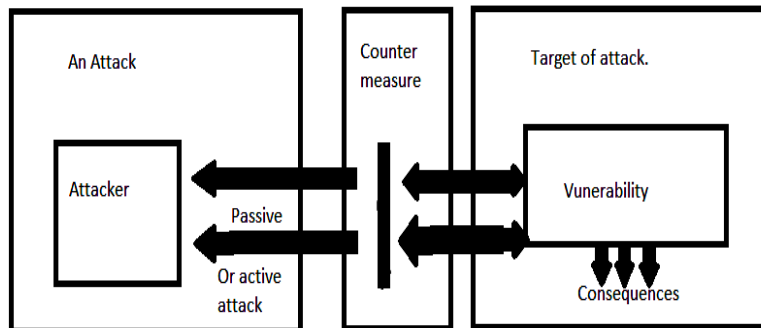
*M.Tech Scholar1, Assistant Professor2, Department of Computer Science and Engineering, All Saints' College of Technology, Bhopal, India*

**Abstract--** Intrusion detection is one of the important security problems in today's cyber world. Countless procedures have been created which depend on AI draws near. So for recognizing the interruption we have planned the AI calculations. By utilizing the calculation we figure out interruption and we can recognize the assailant's subtleties too. IDS are chiefly two sorts: Host based and Organization based. A Network based Intrusion Detection System (NIDS) is generally positioned at network focuses, for example, a door and switches to check for interruptions in the organization traffic. This paper presents an efficient machine learning technique for network intrusion detection system for cyber security application.

**Keywords--** NIDS, C4.5, Decision Tree, accuracy, IoT, IDS, Cyber, Attack, Security.

## I. INTRODUCTION

A cyber attack can be utilized by sovereign states, people, gatherings, society, or associations, and it might start from an unknown source. An item that works with a digital assault is some of the time called a digital weapon. A digital assault might take, modify, or obliterate a predefined focus by hacking into a powerless framework. Digital assaults can go from introducing spyware on a PC to endeavoring to obliterate the foundation of whole countries. Lawful specialists are looking to restrict the utilization of the term to episodes causing actual harm, recognizing it from the more normal information breaks and more extensive hacking exercises.



**Figure 1: Attack system.**

Things have advanced because of the intermingling of different innovations, continuous investigation, AI, omnipresent registering, item sensors, and inserted frameworks. Customary fields of implanted frameworks, remote sensor organizations, control frameworks, mechanization (counting home and building robotization), and others all add to empowering the Web of things.

In the customer market, IoT innovation is generally inseparable from items relating to the idea of the "shrewd home", including gadgets and apparatuses (like lighting apparatuses, indoor regulators, home security frameworks and cameras, and other home machines) that help one or more normal biological systems, and can be controlled by means of gadgets related with that environment, like PDAs and savvy speakers. The IoT can likewise be utilized in medical services frameworks.

II. METHODOLOGY

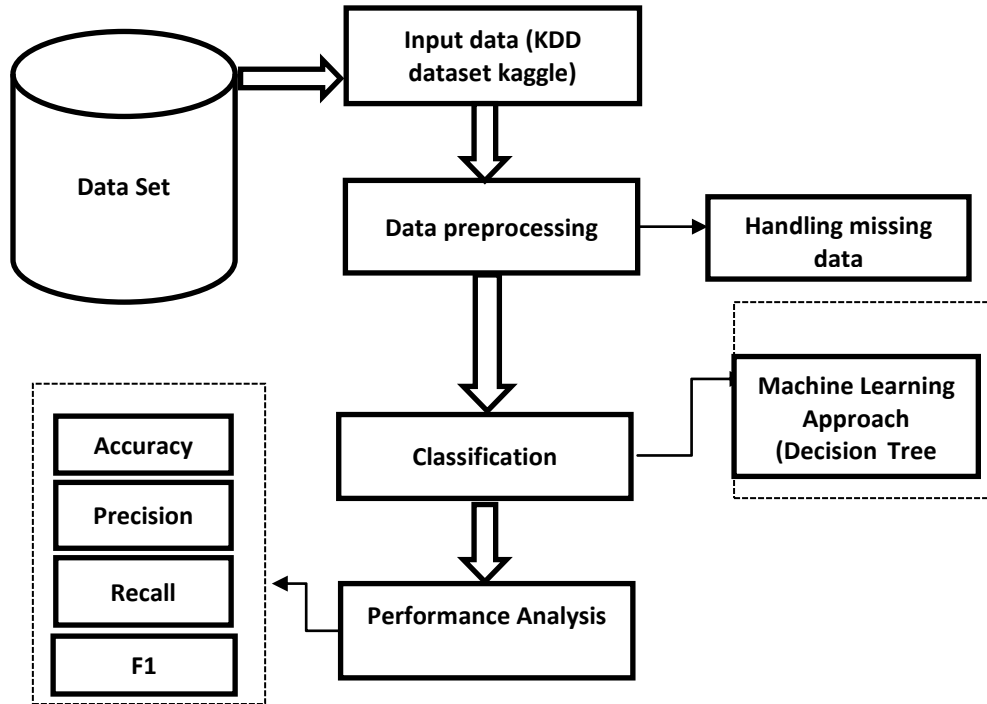


Figure 2: Flow Chart

*Steps-*

- Firstly, download the dataset from KDD dataset kaggle website, which is a large dataset provider company for research.
- Now preprocessing of the data, here handling the missing dataset. Remove the null value or replace from common 1 or 0 value.
- Now apply the classification method based on the machine learning approach. The Decision Tree (DT) with C4.5 machine learning method is applied.
- Now check and calculate the performance parameters in terms of the precision, recall, F<sub>1</sub> measure, accuracy and error rate.

It is fundamental hub that is adjusted as to such an extent that the entropy diminishes with parting downwards. This essentially implies that the more parting is done properly; coming to a distinct choice becomes simpler.

In this way, we really look at each hub against each parting probability. Data Gain Proportion is the proportion of perceptions to the absolute number of perceptions ( $m/N = p$ ) and ( $n/N = q$ ) where  $m+n=N$  and  $p+q=1$ .

Subsequent to parting in the event that the entropy of the following hub is lesser than the entropy prior to parting and in the event that this worth is the least when contrasted with all conceivable experiments for parting, then the hub is parted into its most perfect constituents.

1. Check for the above base cases.
2. For each attribute a, find the normalised information gain ratio from splitting on a.
3. Let a<sub>best</sub> be the attribute with the highest normalized information gain.
4. Create a decision node that splits on a<sub>best</sub>.
5. Recur on the sublists obtained by splitting on a<sub>best</sub>, and add those nodes as children of node.

Advantages of C4.5 over other Decision Tree systems:

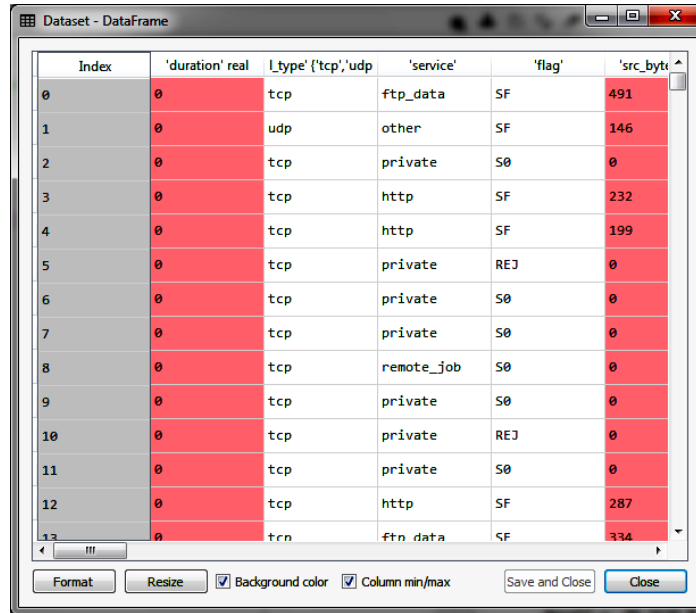
1. The algorithm inherently employs Single Pass Pruning Process to Mitigate overfitting.
2. It can work with both Discrete and Continuous Data
3. C4.5 can handle the issue of incomplete data very well

We should also keep in mind that C4.5 is not the best algorithm out there but it does certainly prove to be useful in certain cases.

**III. SIMULATION RESULTS**

The implementation of the proposed algorithm is done over python spyder 3.7.

The sklearn, numpy, pandas, matplotlib, pyplot, seaborn, os library helps us to use the functions available in spyder environment for various methods like decision tree, random forest, naive bayes etc.

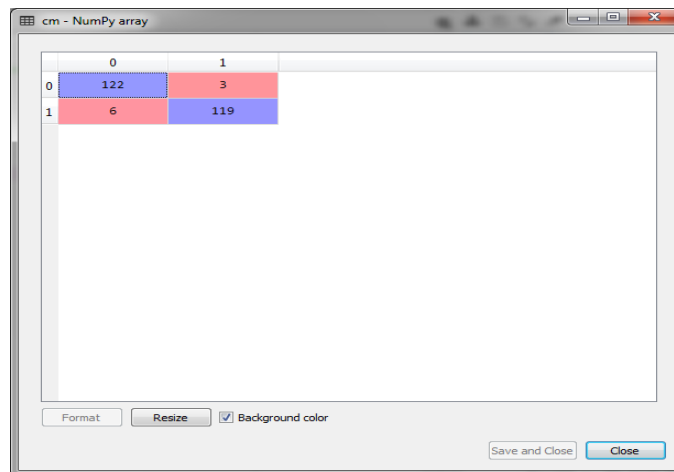


Index	'duration' real	'_type' {'tcp','udp'}	'service'	'flag'	'src_bytes'
0	0	tcp	ftp_data	SF	491
1	0	udp	other	SF	146
2	0	tcp	private	S0	0
3	0	tcp	http	SF	232
4	0	tcp	http	SF	199
5	0	tcp	private	REJ	0
6	0	tcp	private	S0	0
7	0	tcp	private	S0	0
8	0	tcp	remote_job	S0	0
9	0	tcp	private	S0	0
10	0	tcp	private	REJ	0
11	0	tcp	private	S0	0
12	0	tcp	http	SF	287
13	0	tcp	ftp_data	SF	334

**Figure 3: Dataset**

Figure 3 is showing the KDD data set. This dataset contain the total 999 datas with 42 colour features like 'duration' real 'protocol\_type' {'tcp','udp', 'icmp'} 'service'

'flag' 'src\_bytes' real 'dst\_bytes' real 'land' {'0', '1'} 'wrong\_fragment' real 'urgent' real 'hot' etc.



	0	1
0	122	3
1	6	119

**Figure 4: Confusion Matrix**

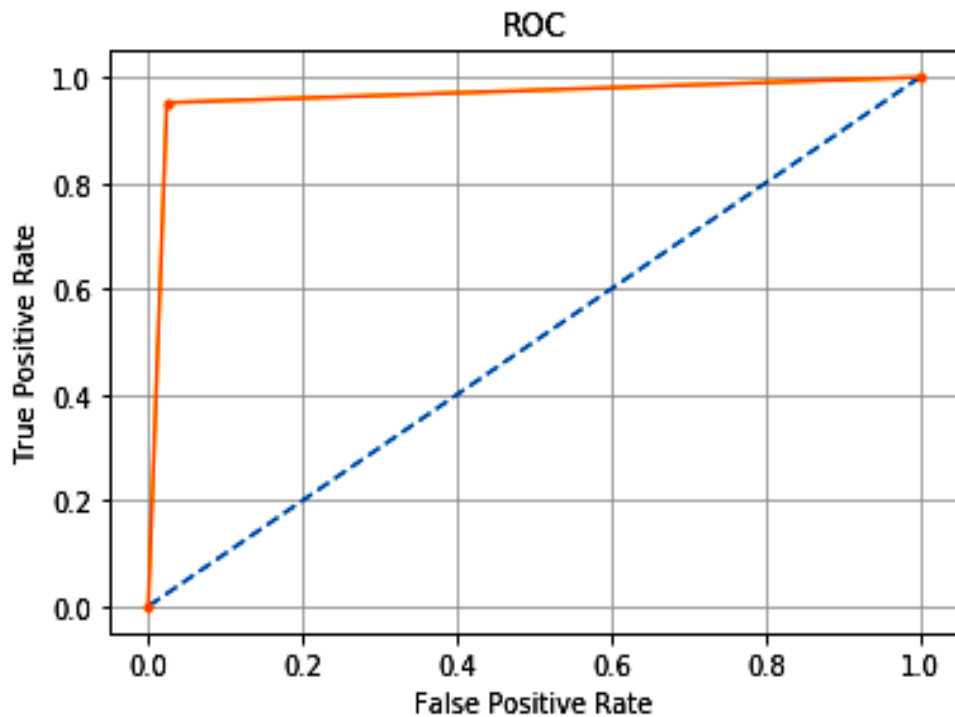
Figure 4 is showing the confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

*TP*: True Positive: Predicted values correctly predicted as actual positive

*FP*: Predicted values incorrectly predicted an actual positive. i.e., Negative values predicted as positive

*FN*: False Negative: Positive values predicted as negative

*TN*: True Negative: Predicted values correctly predicted as an actual negative



**Figure 5: ROC**

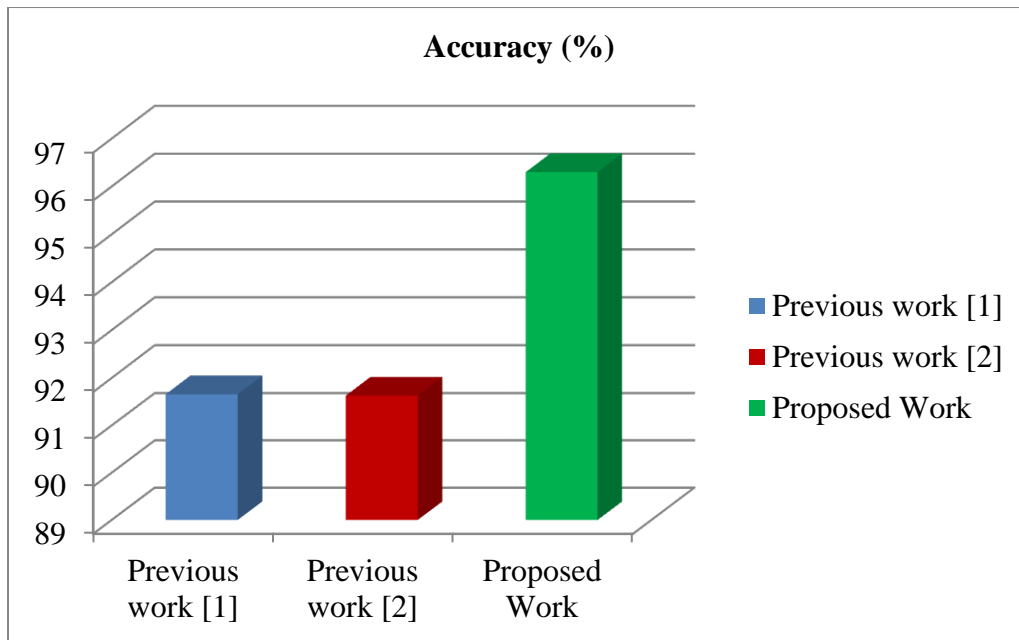
The figure 5 is showing the Receiver Operating Characteristic Curves (ROC) is a plot of signal (True Positive Rate) against noise (False Positive Rate).

**Table 1:  
Simulation Result of DT with C4.5**

Sr. No.	Parameters	Proposed Method (%)
1	Precision	97.6
2	Recall	95.3
3	F-measure	96.4
4	Accuracy	96.3
5	Error Rate	3.5

**Table 2:**  
**Result Comparison**

Sr. No.	Parameters	Previous work [1]	Previous work [2]	Proposed Work
1	Methodology	CNN	LSTM	DT with C4.5
2	Precision (%)	NA	63.76	97.6
3	Recall (%)	NA	66.36	95.3
5	F-measure (%)	NA	65.04	96.4
6	Accuracy (%)	91.66	91.63	96.3
7	Error Rate(%)	8.34	837	3.5



**Figure 6: Accuracy comparison**

Table 2 is showing the results comparison of the previous and proposed research works.

It is clear from the previous and proposed work performance parameters result calculation, the proposed work is achieving significant better results than existing.

#### IV. CONCLUSION

This paper presents an efficient machine learning technique for network intrusion detection system for cyber security application. The dataset is taken from the KDD dataset kaggle. Precision value of existing results is 63.76 and 85.82% while proposed work achieved 97.6%. The recall value achieved by proposed technique is 95.3% while previous achieved is 66.36 and 84.49%. The f measure value is 96.4% and error rate is 3.5% by proposed technique while previous results is 65.04 and 85.14% of Fmeasure and 8 and 16% is error rate. Finally the accuracy achievement is 96.3% by the proposed methodology while previous accuracy value is 91.63 and 83.58%.

#### REFERENCES

- [1] S. Ho, S. A. Jufout, K. Dajani and M. Mozumdar, "A Novel Intrusion Detection Model for Detecting Known and Innovative Cyberattacks Using Convolutional Neural Network," in IEEE Open Journal of the Computer Society, vol. 2, pp. 14-25, 2021, doi: 10.1109/OJCS.2021.3050917.
- [2] V. K. Navya, J. Adithi, D. Rudrawal, H. Tailor and N. James, "Intrusion Detection System using Deep Neural Networks (DNN)," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), 2021, pp. 1-6, doi: 10.1109/ICAECA52838.2021.9675513.
- [3] S. Liu, M. Dibaei, Y. Tai, C. Chen, J. Zhang and Y. Xiang, "Cyber Vulnerability Intelligence for Internet of Things Binary," in IEEE Transactions on Industrial Informatics, vol. 16, no. 3, pp. 2154-2163, March 2020, doi: 10.1109/TII.2019.2942800.
- [4] Y. Jin, M. Tomoishi and N. Yamai, "Anomaly Detection by Monitoring Unintended DNS Traffic on Wireless Network," 2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 2019, pp. 1-6, doi: 10.1109/PACRIM47961.2019.8985052.
- [5] B. Peng, Q. Wang, X. Li, J. Cai, J. Fei and W. Chen, "Research on Abnormal Detection Technology of Real-Time Interaction Process in New Energy Network," 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2019, pp. 433-440, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00092.
- [6] W. Bi, K. Zhang, Y. Li, K. Yuan and Y. Wang, "Detection Scheme Against Cyber-Physical Attacks on Load Frequency Control Based on Dynamic Characteristics Analysis," in IEEE Systems Journal, vol. 13, no. 3, pp. 2859-2868, Sept. 2019, doi: 10.1109/JSYST.2019.2911869.
- [7] K. Liu, Z. Fan, M. Liu and S. Zhang, "Hybrid Intrusion Detection Method Based on K-Means and CNN for Smart Home," 2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), 2018, pp. 312-317, doi: 10.1109/CYBER.2018.8688271.
- [8] Y. Jin, K. Kakoi, N. Yamai, N. Kitagawa and M. Tomoishi, "A Client Based Anomaly Traffic Detection and Blocking Mechanism by Monitoring DNS Name Resolution with User Alerting Feature," 2018 International Conference on Cyberworlds (CW), 2018, pp. 351-356, doi: 10.1109/CW.2018.00070.
- [9] R. Velea and Ş. Drăgan, "CPU/GPU Hybrid Detection for Malware Signatures," 2017 International Conference on Computer and Applications (ICCA), 2017, pp. 85-89, doi: 10.1109/COMAPP.2017.8079736.
- [10] S. Merat and W. Almuhtadi, "Artificial intelligence application for improving cyber-security acquirement," 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), 2015, pp. 1445-1450, doi: 10.1109/CCECE.2015.7129493.
- [11] S. Han, M. Xie, H. Chen and Y. Ling, "Intrusion Detection in Cyber-Physical Systems: Techniques and Challenges," in IEEE Systems Journal, vol. 8, no. 4, pp. 1052-1062, Dec. 2014, doi: 10.1109/JSYST.2013.2257594.
- [12] M. Bousaaid, T. Ayaou, K. Afdel and P. Estrailier, "Hand gesture detection and recognition in cyber presence interactive system for E-learning," 2014 International Conference on Multimedia Computing and Systems (ICMCS), 2014, pp. 444-447, doi: 10.1109/ICMCS.2014.6911197.