IJBPAS, November, Special Issue, 2021, 10(11): 1039-1047
ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)
'A Bridge Between Laboratory and Reader'

www.ijbpas.com

DETECTION AND CLASSIFICATION OF INTRUDER DATA ATTACKS USING INTERNET OF THINGS AND MACHINE LEARNING TECHNIQUES BASED ON ENVIRONMENT SITUATION

SHIVAKUMARA T[1*], PURVANSH JAIN[2], RAJSHEKHAR M PATIL[3], DINESH MAVALURU[4], LOKESWARA REDDY.V[5] AND KANNADASAN .B[6]

1: Assistant Professor in MCA, BMS Institute of Technology and Management, Bangalore

2: B.Tech Graduate in Computer Science Engineering at Faculty of Engineering and Technology, JAIN (Deemed-to-be-University), Kanakpura Road, Bangalore, India

3: Professor, Department of CSE, Amruta Institute of Engineering and Management Sciences, Bangalore- 562109, Karnataka, India

4: Educator in Information Technology, College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia

5: Professor in Computer Science Engineering, K.S.R.M College of Engineering, YerramasuPalli, Tadigotla(village), Chintakommadinne (Mandal),YSR Kadapa (District), Andhra Pradesh

6: Assistant Professor in Civil Engineering at B.S Abdur Rahman Crescent Institute of Science and Technology, GST Road , Vandalur, Chennai, India

***Corresponding Author: Shivakumara T; E Mail: shivakumarat@bmsit.in**

Received 23rd July 2021; Revised 27th Aug. 2021; Accepted 30th Sept. 2021; Available online 1st Nov. 2021

https://doi.org/10.31032/IJBPAS/2021/10.11.1089

ABSTRACT

The number of Internet of Things devices, and the information created by these systems, has exploded in recent years. Because of its constrained resources, contributing systems in Internet of Things networks could be difficult, & safety on these systems is frequently disregarded. As an outcome, attackers now have a stronger motivation to attack Internet of Things. Even as number of assaults that can be launched against a system grows, conventional intrusion detection systems find it much harder to keep up. On the Bot-Internet of Things database, Machine learning methods are compared for consists of multi

categorization. They compared the Machine learning using numerous criteria such as reliability, accuracy, recall, F1 score, & log lost in an experimental. Overall reliability of radio frequency, inside the case of a Hypertext transfer protocol dispersed denial of service assault is 99 percent. Other simulation outcomes, such as accuracy, recall, F1 rate, & logarithmic loss metric, indicate it radio frequency, surpasses all sorts of assaults in classification model.

Keywords: Intruder Data; Internet of Things; Machine Learning; Environment

INTRODUCTION

The Internet of Things envisions a future in which things could comprehend context using detectors & interact with one another via networking capabilities **[1]**. According to the utilise cases **[2]** the gadgets in the Internet of things could be used to collect data. Such businesses include retail, health-care, & manufacturing, which employ Internet of Things gadgets to track bought things, monitor patients remotely, and run completely autonomous ware-houses. According to reports, the number of Internet of Things gadgets is increasing each year, with a projected total of 75.44 billion gadgets in 2025 **[3]**. As an outcome of the tremendous increase in Internet of Things systems, many attackers are targeting Internet of Things networks **[4]**. According to estimates, the majority of attack traffic produced on Internet of Things networks is automated using scripting & malware **[5]**. The rise in attacks, combined also with autonomous character of a attacks, is a challenge for Internet of Things systems, because most gadgets were operated in the blaze & absence manner of decades with-out human engagement **[6]**. Which, paired include the constraints for Internet of Things devices, such as low computing energy & bandwidth, makes it hard to provide effective security, which could lead to network layer assaults like denial of service **[7]**. As a result, it's critical to investigate methods for detecting this type of network traffic that could be employed in detection of intrusion & prevention systems.

An intrusion sensor model is a device form monitors the network of presumably dangerous traffic. Signature based identification & anomaly based detection are the two forms of intrusion detection system that could be used. A signature based intrusion detection system compares inbound traffic to a database with current attack patterns, meaning how an attack could only be recognized if the signature has been present in the system **[8]**. A network traffic monitor that is based on

anomalies & tries with determine some traffic which is out of the ordinary compared to the rest of the network **[9]**. A signature based detection method has a significant drawback in that it is vulnerable with the zero-day assault / a attacker who alters a attack with avoid being detected by signature database **[10]**. Anomaly based intrusion detection system are much better suited to using Machine Learning because they could be trained to distinguish among regular & attack traffic. Integrating Machine Learning with IDS, on the other hand, isn't a silver bullet and could cause issues. **[11]** Found various issues, one of which is that classifiers might give false positives, rendering the intrusion detection system useless because regular information causes the intrusion detection system to notify the system. Although the study is somewhat old, this is still a significant issue when employing machine learning with intrusion detection systems **[12-14]**. As an outcome, it's critical to identify algorithms that generate the fewest false positives.

Related works

Machine Learning was the portion of AI which entails feeding a database to a program, or in this case, a machine, that will be utilised to find patterns that could be utilised to forecast future information. Only a little amount of study has been done on Intrusion detection system utilising machine learning on Internet of Things networks. To this purpose, a recent study employed Machine Learning datasets from the Defense Advanced Research Projects to evaluate svm classifier algorithms **[15]**. This study's findings are reported in terms of RMSE, mean extreme percentage of error, receiver operating curve, & reliability, as well as the outcomes was positive, with Random forest algorithm being among the supermodels. Moreover, there are two major drawbacks to this study: For starters, it relied on DARPA databases that were more than 20years old at the time of publication. Second, this was not done utilising the datasets of multiple class evaluation. The Bot IoT dataset were also employed, including models such as k-closest neighbors, quadratic-discriminant evaluation, three dichotomise of iterative, radio frequency, appropriate gaining, multiple level perceptron, & narrowband **[16]**. In terms of effectiveness, accuracy, memory, F1 rate, as well as time, the study produced excellent outcomes. This research used a current dataset and a number of machine learning techniques **[17]**. However, none of the systems were subjected to multiple class assessment in this study.

The researchers of **[18]** employed multiple Machine Learning approaches for multiple class categorization. This study

examined techniques like logistic with a set of data developed by the researchers that was not open to the public. The research concluded that the optimum model for multiple class categorization were radio frequency, **[19]**. This study demonstrates that high-quality results can be achieved using multiple class categorization. Further algorithm development might assist support the research's findings. Within field of Internet of Things networks, there is presently a dearth of research in intrusion detection. It could be due to a lack of sets of data and real equipment, as all set of data are made up of simulated Internet of Things devices on ordinary computers. Also there is a scarcity of multiple class classification research, possibly owing to the amount of a specialized multiple class set of data. Because all available set of data were built with classification model in mind, multiple class evaluation necessitates merging them into a single dataset with adequate tagging for every class.

Methods

A fresh CSV file is constructed utilising the binary classification set of data to conduct multiple class evaluation. The set of data were gathered, randomised, and afterwards saved in a new folder. Because of the vast size of data, just a small portion of it is employed to avoid long run durations. The category representation of the training & testing datasets in the multiple class set of data is shown in Table 1. In both binary & multiple class sets of data, it is clear that not all categories are equally represented. Seeing the consequences of having proportional participation among the categories, analysis with weighted categories could be performed. The balanced weighted categories option is open for the systems Support vector, radio frequency, Artificial Neural Network, Drive test, & LR and it relates to a category weights as regards:

$$p(q_j) = \begin{Bmatrix} J_j\left[Q_a - \frac{a_p}{2} : Q_a + \frac{a_p}{2} : Q_u - \frac{u_p}{2} : Q_u + \frac{u_p}{2}\right], \\ J_{reference}\left[Q_a - \frac{a_p}{2} : Q_a + \frac{a_p}{2} : Q_u - \frac{u_p}{2} : Q_u + \frac{u_p}{2}\right] \end{Bmatrix} \quad (1)$$

$$F_2\,(J, J^{reference)} = \frac{\sum_{j=1}^{a}\sum_{k=1}^{u}\sum_{i=1}^{v} || J_{jki} - J_{jki}^{reference} ||_2^2}{auv} \quad (2)$$

$$F_{\frac{1}{2}}\,(J, J^{reference)} = \frac{\sum_{j=1}^{a}\sum_{k=1}^{u}\sum_{i=1}^{v} || J_{jki} - J_{jki}^{reference} ||^{\frac{1}{2}}}{auv} \quad (3)$$

$$L_{per}(J, J^{reference)} = \sum_{m=1}^{M} \frac{||\emptyset_m(J) - \emptyset_m(J^{reference)})||_2^2}{a_m u_m c_m} \quad (4)$$

$$a_j = \frac{F(J_{j-1}, J_{reference}) - F(J_j, J_{reference})}{F(J_p, J_{reference})} \quad (5)$$

$$K(\pi) = \sum_{w=1}^{w_{maximum}} a_d \gamma_d \quad (6)$$

$$d_{maximum} = arg_j minimum(a_j > a_{threshold}) \quad (7)$$

wherein Samples is the amount of rows in the set of data, Categories denotes the amount of categories, & Y denotes the amount of variables.

Table 1: Data Representation with Multiple Classes

Dataset	Data Training	Data Test	Total
DDoS HTTP	23	1016	1039
DDoS TCP	11065	1065	12130
DDoS UDP	22356	54982	77338
OS Scan	7896	45632	53528
Service Scan	22198	4632	26830

Implementation

For such development of machine learning methods, they employ the Python 3.7.4 computer language. Sklearn & Keras are the 2 significant components that are utilised to construct the models. The Artificial Neural Network is implemented with Keras, while another types are implemented with sklearn. This should be mentioned that they utilised the default settings of hyper - parameters for every classifiers of comparative purposes. **Table 2** lists the names of the components utilised as well as a brief description of each.

The set of data contains characteristics which either includes no data / contain traces which was irrelevant to the systems classification of the information. The pandas module could be used to eliminate undesired characteristics and during pre - processing stage. Many features was omitted from the set of data, including flgs, prototype, directory, doui, dmac, saddr, smac, soui, sco, daddr, state, srcid, document, group, & sub-category.

Assessment consumes 20% of the information, which is often a significant amount of information. But, if the set of data is tiny, this might lead to a lack of testing data as well as the false impression that the system had performed exceptionally well when, in reality, it's not been well evaluated. Split of train test from the Python component model selection could be utilised to split the set of data into training & testing data. The arbitrary state

variable could be utilised to set the seed of the pseudorandom producer when utilising this method; in this instance, the amount 121 were utilised.

RESULTS AND DISCUSSION

This part comprises all of the outcomes & analyses based on multiple performance measures, combining binary & multiple class test dataset, to determine which Machine learning algorithms are the greatest & worst of categorising assault information on Internet of Things networks. Exfiltration: **Table 2** illustrates the findings for exfiltration information, showing that radio frequency, had the best ratings across the board, including log-loss. Drive test had flawless ratings as well, but it had a significant log loss, showing which a radio frequency, system is more confidence with it's projections.

Table 3 depicts the Radiofrequency confusion matrix, as well as two key pieces of data. The first is that volume of data evaluated is insignificant, as well as the categories are not represented equally. It's likely that the outcomes are influenced by the tiny portion of test data available. In comparison to radio frequency, the other systems, with the exception of Drive test, had comparatively low ratings.

Table 3 demonstrates that raising the testing data to 30percentage points reduces the log-loss, implying which the performance is good with additional information, albeit modestly. The system was able from retain flawless recollect with-up from the 50percentage divide in a learning & analysis information, but outcomes start to deteriorate as the test data approaches 40percent and beyond.

The weighted categories argument might be utilised since the class description is unbalanced. This enables for the correction of the gap between categories, as seen in **Table 4**. When utilising the K-Nearest Neighbor & NB models, this feature is not available. **Table 4** shows that by utilising balanced groups, SVM's performance has improved, with all metrics improving & log loss dropping. Balanced courses have no effect on Artificial Neural Network, but they do have a minor impact on LR, which has perfect accuracy but reduced recall. Drive test loses perfect rates, whereas RF maintains perfect rates while growing slowly log loss.

As compared to Drive test, radio frequency, is the right model without employing balanced classes because of its minimal log loss. With perfect rates as well as a minimal log loss, radio frequency, probably is the best method if weighted classes are utilised, demonstrating that the system is confidence in its predictions.

Model Comparison

The optimal parameters for every the datasets, also including, are shown in **Table 5**. The most common models in the **Table 5** were Drive test & radio frequency, with Artificial Neural Network showing frequently as in weighted groups column. radio frequency, gets the highest results without utilise of weighted groups. Artificial Neural Network gets the highest results with weighted groups. Utilising weighted groups, on the other hand, reduces the model's overall effectiveness.

Table 2: Matrix of DDoS HTTP DT Confusion

Label of Actual		
Label of Prediction	**No Attack**	**Attack**
No Attack	**8**	**1**
Attack	**1**	**3590**

Table 3: Matrix of DDoS HTTP RF Confusion

Label of Actual		
Label of Prediction	**No Attack**	**Attack**
No Attack	**12**	**1**
Attack	**1**	**5892**

Table 4: Description of the modules that were utilised

Module	Description
Numpy	**Used to store the dataset in an array**
Pandas	**Used to read the dataset CSV file**
Preprocessing	**Used to normalize feature data**
Model-Selection	**Used for Splitting the training and test**
Metrics	**Contains the performance metrics**
Neighbors	**Contains KNN model**
SVM	**Contains SVM model**
Tree	**Contains DT model**
Naive bayes	**Contains NB model**
ensemble	**Contains RF model**

Table 5: Features and Descriptions of the Dataset

Features	Description
Stime	**Record Start Time**
Sport	**Port that data is being sent from**
Dport	**Port that data is being received from**
Pkts	**Total number of packets**
Bytes	**Total number of Bytes**
Ltime	**Record last time**
Seq	**Sequence Number**
Dur	**Record Total Duration**
Mean	**Average Duration**
Sum	**Aggregated Records**

CONCLUSIONS

Both on weighted & non weighted Bot- Internet of Things datasets, state for a art Machine Learning program was comparison on the basis for reliability, reality, recollect, F1 rate, & log-loss. With non weighted set of data, it is proven that radio frequency, performs better in terms of precision and accuracy. Artificial Neural Network, on the other hand, has a better accuracy of binary classification in a weighted set of data. For weighted and non-

weighted datasets, K-Nearest Neighbor & Artificial Neural Network was extremely accurate in multiple classification. If all kinds of attacks had weighted sets of data, the findings show that Artificial Neural Network identifies the type of assault with better accuracy.

REFERENCES

[1] Kiran KS, Devisetty RK, Kalyan NP, Mukundini K, Karthi R. Building a intrusion detection system for iot environment using machine learning techniques. Procedia Computer Science. 2020 Jan 1;171:2372-9.

[2] Li Y, Xu Y, Liu Z, Hou H, Zheng Y, Xin Y, Zhao Y, Cui L. Robust detection for network intrusion of industrial IoT based on multi-CNN fusion. Measurement. 2020 Mar 15;154:107450.

[3] Ullah I, Mahmoud QH. Design and development of a deep learning-based model for anomaly detection in IoT networks. IEEE Access. 2021 Jul 1;9:103906-26.

[4] Nivaashini M, Thangaraj P. A framework of novel feature set extraction based intrusion detection system for internet of things using hybrid machine learning algorithms. In2018 international conference on computing, power and communication technologies (GUCON) 2018 Sep 28 (pp. 44-49). IEEE.

[5] Alsariera YA. Detecting Generic Network Intrusion Attacks using Tree-based Machine Learning Methods. Inter. J. of Adv. Comp. & Science and Applications. 2021;12(2):597-603.

[6] da Costa KA, Papa JP, Lisboa CO, Munoz R, de Albuquerque VH. Internet of Things: A survey on machine learning-based intrusion detection approaches. Computer Networks. 2019 Mar 14;151:147-57.

[7] Gao X, Shan C, Hu C, Niu Z, Liu Z. An adaptive ensemble machine learning model for intrusion detection. IEEE Access. 2019 Jun 19;7:82512-21.

[8] Karatas G, Demir O, Sahingoz OK. Deep learning in intrusion detection systems. In2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT) 2018 Dec 3 (pp. 113-116). IEEE.

[9] Salloum SA, Alshurideh M, Elnagar A, Shaalan K. Machine Learning and Deep Learning Techniques for Cybersecurity: A Review. InAICV 2020 Mar 23 (pp. 50-57).

[10] Baraneetharan E. Role of machine learning algorithms intrusion detection in WSNs: a survey. Journal of Information Technology. 2020;2(03):161-73.

[11] Taher KA, Jisan BM, Rahman MM. Network intrusion detection using supervised machine learning technique with feature selection. In2019 International conference on robotics, electrical and signal

processing techniques (ICREST) 2019 Jan 10 (pp. 643-646). IEEE.

[12] Kalimuthan C, Renjit JA. Review on intrusion detection using feature selection with machine learning techniques. Materials Today: Proceedings. 2020 Jan 1;33:3794-802.

[13] Ezhilarasi, T. P., Dilip, G., Latchoumi, T. P., & Balamurugan, K. (2020). UIP—A Smart Web Application to Manage Network Environments. In *Proceedings of the Third International Conference on Computational Intelligence and Informatics* (pp. 97-108). Springer, Singapore

[14] RM SP, Maddikunta PK, Parimala M, Koppu S, Gadekallu TR, Chowdhary CL, Alazab M. An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture. Computer Communications. 2020 Jul 1;160:139-49.

[15] Verma A, Ranga V. Machine learning based intrusion detection systems for IoT applications. Wireless Personal Communications. 2020 Apr;111(4):2287-310.

[16] Ravipati RD, Abualkibash M. Intrusion detection system classification using different machine learning algorithms on KDD-99 and NSL-KDD datasets-a review paper. International Journal of Computer Science & Information Technology (IJCSIT) Vol. 2019 Jun;11.

[17] Oliveira N, Praça I, Maia E, Sousa O. Intelligent cyber attack detection and classification for network-based intrusion detection systems. Applied Sciences. 2021 Jan;11(4):1674.

[18] Rahman MA, Asyhari AT, Leong LS, Satrya GB, Tao MH, Zolkipli MF. Scalable machine learning-based intrusion detection system for IoT-enabled smart cities. Sustainable Cities and Society. 2020 Oct 1;61:102324.

[19] Kumar V, Das AK, Sinha D. UIDS: A unified intrusion detection system for IoT environment. Evolutionary Intelligence. 2021 Mar;14(1):47-59.