iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45697

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Attack and Anomaly Detection in IoT Sites Using Machine Learning Techniques

Aziza Khalilahmed Shaikh[1], Prof Govind Negalur[2]
[1]*PG Scholar,* [2]*Professor, Department of Computer Science Engineering, SDM College of Engineering & Technology, Dharwad, India*

Abstract: ***A growing problem in the IoT space is the attack and anomaly detection in the infrastructure of the Internet of Things (IoT). Every domain is using IoT infrastructure more and more, and with that use comes a surge in risks and attacks against those infrastructures. Such attacks and anomalies that can lead to an IoT system failure include Denial of Service, Data Type Probing, Malicious Control, Malicious Operation, Scan, Spying, and Wrong Setup. Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) are the machine learning (ML) methods that have been employed in this. Accuracy, precision, recall, f1 score, and area under the receiver operating characteristic curve are the evaluation measures used in performance comparison. For Decision Tree and Random Forest, the system received test accuracy results of 99.4 %. Despite the same accuracy of these algorithms, other criteria show that Random Forest performs significantly better.***
Keywords: Anomaly, Attack, Logistic Regression, Decision Tree, Random Forest, Internet of Things

I. INTRODUCTION

Anomaly means identifying the objects, items or observations that are dubious. Finding unusual occurrences or observations that are suspicious because they deviate considerably from expected patterns or behaviour is known as anomaly detection. The Internet of Things (IoT) is presently experiencing a period of fast expansion. The Internet of Things is anticipated to become the "next big thing" in the coming years, according to analysts. People are becoming used to infrastructure that is data-driven, and this is pushing greater research into machine learning-based applications for the Internet of Things. Today, technologies based on the Internet of Things and machine learning are employed in every aspect of human existence. Image recognition - Give a name to a face in a photo (also known as "tagging" on social media). Speech Recognition - appliance management. Example: Alexa. Medical Diagnosis - Machine learning is used in oncology and pathology to identify malignant tissue. Statistical arbitrage - Manage a huge number of securities by analysing massive data sets utilised in finance. Predictive Analysis - determining if a transaction is genuine or fraudulent. Security risks and anomalies with IoT devices are becoming increasingly frequent. An attack on a local network that uses normal communication is restricted to nearby nodes or a small local domain, whereas an assault on an Internet of Things system covers a much wider region and has catastrophic implications for IoT sites. While an attack on an Internet of Things system affects a much larger area and has disastrous effects for IoT sites, a local network attack that employs conventional communication is limited to adjacent nodes or a small local domain.

Different sources of anomalies are Intrusion Detection System, Fraud Detection and Data Leakage. Intrusion Detection System - IoT devices are susceptible to security-related assaults as they are linked to the internet. Assaults that significantly impair IoT services and apps for smart environments come into this category and include denial-of-service (DoS) and distributed denial-of-service (DDoS) attacks. Fraud detection - Credit card numbers, bank account numbers, and other sensitive data can still be stolen via IoT networks during logins and online transactions. Data Leakage - Databases, file servers, and other sources of information might expose private data to unauthorised parties. Information is lost as a result of such leaks, but there is also a potential that the attacker would remove sensitive data from the system. Utilizing the right encryption methods will stop such leaks. Anomaly detection seeks to locate each of these data points in a data-driven way. Anomalies and outliers are not always unfavourable. These are observations that don't follow the pattern of the other observations.

According to their kind, IoT system anomalies may be divided into point-wise, contextual, or collective abnormalities. Point-wise Anomaly - A point anomaly occurs when a single point stands out from the rest of the data. Collective Anomaly - A collective anomaly occurs when several connected data examples are aberrant when compared to the total data set. Contextual Anomaly - A data instance that is anomalous in a certain context is referred to as a contextual anomaly. The system's main objective is to build an intelligent, safe, and trustworthy IoT-based infrastructure that can identify its risk. Here, a machine learning-based approach that can identify whether a system is acting oddly is used [1].

II. LITERATURE REVIEW

Numerous IoT industries have undertaken research similar to this one. This study is still being conducted by researchers [2].

Mahmudul Hasan [1] have developed the system that aims to build an IoT-based infrastructure that is intelligent, secure, and trustworthy that can recognise its weaknesses, have a robust firewall against any threats, and automatically recover.

The algorithms that have been used are Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Artificial Neural Network (ANN). Accuracy, precision, recall, f1 score, and area under the receiver operating characteristic curve serve as performance comparison measurements.

The system got 99.4% accuracy results for the Decision Tree, Random Forest, and ANN.

Pahl et al. [2] have devised a detection and firewall for an anomaly of IoT microservices in IoT sites.

In this study, several microservices have been clustered using K-Means and BIRCH, respectively [7].

When clustering, separate clusters were clustered together if their centres were three times as far apart as the standard deviation. Online learning has been used to update the clustering model.

The system's total accuracy after the algorithms are in place is 96.3%.

In an industrial IoT site, Liu et al. [3] suggested a detector for On and Off attacks by a rogue network node.

A rogue node might attack an IoT network while it was in an active state, or "On state," as they described by the phrase "On and Off assault."

The system's development involved the use of a light probe routing mechanism with the computation of each neighbour node's trust estimation for the purpose of anomaly detection.

Attack detection utilising fog-to-things architecture was explored by Diro et al. [4].

Using an open source dataset, the authors of the publication conducted a comparison of deep and shallow neural networks.

The target of this experiment was to identify four kinds of attack and anomaly. In terms of accuracy, the strategy achieved 96.75 % for shallow neural network models and 98.27 % for deep neural network models for four classes.

An IoT intrusion detection system was addressed by Anthi et al. [5]. Several ML classifiers have been used to effectively recognise network scanning probing and basic Denial of Service (DoS) attacks for this purpose.

Utilizing the programme Wireshark, network traffic for four additional days is collected to create the data set.

Weka was the programme used for implementing ML classifiers.

Real Traffic Data and the binary NSLKDD dataset were subjected to D'Angelo et al [6] .'s application of Uncertainty-managing Batch Relevance-based Artificial Intelligence (U-BRAIN) (from Fredrico II University of Napoli).

A dynamic model with the ability to manage missing data, the U-Brain is run on several computers. 41 features make up the NSL-KDD dataset. Using a classification technique based on J-48, 6 features were chosen from a total of 41 features.

The accuracy percentages for NSL-KDD and Real Traffic Data, respectively, were 94.1% and 97.4% (10-fold training mean).

III. METHOD

A number of separate procedures are combined to create the overall structure.

The system's overarching structure is shown in figure 1.

This framework's dataset collecting procedure is its initial step.

The dataset was gathered and carefully scrutinised during this procedure to determine the different categories of data.

Normal values are analysed because algorithms never detect string values, so we need to convert string values to integer.

When working with Machine Learning algorithms, we need to split the data into Training data and Testing Data.

From the dataset the data is split into 82% of training data and 18% of testing data.

The learning algorithm was applied to the training set, and an optimization method was utilised to create the final model.

The many classifiers utilised in this study made use of various optimization strategies.

Logistic regression made use of coordinate descent [8].

Due to the non-parametric nature of the DT and RF models, the optimizer is not utilised in these cases. Using various evaluation measures, the resulting model was compared to the testing set.[1]

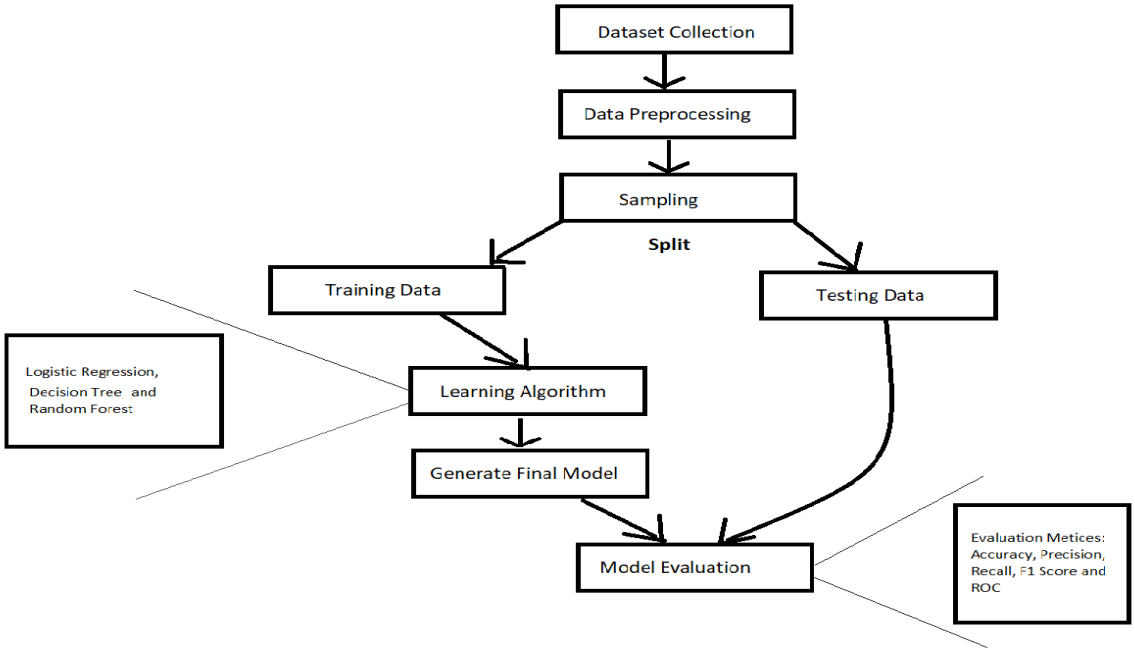


Fig. 1 Overall framework for attack and anomaly detection in IoT

A. Dataset Collection and Description

The publicly available data collection was gathered from Kaggle and contributed by FrancoisXA. To generate synthetic data, they built a virtual Internet of Things (IoT) environment using the Distributed Smart Space Orchestration System (DS2OS). Communications between various IoT nodes are included in this collection. Dataset is collected using four simulated IoT locations with various services, including thermostats, washing machines, batteries, movement sensors, smart doors, and smart phones. This dataset was developed to evaluate anomaly detection techniques. Table 1 depicts the distribution of various assaults and anomalies over the whole dataset in detail. Besides these, Fig. 2 depicts the frequency distribution of a number of features [1]

TABLE I
DISTRIBUTION OF REGARDED FREQUENCY

SN.	Attacks	Frequency Count	% of Total Data	% of Anamolus Data
1	Denial of Service	5780	01.61%	57.70%
2	Data Type Probing	342	00.09%	03.41%
3	Malicious Control	889	00.24%	08.87%
4	Malicious Operation	805	00.22%	08.03%
5	Scan	1547	00.43%	15.44%
6	Spying	532	00.14%	05.31%
7	Wrong Setup	122	00.03%	01.21%

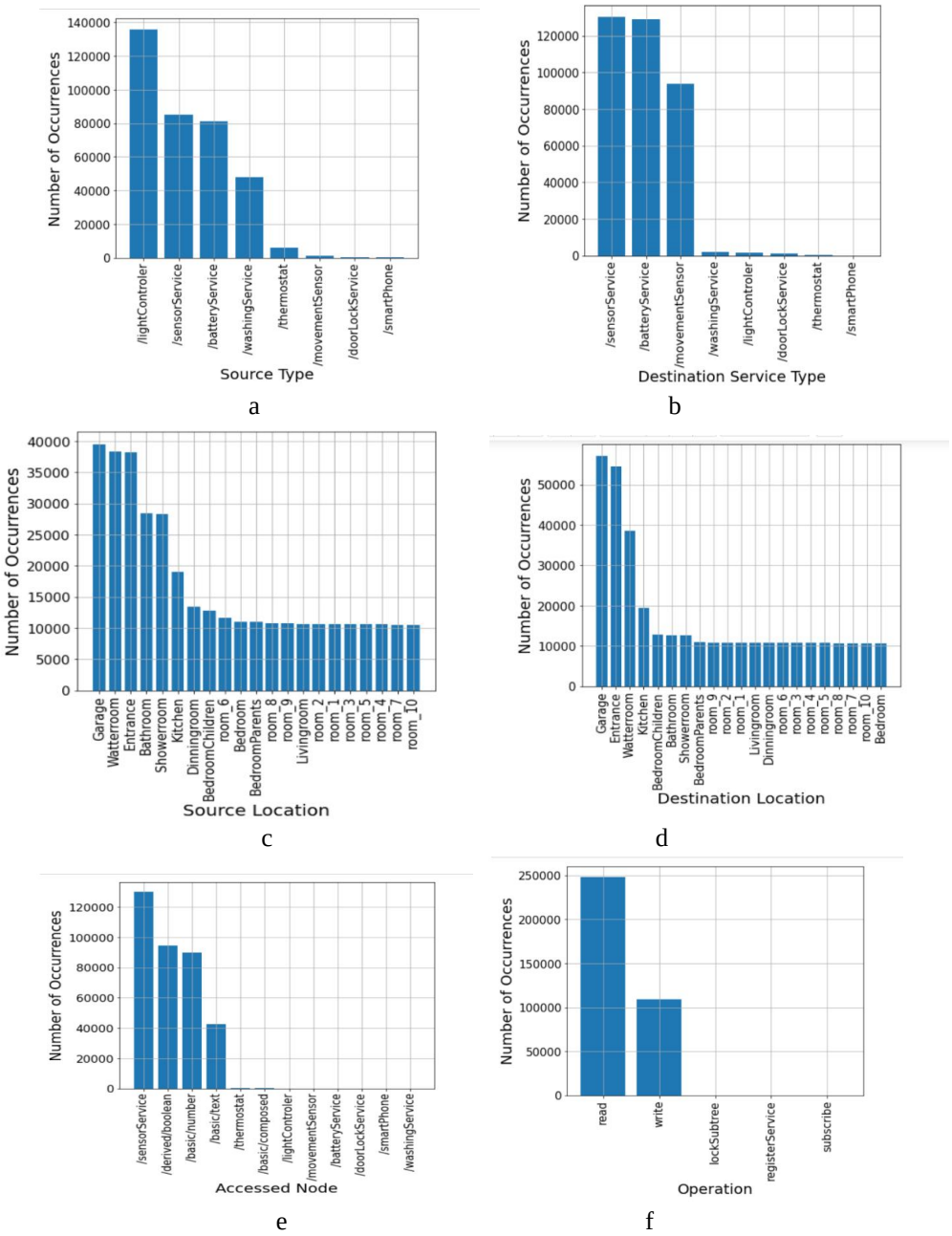


Fig. 2. Frequencies of (a) Source Type (b) Destination Type (c) Source Location (d) Destination Location (e) Accessed Node (f) Operation

B. Attacks

1) Denial of Service (DOS) - Too much unsolicited traffic coming from one source or recipient leads to a DoS attack. The attacker bombards the victim with too many confusing packets, rendering its services inaccessible to other services [9]. There are 5780 samples of Denial of Service in the dataset.

2) Data Type Probing (D.P): Here, a rogue node writes data that isn't the specified data type [2]. There are 342 samples of Data Type Probing in the dataset.

3) Malicious Control (M.C): Through software flaws, an attacker may occasionally be able to get a functioning session key or intercept network data. Using this technique, a hostile party can take control of the entire system [10, 11]. There are 889 samples Malicious Control in the dataset.

4) Malicious Operation (M.O): Malware is usually the root of malicious operations. Malware refers to bogus activity that diverts attention from the primary action [12]. There are 342 samples of Malicious Operation in the dataset.
5) Scan (SC): When data is obtained by hardware and the system is scanned, this procedure occasionally results in damaged data [13]. There are 1547 samples of Scan in the dataset.
6) Spying (SP): Utilizing a backdoor route to get access to the system, the attacker spying uses the system's flaws to their advantage and gathers crucial data [4]. There are 532 samples of Spying in the dataset.
7) Wrong Setup (W.S): Incorrect system configuration might potentially cause the data to become corrupt [14]. There are 122 samples of Wrong Setup in the dataset.
8) Normal (NL): The term "normal data" refers to information that is totally true and accurate [1]. There are 347935 samples of normal data in the dataset.

C. Theoretical Consideration

Several machine learning methods were employed for the data analysis portion. Here are lists of algorithms along with a brief description of each.

1) *Logistic Regression:* A discriminative model that is dependent on the calibre of the dataset is logistic regression (LR). The equation for estimating the posterior is presented below, taking into account the features X = X1, X2, X3,..., Xn (where X1 Xn = Distinct f eatures), weights W = W1,W2,W3,...,Wn, bias b = b1, b2,..., bn, and Classes C = c1, c2,..cn [15]

$$\text{Predicted Value: } p(y = C|X; W, b) = \frac{1}{1 + \exp(-W^{\text{transpose}}X - b)}$$

2) *Decision Tree (DT):* Each node in a decision tree is given the ability to compare potential courses of action based on the advantages, costs, and probabilities associated with each. Overall, it represents a road map of potential consequences of a number of connected decisions [19]. A DT typically begins with a single node before branching out to other possibilities. Each of these results generates new nodes that branch out into other instances. The objective of DT is to maximise the following Information Gain.

$$Information\ Gain(P_a, x) = I(\mu_{rl}) - \frac{LC_n}{P_n}(LC_d) - \frac{RC_n}{P_n} I(RC_{rl})$$

There are three methods for calculating the Impurity Measure I(data). Gini Index, Entropy, and Classification Error are three metrics.

$$I_H(n) = -\sum_{i-1}^{c} p(c|n) log_2 p(c|n)$$

$$I_G(n) = 1 - \sum_{i=1}^{c} p(c|n)^2$$

$$I_E(n) = 1 \quad \max\{p(c|n)\}$$

Where p(c|n) represents the ratio of c to n and c symbolises classes or labels, n denotes any node [1].

3) *Random Forest (RF):* The random forest method, as the name suggests, produces a forest with several decision trees. It is an algorithm for supervised classification. A random forest is made up of several decision trees that are ensembled together, and it makes predictions by averaging the predictions of each component tree [21]. Compared to just one decision tree, it often offers substantially higher prediction accuracy.

D. Evaluation Criteria

The developed system's performance was measured using the metrics listed below. These metrics can be used to determine which approach is most appropriate for this task.

1) *Confusion matrix:* A technique's performance may be seen using the confusion matrix. It is a table that is frequently employed to summarise how well a classification model performs when applied to a set of test data for which the real values are known. A classification problem's prediction outcomes are summarised in a confusion matrix [22].
   True positive (TP): An observation that was expected to be positive and turned out to be such.
   False positive (FP): An observation that was expected to be positive but turned out to be negative.
   True negative (TN): True negative observations are those that are both expected and real negative.
   False negative (FN): False negative observations are those that are projected to be negative but are really positive.

2) *Accuracy:* A model's performance includes more factors than just accuracy. One parameter for assessing classification models is accuracy. Single class accuracy measurement is shown in equation (14).

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

3) *Precision:* Precision is the ability to make accurate predictions. It is a measurement of the ratio of the model's claimed positives to the number of genuine positives. The following equation gives the precision value for a single class:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

4) *Recall:*The recall is often referred to as the real positive rate, which measures the proportion of positives in model assertions to the overall positive rate of the data. The following equation provides the recall value for a single class:

$$Recall - \frac{True\ Positive}{True\ Positive + False\ Negative}$$

5) *F1 score:* A model's performance may also be evaluated using the F1 score. It is a weighted average of a model's recall and accuracy. The following equation gives the F1 Score value for a single class:

$$F1\ Score - \frac{2 * True\ Positive}{2 * True\ Positive + False\ Positive + False\ Negative}$$

6) *Receiver operating characteristic curve:*It is a widely used graph that summarises a classifier's performance over all feasible thresholds. It is created by graphing the True Positive Rate vs the False Positive Rate while the threshold value is changed to categorise observations [22]. The following equation provides the computation of True Positive Rate and False Positive Rate.

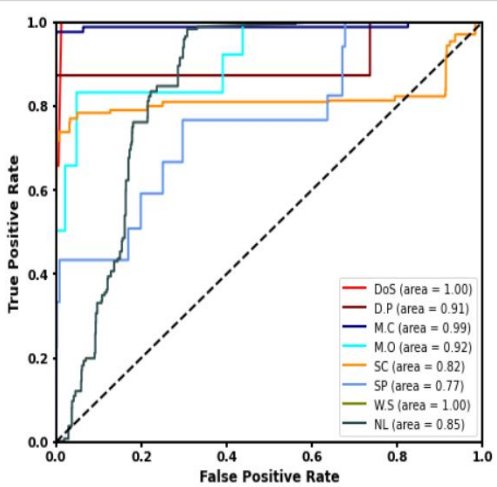
$$False\ Positive\ Rate = \frac{Number\ of\ False\ Positive\ Samples}{Total\ Number\ of\ Samples}$$

$$True\ Positive\ Rate = Recall = \frac{Number\ of\ True\ Positive\ Samples}{Total\ Number\ of\ Samples}$$

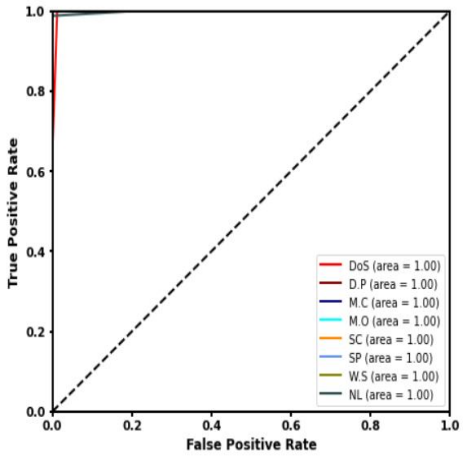
E. Result Analysis

Several Machine Learning techniques have been applied to the dataset. Five-fold cross-validation has been applied to each of the technique. It is clear from the cross-validation that RF have demonstrated the best accuracy in both training and testing. In terms of training, DT's performance was roughly comparable to that of RF. But with testing, DT performed poorly at initially and had more deviations than other approaches. However, it performed similarly to RF in the final three folds. Compared to other techniques, LR had poorer training results. In the testing scenario during the first two folds, LR outperformed other techniques, with logistic regression coming out on top, however at the final three folds, they underperformed.

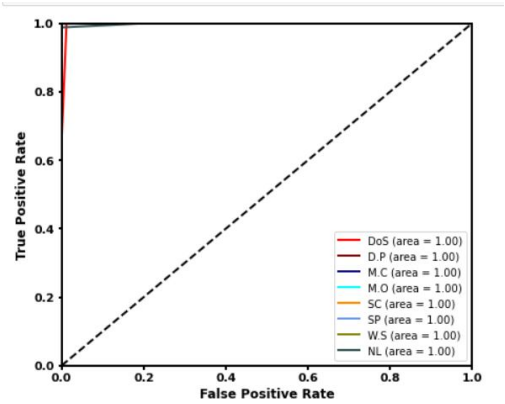
RF is the best technique for this work, according to the confusion matrices.



a

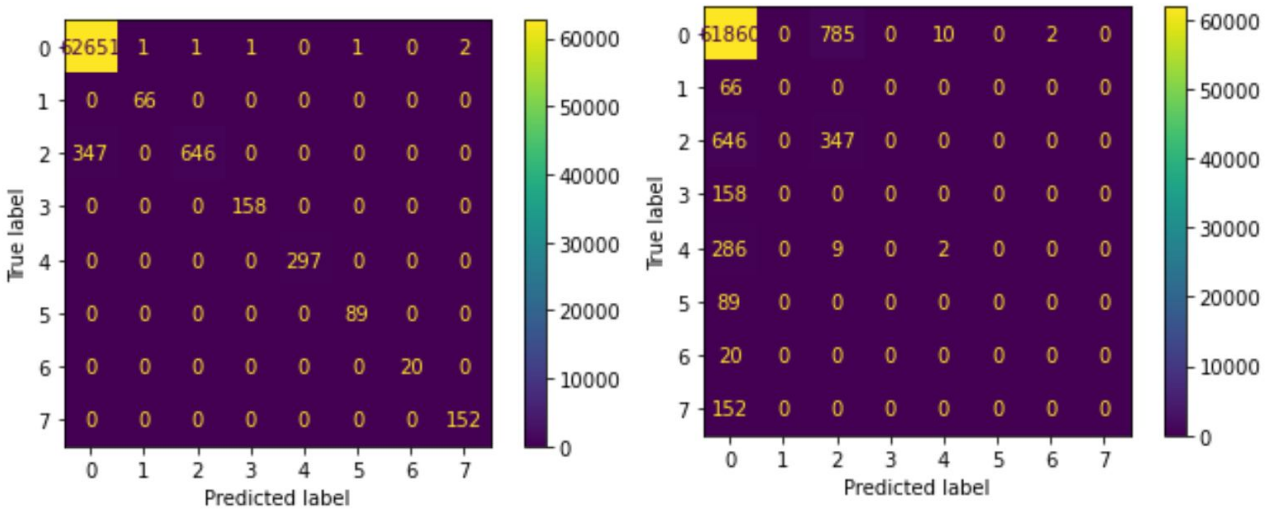


b

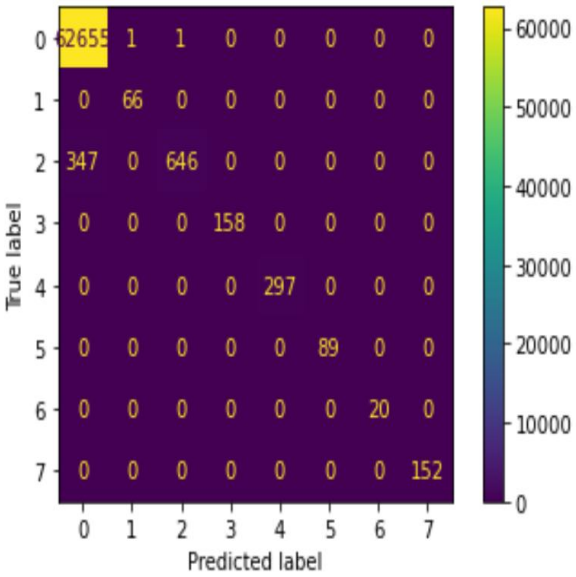


c

Fig. 3. ROC Curve of (a) Logistic Regression (b) Decision Tree (c) Random Forest



a b



c

Fig. 4. Confusion Matrix of (a) Decision Tree (b) Logistic Regression (c) Random Forest

IV. CONCLUSION

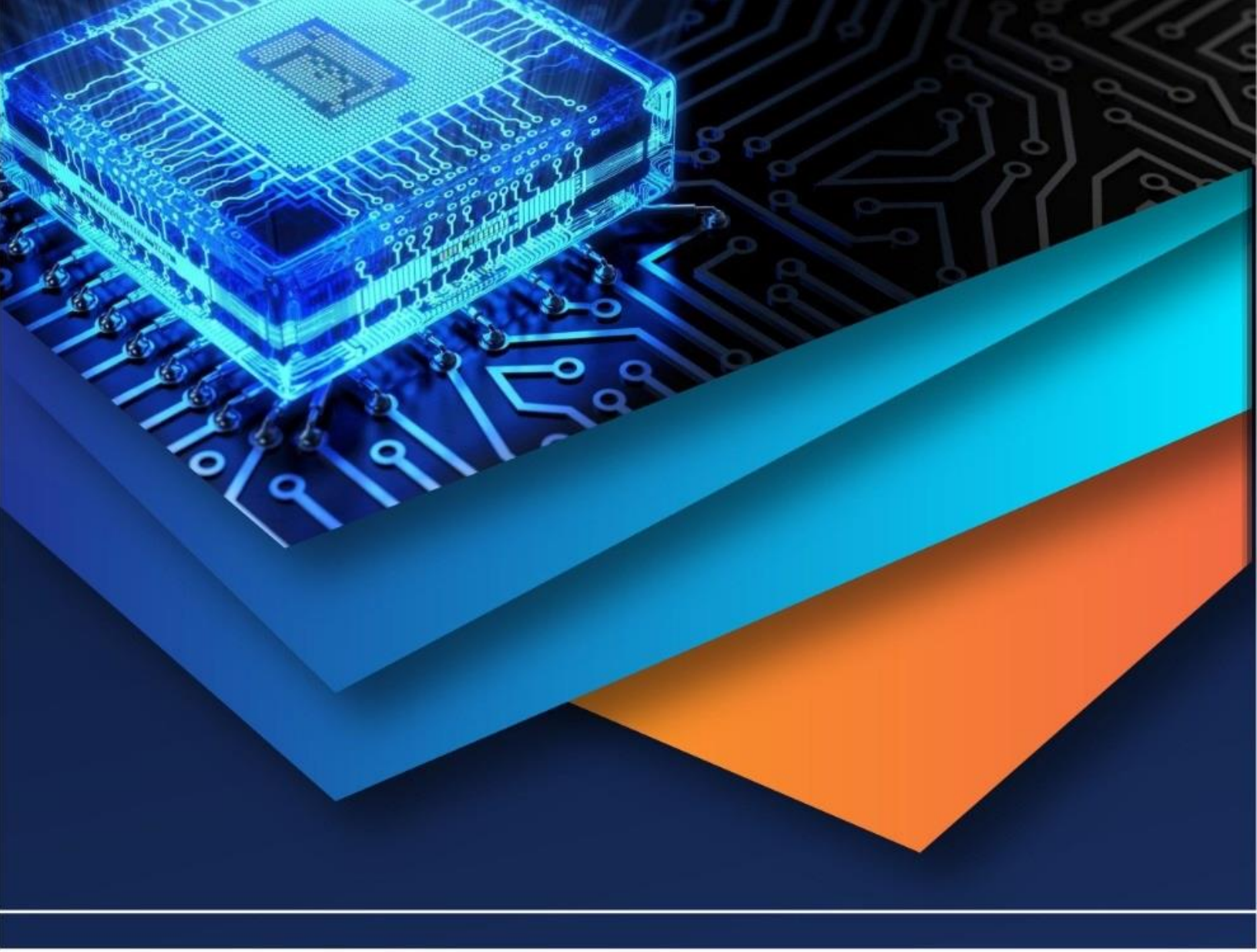
Based on the results of the entire study, it was determined that Random Forest technique should be used on these types of datasets to address assaults on IoT networks since, in comparison to other methods, RF accurately predicted Data Probing, Malicious Control, Malicious Operation, Scan, Spying, and Wrong Setup attacks. However, only traditional machine learning techniques are used on the dataset in this instance, and a comparative study is provided. On this dataset, no additional algorithm is developed. Therefore, more research is required to create a reliable detection method. More research should be done on designing the entire framework. Additionally, this work is based on data from a virtual environment. Additionally, data from a virtual environment was used to create this work. Different issues could arise when dealing with real-time data. This issue requires a more in-depth empirical investigation that focuses on real-time data. Microservices in the IoT network operate differently at different times, which leads to deviations from the usual behaviour of IoT services and ultimately creates an anomaly. To interpret these issues in greater detail, more research is required. With an accuracy of 99.4%, Random Forest performs significantly better in this study. However, it does not guarantee that Random Forest will function in this manner in the case of huge data or other unidentified issues. More research will therefore be required.

V. ACKNOWLEDGMENT

We appreciate FrancoisXA, Chyneya, Sahar Lazem for contributing the open-source dataset we used for this research.

REFERENCES

[1] [Mahmudul Hasan, Md. Milon Islam, Md Ishrak Islam Zarif, M.M.A. Hashem, Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet of Things 7 (2019) 100059
[2] M.-O. Pahl, F.-X. Aubet, All eyes on you: distributed multi-dimensional IoT microservice anomaly detection, in: Proceedings of the 2018 Fourteenth International Conference on Network and Service Management (CNSM) (CNSM 2018), 2018. Rome, Italy.
[3] X. Liu, Y. Liu, A. Liu, L.T. Yang, defending on–off attacks using light probing messages in smart sensors for industrial communication systems, IEEE Trans. Ind. Inf. 14 (9) (2018) 3801–3811.
[4] A.A. Diro, N. Chilamkurti, Distributed attack detection scheme using deep learning approach for internet of things, Future Gen. Comput. Syst. 82 (2018) 761–768. 14 M.
[5] E. Anthi, L. Williams, P. Burnap, Pulse: an adaptive intrusion detection for the internet of things (2018).
[6] G. D'Angelo, F. Palmieri, M. Ficco, S. Rampone, An uncertainty-managing batch relevance-based approach to network anomaly detection, Appl. Soft Comput. 36 (2015) 408–418.
[7] C.C. Aggarwal, J. Han, J. Wang, P.S. Yu, A framework for clustering evolving data streams, in: Proceedings of the Twenty-ninth International Conference on Very Large Data Bases-Volume 29, VLDB Endowment, 2003, pp. 81–92.
[8] Z. Allen-Zhu, Z. Qu, P. Richtárik, Y. Yuan, even faster accelerated coordinate descent using non-uniform sampling, in: Proceedings of the International Conference on Machine Learning, 2016, pp. 1110–1119.
[9] O. Brun, Y. Yin, E. Gelenbe, Y.M. Kadioglu, J. Augusto-Gonzalez, M. Ramos, Deep learning with dense random neural networks for detecting attacks against IoT-connected home environments, in: Proceedings of the 2018 ISCIS Security Workshop, Imperial College London. Recent Cybersecurity Research in Europe. Lecture Notes CCIS, in: 821, 2018.
[10] J. Liu, Y. Xiao, C.P. Chen, Authentication and access control in the internet of things, in: Proceedings of the 2012 Thirty-second International Conference on Distributed Computing Systems Workshops (ICDCSW), IEEE, 2012, pp. 588–592.
[11] Z.-K. Zhang, M.C.Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, S. Shieh, Iot security: ongoing challenges and research opportunities, in: Proceedings of the 2014 IEEE Seventh International Conference on Service-Oriented Computing and Applications (SOCA), IEEE, 2014, pp. 230–234.
[12] J. Milosevic, N. Sklavos, K. Koutsikou, in: Malware in IoT software and hardware, 2016.
[13] W. Ding, Study of smart warehouse management system based on the IoT, in: Proceedings of the Intelligence Computation and Evolutionary Computation, Springer, 2013, pp. 203–207.
[14] W. Leister, T. Schulz, Ideas for a trust indicator in the internet of things, in: Proceedings of the First International Conference on Smart Systems, Devices and Technologies, SMART, 2012, pp. 31–34.
[15] U.S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, M. Stanley, A brief survey of machine learning methods and their sensor and IoT applications, in: Proceedings of the 2017 Eighth International Conference on Information, Intelligence, Systems and Applications (IISA), IEEE, 2017, pp. 1–8.
[16] L. Wang, Support Vector Machines: Theory and Applications, 177, Springer Science & Business Media, 2005.
[17] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[18] C.A. Burges, A tutorial on support vector machines for pattern recognition, Data Mining Knowl. Discov. 2 (1998) 121–167.
[19] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Trans. Syst. Man. Cybern. 21 (3) (1991) 660–674
[20] S. Raschka, Python Machine Learning, Packt Publishing Ltd, 2015.
[21] T.K. Ho, Random decision forests, in: Proceedings of the third international conference on Document analysis and recognition, 1995, 1, IEEE, 1995, pp. 278–282.
[22] M. Hossin, M. Sulaiman, A review on evaluation metrics for data classification evaluations, Int. J. Data Mining Knowl. Manag. Process 5 (2) (2015).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY