EFFECTIVE IDS CLASSIFICATION FOR IOT

1st Aryan
SCORE, VIT, Vellore, India
aryan.2021b@vitstudent.ac.in

2nd Pratyush Abhi
SCORE, VIT, Vellore, India
pratyush.abhi2021@vitstudent.ac.in

3rd Vuggi Siddhartha
SCORE, VIT, Vellore, India
vuggi.siddhartha2021@vitstudent.ac.in

4th Chandrasegar T
SCORE, VIT, Vellore, India
chandrasegar.t@vit.ac.in

Abstract – While the concept of the Internet of Things affords ubiquity and connection as its strength, it depends heavily on the reliability and the security of both the devices and the network it employs. Intrusion Detection Systems (IDS) play the role of a first barrier to protect IoT systems from the evil deeds. However, due to the dynamism and heterogeneity of the data that is captured across the IoT network, traditional methods of intrusion detection is not possible and therefore different advanced machine learning techniques is required. This research aims at assessing the effectiveness of the leading machine learning algorithms such as Random Forest, KNN, SVC, XG Boost, Decision Tree, and Linear Regression on the classification of IDS data created for the IoT purpose. The method of this research is a path towards to understanding the organic applicability of these models in differentiating intrusions in a blend of data from IoT devices.

They are: KNN, SVC, XG Boost, Random Forest, Linear Regression and Decision tree.

I. INTRODUCTION

The relationship between humans and new technologies has become different nearly abruptly due to the IoT becoming mainstream. Smart device environments interface with each other creating a virtual amalgam of the physical and the online. At the same time, connectivity introduces an unprecedented amount of risk because threat actors seeking to exploit vulnerabilities seek out IoT systems as their primary targets. In this paradigm, Intrusion Detection Systems (IDS) assumes roles akin to that of a sentry, a post that is absolutely crucial in ensuring that an organization monitors the traffic on its networks and those that may point to a possible intrusion. Networks have revealed that conventional classic IDS operate efficiently in ordinary environments, but the peculiarities of IoT conditions offer new challenges. More sophisticated and flexible solutions must be employed because the number of new objects, their diversity, and especially the fact that the data they produce is real-time, makes it impossible to address the problem with simple solutions.

IOT Dataset

In this case we have moved the IOT dataset to classify the problem. This dataset was comprised of 123118 samples. In our research study, the attributes present in the dataset are as follows: Important attributes have been explained below

- Backward packets: Max. payload, Standard payload, Average payload, Minimum payload, Initial window size bulk rate.
- Carry packets – Standard payload, Sub flow, Initial window size.
- Flow packets; maximum payload, standard payload, flag count.

A. Random Forest

It generates several models known as decision trees using the process of bagging and it combines it to get the best outcome. Nonetheless, within each tree, the variable selection is divided through the split of each tree. From this construction, the best tree is selected by having the least standard deviation. RF can be ranked relatively high when it comes to classification problems.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \tag{1}$$

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}} \tag{2}$$

B. Logistic Regression

A classification algorithm is used for the binary classification or for the multiclass classification. It is mostly employed when the output is binary as is the case with sentiment analysis, image and video classification, and a lot more.

$$P(x) = \frac{1}{1+e^{-(\beta_0+\beta_1)}} \tag{3}$$

C. Naïve Bayes

A classification algorithm that calculated the probability of each class and make decision by comparing probability of each class based on the equation of Bayes' theorem,with the underlying assumption that all features are independent of each other given the class itself. Nonetheless, due to its simplistic assumption that the features are independent, This algorithm is very effective for all forms of classification, including text classification and spam filtering.

$$P\left({}^{y}/_{X}\right) = \frac{P\left({}^{X}/_{y}\right)P(y)}{P(X)} \tag{4}$$

D. KNN

K Nearest Neighbors (KNN) is a basic machine learning algorithm which falls under the classification of non parametric methods. Of the nearest 'k' neighbors to a query point, the label or value is determined by the most frequent class or the average of the neighbors.

E. SVC

Support Vector Classifier or SVC is a Machine Learning algorithm used for classification purpose. This is by virtue of the fact that it identifies the separating hyperplane with the maximum margin between different classes of the data in a high dimensional space.

F. XG Boost

It is an ML algorithm with roots in decision trees and that is especially known to be fast and efficient for classification and regression. It improves model accuracy by incorporating ways such as; feature selection, decision tree pruning, use of parallel computing.

G. Linear Regression

For the given input features x_i, y_i with an input vector x_i of data D the linear form of solution $f(x) = mx + b$ is solved by subsequent parameters:

$$m = \frac{\left(\sum_i x_i y_i\right) - n\overline{x_i y_i}}{\left(\sum_i x_i^2\right) - n\overline{x_i}^2} \qquad (5)$$

$b = \overline{y} - m\overline{x}$ where $\overline{x}, \overline{y}$ are mean.

H. Decision Tree

It symbolizes every possible decision for the outcome. First, to decide on the features, the entropy or information gain is used for it, and then depending on the acquired result, the appropriate input features are chosen. Based on performing features, the datasets are divided with some restrictions.

$$E(s) = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (6)$$

II. Results and Discussions

The results are experimented with using Scikit Python using Win. 11 OS, 8 GB RAM, and an i5 12th generation processor. We test the ML models using full validation. We have taken the IDS data from CIC.

A. HeatMap

Heatmap is a graphical technique that implements data values using color intensities. In data visualization, it is rather used to indicate the intensity or spread of values in two dimensions to form some pattern.

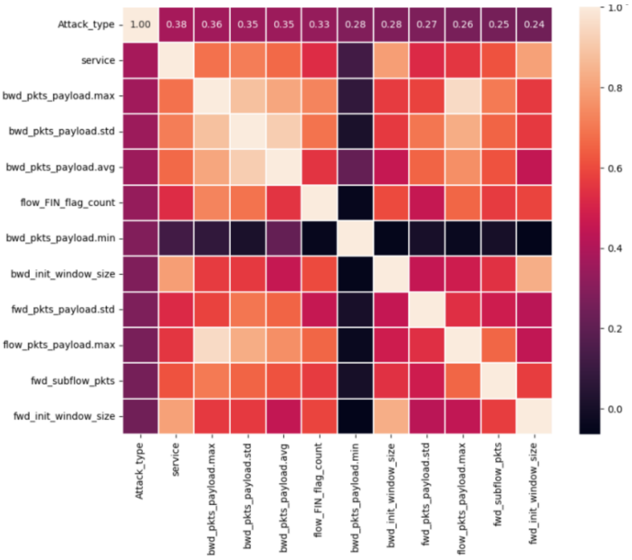
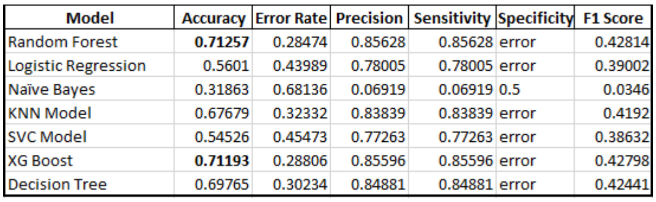


Figure 1: HeatMap



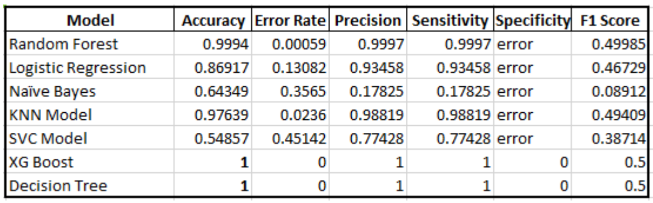
Model	Accuracy	Error Rate	Precision	Sensitivity	Specificity	F1 Score
Random Forest	**0.71257**	0.28474	0.85628	0.85628	error	0.42814
Logistic Regression	0.5601	0.43989	0.78005	0.78005	error	0.39002
Naïve Bayes	0.31863	0.68136	0.06919	0.06919	0.5	0.0346
KNN Model	0.67679	0.32332	0.83839	0.83839	error	0.4192
SVC Model	0.54526	0.45473	0.77263	0.77263	error	0.38632
XG Boost	**0.71193**	0.28806	0.85596	0.85596	error	0.42798
Decision Tree	0.69765	0.30234	0.84881	0.84881	error	0.42441

Table 1 : Raw Data

According to the definitions given under raw data, it is also called as primary or source data which states all data that are collected in a natural form from different sources without any further processing.es. This data is in raw format and no effort has been made to clean, sort, analyze, or enrich this data in any way or form. They are often inaccurate, lack some records and contain other unwanted records.

B. Models vs. Performance

Models are the mathematical or computational structures involved in data analysis to model, predict or classify events in the real world based on input data whereas, performance is a measure of how well the models carrying out the aforesaid objectives. One of the important criteria in model building is the trade between the models' sophistication and their accuracy.



Model	Accuracy	Error Rate	Precision	Sensitivity	Specificity	F1 Score
Random Forest	0.9994	0.00059	0.9997	0.9997	error	0.49985
Logistic Regression	0.86917	0.13082	0.93458	0.93458	error	0.46729
Naïve Bayes	0.64349	0.3565	0.17825	0.17825	error	0.08912
KNN Model	0.97639	0.0236	0.98819	0.98819	error	0.49409
SVC Model	0.54857	0.45142	0.77428	0.77428	error	0.38714
XG Boost	**1**	0	1	1	0	0.5
Decision Tree	**1**	0	1	1	0	0.5

Table 2: Refined Data

Processed or cleaned data refers to the kind of data derived from raw data and arranged systematically in a comprehensible format. This process involves data scrubbing, where the data is first cleaned then filtered and put to the right format to eliminate error, duplicate and inconsistent data entries. Derived data is for analysis, reporting or for further use in models or another model.

III. CONCLUSION

We found the results on IDS data by applying the RF, LR, NB, KNN, SVC, XG boost, and DT models. We fully validate the models using both raw and refined IDS samples. In raw samples, XG boost and RF attain the highest accuracy of 71.25 and 71.19%, respectively. Whereas in the case of refined samples, XG boost and DT outperform other models.

REFERENCES

[1] Performance Evaluation of Parametric and Non-Parametric Machine Learning Models using Statistical Analysis for RT-IoT2022 Dataset. Sharmila B S1 *, Nandini B M2 , Kavitha S S1 & Anand Srivatsa1 1Department of Electronics and Communication Engineering, 2Department of Information Science and Engineering, The National Institute of Engineering, Mysuru 570 008, Karnataka, India Received 19 December 2023; revised 18 April 2024; accepted 14 June 2024

[2] Optimized common features selection and deep-autoencoder (OCFSDA) for lightweight intrusion detection in Internet of things. Uneneibotejit Otokwala1 · Andrei Petrovski1 · Harsha Kalutarage1

[3] Enhancing internet of things security: evaluating machine learning classifiers for attack prediction. Areen Arabiat, Muneera Altayeb Department of Communications and Computer Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan

[4] Deep-IDS: A Real-Time Intrusion Detector for IoT Nodes Using Deep Learning. SANDEEPKUMAR RACHERLA 1, (Senior Member, IEEE), PRATHYUSHA SRIPATHI 1, (Member, IEEE), NURUZZAMAN FARUQUI 2, MD ALAMGIR KABIR 3, (Member, IEEE), MD WHAIDUZZAMAN 4, (Senior Member, IEEE), AND SYED AZIZ SHAH 5

[5] A Comprehensive Analysis of the Machine Learning Algorithms in IoT IDS Systems. Ozdogan E.

[6] ABCNN-IDS: Attention-Based Convolutional Neural Network for Intrusion Detection in IoT Networks. Momand A.; Jan S.U.; Ramzan N.

[7] CICIoV2024: Advancing realistic IDS approaches against DoS and spoofing attack in IoV CAN bus. Momand, Asadullah (57863895300); Jan, Sana Ullah (56740235200); Ramzan, Naeem (16069861100)

[8] CICIoV2024: Advancing realistic IDS approaches against DoS and spoofing attack in IoV CAN bus Euclides Carlos Pinto Neto a, Hamideh Taslimasa a, Sajjad Dadkhah a, Shahrear Iqbal b, Pulei Xiong b, Taufiq Rahman b, Ali A. Ghorbani a

[9] M-RL: A mobility and impersonation-aware IDS for DDoS UDP flooding attacks in IoT-Fog networks Saeed Javanmardi a, Meysam Ghahramani b, Mohammad Shojafar c, Mamoun Alazab d, Antonio M. Caruso

[10] Effective Intrusion Detection in Highly Imbalanced IoT Networks with Lightweight S2CGAN-IDS Caihong Wang, Du Xu, Zonghang Li, Dusit Niyato, Fellow IEEE.

[11] Intrusion detection systems for IoT-based smart environments: a survey. Mohamed Faisal Elrawy, Ali Ismail Awad, and Hesham F. A. Hamed5 Elrawy et al. Journal of Cloud Computing: Advances, Systems and Applications (2018)

[12] Quantized autoencoder (QAE) intrusion detection system for anomaly detection in resource-constrained IoT devices using RT-IoT2022 dataset. S Sharmila and Rohini Nagapadma,Sharmila and Nagapadma Cybersecurity (2023)

[13] Enhancement of an IoT hybrid intrusion detection system based on fog-to-cloud computing. Doaa Mohamed and Osama Ismael Mohamed et al. Journal of Cloud Computing Published: 22 March 2023

[14] FEDDBN-IDS: federated deep belief network-based wireless network intrusion detection system. M Nivaashini 1, E. Suganya 2, S. Sountharrajan 3, M. Prabu3 and Durga Prasad Bavirisetti Nivaashini et al. EURASIP Journal on Information Security (2024) Published: 04 April 2024

[15] Intrusion Detection System for IoT Based on Deep Learning and Modified Reptile Search Algorithm. Abdelghani Dahou, ¹2 Mohamed Abd Elaziz,3,4,5 Samia Allaoua Chelloug, Mohammed A. Awadallah,4,7 Mohammed Azmi Al-Betar,4,8 Mohammed A. A. Al-qaness, and Agostino Forestiero 10 Hindawi

[16]) A survey on intrusion detection system and prerequisite demands in IoT networks. Parthiban Aravamudhan and T Kanimozhi 2021 J. Phys.A. Aleroud and L. Zhou, "Phishing environments, techniques , and countermeasures : A survey," Comput. Secur., vol. 68, pp. 160–196, 2017.

[17] IoT Intrusion Detection System Based on Machine Learning. Bayi XuLei , SunXiuqing ,MaoRuiyang, Dingand Chengwei Liu .Submission received: 18 September 2023 / Revised: 10 October 2023 / Accepted: 12 October 2023 / Published: 17 October 2023

[18] IoT Protocol-Enabled IDS based on Machine Learning. Rehab Alsulami , Batoul Alqarni, Rawan Alshomrani , Fatimah Mashat , Tahani Gazdar Received: 21 September 2023 | Revised: 23 October 2023 | Accepted: 4 November 2023

[19] A machine learning-based intrusion detection for detecting internet of things network attacks. Yakub Kayode Saheed a, Aremu Idris Abiodun b, Sanjay Misra c, Monica Kristiansen Holone c, Ricardo Colomo-Palacios. Received 21 October 2021; revised 5 February 2022; accepted 27 February 2022 Available online 28 March 2022

[20] A Comprehensive Analyses of Intrusion Detection System for IoT Environment. Sushil Kumar Singh , Jose Sicato , and Jong Hyuk Park J Inf Process Syst, Vol.16, No.4, pp.975~990, August 2020