



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 2, February 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.542



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

IOT Based Intrusion Detection System using ML Techniques

Vishal Kaushik¹, Swati Goel², Dr. A.K. Gautam (Guide)³, Dr. Pragati Sharma (Co-Guide)⁴

M.Tech Student, Department of Electronics & Communication, S.D. College of Engineering & Technology, Muzaffarnagar, U.P., India ¹

Former Assistant Professor, Department of Computer Science, AKG Engineering College, Ghaziabad, U.P., India²

Principal, Department of Electronics & Communication, S.D. College of Engineering & Technology, Muzaffarnagar, U.P., India ³

Professor, Department of Electronics & Communication, S.D. College of Engineering & Technology, Muzaffarnagar, U.P., India ⁴

ABSTRACT: In the era of unaccountable advanced smart services, each person is surrounded by plenty of smart devices and comes in contact with these on a daily basis. The enhancement of IoT leads to more security challenges as compared to before as it also introduces different types of attacks and its severity is concerned for combating the cybersecurity attacks and threats that target them, including malware, privacy breaches, and denial of service attacks, among others. To target and analyze such challenges, we introduce a model that uses machine learning algorithms and can be integrated with edge devices to detect such anomalies. The proposed model has been applied to the KDD Cup 1999 Data to validate its ability to detect attacks. These satisfactory objectives can be demonstrated through the result obtained from the proposed model.

I. INTRODUCTION

The Internet of Things is diversifying in all places, offering a variety of benefits to almost every aspect of our lives, such as health, entertainment, commerce, industry, intelligent institutions, and workplaces. Every device can connect to the Internet and connect to the web or mobile devices, or share data (Atzori, L., Iera, A., & Morabito, G., #). This developing IoT technologies deal with various security attacks and Denial of Service (DOS) attacks, security threats to resources, and disruption of IoT networks and devices as well as the connections between them, must be protected: massive amounts of data can be shared between devices, as well as between customers, and also the compromised availability and confidentiality of the information which can have a significant impact. Along with all the other benefits, the Internet has also developed many ways to jeopardize the stability and security of the systems and servers connected to it. Although static security measures such as security walls and software development can provide a decent level of security, more robust security measures such as intrusion detection systems should also be used.

Machine learning algorithms offer a viable alternative for securing IoT devices. ML is a powerful artificial intelligence technology that can outperform dynamic networks and does not require precise coding. Machine Learning and deep learning approaches may be used to train the machine that can detect various anomalies and provide appropriate defence measures and insights. Hence, the attacks and threats can be detected at a very early stage. Moreover, ML techniques turn out to be promising in detecting new types of attacks using learning skills and effectively handling them intelligently.

We propose a combination of already built architecture and ML strategies that provides real time detection and protection of IoT security attacks, which generates an improved architecture. This improved architecture can be attached with any model-driven solution to provide graphical definition of the defensive threats. This can be done using ESB, a middleware. Its function is to create data connections between IoT networks and ML algorithms. Not only limited to this, we can even connect to third parties such as servers, cloud, and alarming sources using ESB. We are working on expanding the forecast with the ability to automatically detect which package features are important in our domain context. The improved architecture with ML techniques to an IoT network prototype was constructed on KDD Cup 1999 Data. The data contains a standard set of data to be analyzed, which includes a variety of simulated interventions in the military network environment.

This dataset contains 41 features, 5 different attacks (DOS, normal, probe, r2l, u2r), and 3 protocol types (ICMP, TCP, UDP). The other way which can be used to generate the dataset to study and design a model is by simulating the data using the help of Node-Red-Mqtt.

Different machine learning algorithms can be used and then verified to categorize the data as normal and abnormal threat data. Evaluation of the classifiers (ml models) is done on the accuracy, error rate, recall, precision, score rate, and F1 score. The K-fold method is used for data sampling which generates 5 independent sets: training (80% of data) and testing classifiers (20% of data), where each set contains all kinds of attacks. Confusion Matrix is generated for every classifier implemented. The result generated provides the potential capability of four classifiers models being tested. E.g.: SVM, naive Bayes, decision tree, Logistic Regression. The paper is arranged to study the different attacks on the IoT Networks and present an overview of the security of IoT and why it is important is also illustrated. The analysis of the dataset on ML-based security is also discussed and going to study and try to build a better model to detect anomalies in the IoT Network.

II. LITERATURE REVIEW

Rüdiger Gad [1] Various uses of events in network research and surveillance were investigated, as well as four actual use-cases and how they may be addressed utilizing CEP and event pattern matching.

A model driven strategy for real-time decision making in SOA 2.0. The Knowledge-Based Systems research calculates the standard deviations in relation to the detection rate of each assault type to assess the methods' performance and resilience. Different predictive machine learning algorithms were examined, including Decision Trees and Random Forests, as well as probabilistic techniques like Naive Bayes and Gaussian. Farnaz Gharibian and Ali A. Ghorbani discovered that probabilistic processes are more robust than predictive ones when trained utilizing a variety of training datasets. [2]

This paper gives a detailed assessment of ML based security solutions for the Internet of Things. The major purpose of the research was to look at various types of security attacks, attack surfaces with repercussions, different types of machine learning algorithms, and machine learning security solutions. In addition, there is a comparison of numerous supervised and unsupervised learning approaches. [3]

Creating an Intrusion Detection System for an IoT Environment Using ML Techniques Mr. R. Karthi and his colleague's designed adversarial systems to produce attacks using Node MCU and DHT11, and devised and built a machine-learning approach to recognize and categorize network assaults (humidity and temperature sensor). They were able to develop the most accurate decision tree model possible. [4]

In the paper, they propose a hybrid approach and a unique framework model to solve Bot-IoT attacks and IoT traffic detection in a smart city. They investigated five well known ML classifiers and used a shoddy data mining tool to do so. All of the chosen ML classifiers are run on the Weka application using ten-fold cross-validation. For identifying abnormalities and intrusions in IoT networks, the Naive Bayes ML algorithm was demonstrated to be far better than the other ML methods. [5]

Mohamed Faisal Elrawy did a comparative analysis of the most current IDSs developed for the IoT paradigm, concentrating on the methodology, features, and processes that were relevant. This research looked at a number of articles. This research is focused on the design and development of IDSs for use in the IoT that is applied on smart environments. This article also looks at the IoT architecture, as well as security vulnerabilities and their interactions with the layers of the IoT architecture. [6]

In order to have a better knowledge of strategic investigations, Nadia Chaabouni and Mohamed Mosbah conducted a poll on IoT risk classification. Based on learning

approaches and state-of-the-art intrusion detection findings, a complete review of NIDS for IoT is presented. [7]

The present trends in IoT research, which are driven by applications and the need for convergence in a number of interdisciplinary technologies, as well as the overall IoT vision and technologies, are outlined in Rajkumar Buyya's study. [8]

This paper presents a study of linear regression poisoning attacks and counter measures, which includes a design for the statistical attack that need minimal knowledge of the learning process, as well as extensively evaluated attacks and

defenses on four regression models (OLS, ridge, LASSO, and elastic net) and variety of datasets from different domains. [9]

For intrusion detection systems, examine various forms, repercussions, and surface attacks on IoT networks. The major focus is on machine learning classification algorithms used to IoT system networks to increase the efficacy of identifying threats.

This paper focuses on the many types of assaults and attack classes. The accuracy of the classification model is 94.57 percent. The whole classification report contains classification based on accuracy, recall, and F1-score over the result, which ranges from 0 to 1, as well as applying the model to training and test datasets, which yields considerably more accurate results based on average macro and average weight of TCP packets. The selection of threshold points, as well as the optimization of anomaly detection, should be precise to the dataset. As demonstrated in the table above, the model accuracy was projected using Machine Learning throughout the IoT network, with a comprehensive classification chart based on data obtained from an IDS.[10]

III. DATA SOURCE

The selected Data Set for our study is KDD Cup 1999 Data. The data has a generalised collection of analysed data that generates a variety of network attacks in a military context. The intrusion detection datasets developed by KDD 99 are taken from a 1998 DARPA initiative. It provides a baseline for intrusion detection system (IDS) builders to compare different approaches.

Simulation is done through a virtual military network that includes three 'targeted' machines using various operating systems and services. There are three additional machines are then used to generate traffic by manipulating various IP addresses. And then, a sniffer uses the TCP dump format to record all network communication. The estimated total time in seven weeks. Normal communication is built on the expected profile of the military network and the attack falls

DIFFERENT LAYERS IN IOT

IoT uses a wide range of internet connections to send data from very small devices such as switches and sensors to the cloud, local feed farms, or large data platforms to make the world model more accurate.

IoT Architecture is a gateway to a variety of hardware applications, which helps to establish links and make life much easier.

Various communication systems, like Bluetooth, WiFi, LPWAN, compact and ZigBee, RFID, are adopted by different categories of IoT architectures in order to transmit and receive various data.

The physical layer, application layer, and network layer are three layers that form the standard IoT structure (Fig 1)

Application layer

Mobile and web-based applications are used by the application layer to give services to users. The application layer works as an intermediary between the IoT device and the network it will communicate with.

It controls data formatting and presentation and acts as a visual link between what an IoT device does and the data it generates is transmitted over a network. IoT applications can be smart homes, smart cities, smart health, animal tracking, etc. It is responsible for providing services to applications.

Perception/physical layer

The perception layer is the first layer in the IoT architecture, which actually works like the human eyes, ears, and nose. It has the responsibility of identifying objects and collecting information from them. RFID, 2-D barcodes and sensors are only some of the sorts of sensors that may be attached to things to gather data, which consists of the physical

(PHY) and medium access control (MAC) layers. The MAC layer creates a connection between physical devices and networks so that they may communicate properly.

Network layer

The network layer is also known as the transmission layer.

The network layer serves as a bridge between the perception layer and the application layer. Information is transmitted through this layer, which collects data from sensory material and distributes it to various levels as needed. The communication device may be wireless or phone-based. Responsibility for connecting smart devices, network devices, and networks to each other is also required. Therefore, it is very sensitive to attacks from the attackers' side. It has significant security issues regarding the integrity and authentication of information transmitted over the network.

into one of four categories. The 4 types of attacks are User to Root, Remote to Local, Denial of Service, and Probe. There are 3 parts of the selected dataset, namely “10% KDD”, “Corrected KDD”, and “Whole KDD”. The 10% of KDD dataset contains only 22 attack types which is a comparatively shorter version of the “Whole KDD” database. It contains more examples of intrusions than normal connections and the types of attacks are not equally represented. Rejection of service attacks accounts for the majority of database attacks due to their nature.

The "KDD Fixed" data set contains data with a statistical distribution that is different from "10% KDD" or "KDD Total" and it contains 14 additional attacks. The list of class labels and their corresponding categories present in the dataset used for our analysis for detecting types of attacks are described in Table 2. We have done our analysis on the KDD dataset described here.

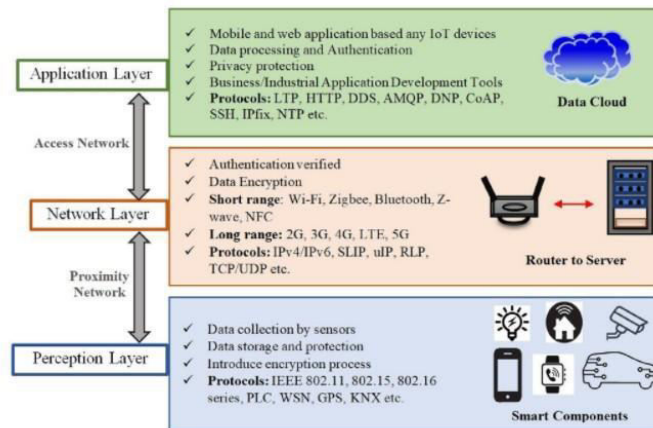


Fig I

IV. METHODOLOGY

The goal of our paper is to compare the performance of different supervised learning techniques in intrusion detection. Our goal is to analyze the sensitivity and performance of supervised techniques such as Decision Tree, Naive Bayes, Logistic Regression, SVR and by using different distributions of training datasets. We are also trying to investigate and tabulate different accuracy of the techniques used to detect the different types of attacks in the dataset.

Description of the KDD99 data set

There are mainly four types of attacks that were used in our simulations.:

- Probe: Attacks that are deliberately crafted and automatically scan a network of computers and hardware to gather information or find known vulnerabilities i.e., the probable weak point in the computer system.
- Denial of Service (DoS): These categories of attacks are meant to shut down a machine or network, making it inaccessible to its specified users. This attack denies legitimate requests from legal users of the system by excessive consumption of resources such as flooding the target with traffic.
- User to Root (U2R): Beginning attacker tries to access the normal user account then gain access to the root by exploiting the weakness of the systems these attacks would be a threat to superuser privileges from a normal user privilege.
- Remote to Local (R2L): The attacks which result in a local user account launching a remote exploit.

These four attack types were ordered sequentially. In each turn, a different attack was randomly picked up from the next category. Since the number of the Probe, R2L, and U2R attacks are very few compared to the DoS attack.

Normalization

Normalization is the important step inside the preparation of the training datasets because the features in KDD are of various natures and their scales are of excessive variance.

Standardization of the training dataset is a basic requirement for many machine learning and deep learning algorithms; they might behave badly if there exists a class imbalance. minority sampling can be performed to avoid higher class imbalance.

The standard normal distribution is a form of the normal distribution. When a normal random variable has a mean equal to 0 and a standard deviation equal to 1.

The random variable in the standard normal distribution is known as the standard score(zscore).It is possible to transform every normal random variable X into a z score using the following formula:

$$z = (X - \mu) / \sigma$$

Here, X is a normal random variable, μ is the mean of X, and σ is the standard deviation of X.

In this study, we have only considered four major supervised machine learning techniques: Decision Tree, Naive Bayes, Logistic Regression, SVR. Following is a brief explanation of each technique.

Decision Trees are powerful and popular techniques for classification and prediction problems. Classification is done in a hierarchical order in the form of a tree. It follows a topdown approach. Each internal node acts on a particular attribute and the leaf node represents the value of the target attribute.

Building the decision tree: Decision trees are generally built based on a set of training data. In this case, a particular attribute will be chosen for each node. Also, the leaf nodes would be labeled according to the appropriate class.

Classification: To classify a new event, an attribute for each node (top to bottom) is considered. Based on the value of the attribute, the tree branch is selected. This process is repeated until the algorithm reaches the leaf node.

Naive Bayes Classifier Algorithm is a simple and strong probabilistic classifier algorithm based on the Bayes rule. The main aim of the Naive Bayes classification is to frame a rule which will allow assigning future objects to a class when only provided with the vectors of variables that describe the future objects.

Logistic Regression is a regression algorithm that can be used for classification and segmentation. Logistic Regression is a statistical method that allows for the analysis and prediction of many events, especially

dichotomous events. With respect to the number of values in the dependent variable, a logistic regression model is divided into binomial regression analysis and multinomial regression analysis. Support Vector Machine (SVM) is a robust and flexible machine learning model which possesses remarkable robust performance with respect to sparse and noisy data. It is mainly used in classification problems. SVM can perform both regression and classification tasks and can also handle continuous and categorical data. SVM creates an optimal hyperplane, a hyperplane between two classes. SVM follows an iterative training algorithm. The goal of the hyperplane is to maximize the distance from each of the available classes and hence to distinguish each class with a minimum error at the maximum margin

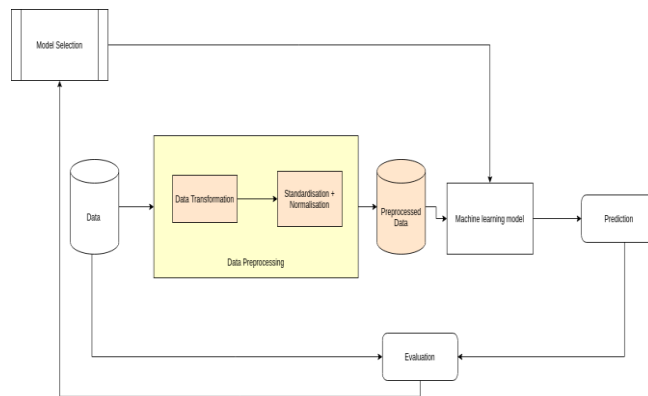


Fig 2

V. DATA PREPARATION

In 42 features present the dataset, protocol type, service, and status flags are categorical features and the remaining 38 features are continuous in nature. These categorical variables are converted to continuous variables by using cat.codes function. The value counts of different types of attacks are presented in Table II.

Attack Types	Value Counts
dos	391458
normal	97278
probe	4107
r2l	1126
u2r	52

Table II

On treating the duplicate and null values in the KDD99 data set, we performed a series of feature extraction and standardization procedures. Some of the features were coupled together and replaced with specific variables for easy computation. Certain features had to be converted from categorical to numerical.



Well-formatted and validated data improves the data quality and protects applications from encryption such as empty values, unexpected duplicates, incorrect identification, and inconsistent formats.

Basic Exploratory data analysis was performed to visualize different patterns and important features.

In order to analyze the sensitivity of surveyed strategies in the distribution of training data, data sets with a different relative number of attack records were prepared based on defined demographic categories. Data is then selected using K-fold from the training set. The selection is done in such a way that each training set contains all the attack categories.

VI. EXPERIMENTAL RESULT

The outcome of our model is tested using K-fold verification and this is done by the use of scikit-learn. This is a re-sample method that will provide a measure of model performance. It does this by dividing the data into k segments, training the model into all components except one held as a test set to test model performance. This process is repeated k-times and the score scale for all built-in models is used as a solid performance measure. Techniques were used in each training set. The same test data was used in different training sets for each phase. The results of each of the 5 different set strategies (in each phase) were combined to evaluate the method.

We have summarized the result for different methods in table III.

Models	Accuracy	Precision
Decision Tree	0.9202	0.9867
Naïve Bayes	0.6171	0.9664
Logistic Regression	0.9609	0.9578
SVR	0.9746	0.9691

Table III

The results from Decision Tree and SVR show an accuracy rate of more than 90% for the detection of attacks. Naive Bayes has a detection accuracy rate of 61%.

Classification reports for different attacks were also summarised in the table.

We have compared our work with the latest work done in the same and the reports generated by other papers are summarised in Table IV.

CLASSIFICATION REPORT

	Date of publishing	Detection methodology	Treated threats	Algorithms used	Accuracy
Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 Datasets using Deep Learning in IoT.	2019	Deep learning neural networks	DoS, R2L, U2R, Probe.	DNN, ANN	97.5%
A Detailed Analysis of the KDD CUP 99 Data Set.	2009	Feature extraction and building new data set (NSL-KDD)	-	-	-
Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives.	2018	-	DoS, R2L, U2R, Probe, Normal	Naive Bayes, SVM, Decision Tree, RF, ANN, K-Means	Naive Bayes-86 % SVM-94% Decision Tree-94% RF-94% ANN-94% K-Means-93%
Intrusion Detection System using KDD Cup 99 Dataset.	2020	J48, Naive bayes, RF	DoS, R2L, U2R, Probe, Normal	J48, Naive bayes, RF	Naive-92.90% RF- 99.99% J48-99.98%
Integrating complex event processing and machine learning: An intelligent architecture for detecting IoT security attacks	2020	Complex Event Processing (CEP) technology and the Machine Learning(ML)	TCP, UDP and Xmas port scans, and DoS	-	-
Building a Intrusion Detection System for IoT Environment using Machine Learning Techniques	2020	Machine Learning	-	Naive Bayes, SVM, decision tree, Adaboost	Naive Bayes -97% SVM-98% Decision tree-100% Adaboost- 98%

Table IV

VII. CONCLUSION

In the paper, datasets with a variety of attack types and percentages are used to evaluate supervised intrusion detection techniques. In this paper, training data sets have a different number of attacks and percentages used to evaluate intervention strategies for finding intervention. The simulation results show that the maximum acquisition rates of the three categories are the same at each stage of the attack. In analyzing the high detection rate, Decision Trees and SVR show good results in the detection of DoS. For detecting other stages of the attack, Naive Bayes gives better results as compared to the other models being experimented with here in the paper. We also considered a general deviation from the level of acquisition of different strategies in each class of people. Strategic effectiveness is assessed based on databases with different percentages of attacks. Based on the results obtained from this paper, the strategies are likely to show stronger than predictability when training using different training data sets. It was also noted that the strategies are likely to reflect different detection rates in the data that have fewer samples such as R2L, U2R, and Probe.

As part of our future work, we aim to find the right combination of these accessibility features. We will work on forming a model with proper integration with CEP and ml models to try to build a better model to detect anomalies in the IoT Network.

REFERENCES

- [1]. Gad, Rüdiger, Juan Boubeta-Puig, Martin Kappes, Inmaculada Medina-Bulo, "Hierarchical Events for Efficient Distributed Network," 2012.
- [2]. Farnaz Gharibian and Ali A. Ghorbani, "Comparative Study of Supervised Machine Learning Techniques for Intrusion Detection".
- [3]. Syeda Manjia Tahsien, Hadis Karimipour, Petros Spachos, "Machine learning-based solutions for the security of Internet of Things (IoT): A survey," in Elsevier Ltd, 2020.
- [4]. K. V. V. N. L Sai Kirana, R. N. Kamakshi Devisetty, Pavan Kalyana, Mukundini, "Building an Intrusion Detection System for IoT Environment using Machine Learning Techniques," in Elsevier B.V, 2020.
- [5]. Z. T. Muhammad Shafiq, "Selection of effective machine learning algorithm and Bot-IoT attacks traffic identification for the internet of things in smart city," in 0 Elsevier, 2020.
- [6]. Mohamed Faisal Elrawy, Ali Ismail Awad, Hesham F. A. Hamed, "Intrusion detection systems for IoT-based survey," 2020.
- [7]. Nadia Chaabouni, Mohamed Mosbah, Akka Zemhari, Cyrille Sauvignac, and Parvez Faruki, "Network Intrusion Detection for IoT Security based on Learning Techniques," 2018.
- [8]. Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," in Elsevier, 2013.
- [9]. Matthew Jagielski*, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li, "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," 2018.
- [10]. Mandal, K., Rajkumar, M., Ezhumalai, P., Jayakumar, D., & Yuvarani, R. (2020). Improved security using machine learning for IoT intrusion detection systems. *Materials Today: Proceedings*. <https://doi.org/10.1016/j.matpr.2020.10.187>
- [11]. Choudhary, S. and Kesswani, N., 2021. Analysis of KDD-Cup'99, NSL-KDD, and UNSW-NB15 Datasets using Deep Learning in IoT.
- [12]. Ieeexplore.ieee.org. 2021. Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives. [online] Available at: <<https://ieeexplore.ieee.org/document/8586840>> [Accessed 24December 2018].
- [13]. M. Tavallae, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.
- [14]. "Intrusion Detection System using KDD Cup 99 Dataset," International Journal of Innovative Technology and Exploring Engineering Regular Issue, vol. 9, no. 4, pp. 3169–3171, 2020.
- [15]. Roldán, José, et al. "Integrating Complex Event Processing and Machine Learning: An Intelligent Architecture for Detecting IoT Security Attacks." *Expert Systems with Applications*, vol. 149, 2020, p. 113251., <https://doi.org/10.1016/j.eswa.2020.113251>.



INNO  **SPACE**
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542



ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 **9940 572 462**  **6381 907 438**  **ijircce@gmail.com**



www.ijircce.com

Scan to save the contact details