Supervised Machine Learning Based Network Intrusion Detection System for Internet of Things

1st Deepa Rani
Department of Computer Science and Engineering
National Institute of Technology
Hamirpur, India
dr.deeparani1@gmail.com

2nd Dr. Narottam Chand Kaushal
Department of Computer Science and Engineering
National Institute of Technology
Hamirpur, India
nar@nith.ac.in

***Abstract*—The Internet of Things (IoT) is an innovative invention that can combine physical object to the Internet with an ability to transfer and access of the data through Internet, however with the rapid growth in the application and services of the IoT, the scope of network attack is also increasing exponentially. To secure data, device and IoT network, there is a need of an efficient, secure and accurate Intrusion Detection System (IDS). IDS basically monitors network and system activities and raises alarm when anything deviated from its normal behaviour is found. Classical intrusion detection system follows rule based detection approaches that fail to detect zero day or unknown attack is not suitable for dynamic and insecure IoT environment. This paper mainly proposes an efficient method with uniform detection system based on supervised machine learning technique by using Random Forest classifier. Also two different datasets, NSL-KDD and KDDCUP99 with minimal feature sets have been used that give lightweight attack detection strategy for IoT network. Simulation of proposed method with theses datasets has 99.9 percentage accuracy in intrusion detection with less amount of time and energy.**

***Index Terms*—Internet of Things, IDS, Machine Learning, Random Forest, Network Security**

I. INTRODUCTION

Internet of Things was previously known as ”Internet of Everything”. Using Internet of Things (IoT), physical object can sense, store, process and transfer data with the help of Internet without any human and machine interference. Its applications and services increase rapidly even in our common household appliances. The main purpose of IoT is to collect, process and then transfer data to the application[1]. Sensors and devices (small and memory constraint) are responsible for the collection and processing of data is done by the application process and for data transfer BLE (Bluetooth Low Energy), Wi-Fi, Bluetooth, Zigbee etc are used. Deployment of devices is important with respect to the security and privacy to ensure safety of data and secure network.

According to Gartner report August 2019, use of IoT is expected to increase to 5.8 million at the end of 2020 and 20.4 million to 2022[2]. Because of wireless medium, more prone to network attack, is being used for data transmission, among many challenges of IoT, ensuring security and privacy is critical concern. With the increased demand of IoT and connected objects and devices which are heterogeneous in nature, IoT network has became soft target for attackers. In addition, many nodes in IoT use to store and share private information, they are also became obvious target for attackers. Therefore it requires lightweight algorithm to perform task and it should be heterogeneous in nature. IoT system consist of four layered [3] architecture, shown in Figure 1, and Intrusion can be deployed in any single layer or in multiple layers of IoT but most vulnerable is network layer [4].

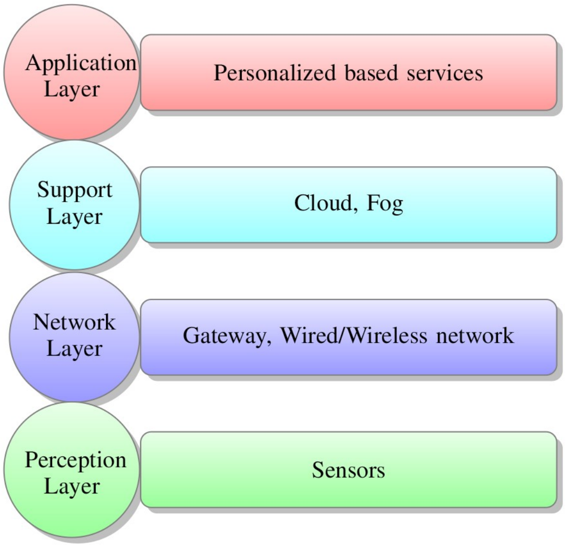

Fig. 1: IoT Architecture

Intrusion Detection system (IDS) plays a vital role in IoT environment for intrusion detection by raising an alarm when anything deviates from its normal behaviour in a system. Basically it deal with four approaches namely,[5] Anomaly Based Detection System, Signature Based Detection System, Specification Based Detection System and Hybrid Based Detection System. Anomaly based IDS uses normal profile of a system and generate alert when anything deviate from its normal profile. It is best suited for unknown attack but generates lots of false alarm. On the other hand, Signature based IDS uses pattern, signature, behaviour of a known attack and generates an alarm if any matches is found. It is mainly suitable for known attacks only and fails to recognize unknown one. Specification based IDS is used to detect only specific type of attack and [6] is hard to develop and verify. However, Hybrid based IDS is the combination of all three.

Traditional IDS fails to provide privacy and security to the IoT system as new attacks are launched day by day and not having any fixed pattern. Therefore an architecture with the combination of IoT and Artificial Intelligence (AI) can lead to self- healing and self corrective in nature. In this combination AI can act as a brain that helps in taking decision and IoT can act as a nervous system which perform the defined action against the attack. Due to the small storage capacity and computational limitation of IoT, Cloud based IoT became more popular that can sort out the issue related to connectivity, storage capacity and data computation. Combining Cloud based IoT with AI system helps to take action upon certain data i.e make a system to take appropriate decision on some condition.

In AI, Machine Learning (ML) helps in reducing analysis time of any node. For Example, camera (IoT without ML) acting as a sensor, will forward every image for analysis but camera having embedded ML will send only those images that contain specific object for analysis. It reduces the analysis time by providing only matched picture frames. So, Machine Learning algorithm give better result in intrusion detection in least amount of time. In 21st century, Machine Learning becomes most powerful tool for tuning information into knowledge. It can take large volume of multi-dimensional and multi-variety of data and can easily discover specific trends and pattern. It is the best method for precisely detecting network intrusions.

In this paper, network intrusion detection system is proposed based on supervised machine learning method that uses random forest classifier with the minimal feature set. And these features helps to train the model and predict intrusions for IoT environment accurately. This method is very much suitable for both KDDCUP99 and NSL KDD dataset. Manually selected feature set results in selection of minimum and relevant features which helps in quick learning and fast and accurate detection.

The remaining paper is organized as follows. Section II discusses research related works. Section III introduces proposed algorithm that contains information about dataset, pre-processing of data, optimal selection of features, balanced splitting and random forest classifier. Section IV consist of performance evaluation, in which comparison with different machine learning classifier and with other existing approaches are shown and Section V contains Conclusion and scope of Future Work.

II. Related Work

There are many intrusion detection methods proposed and developed to provide security and privacy requirements in IoT but still it is critical and severe for IoT environment as new attacks are quickly generated. Most of the research is based on network and host based system and uses anomaly and signature based detection approaches. Recently, Zhao et al.[7] introduced a two stage IDS for SD-IoT (Software Defined IoT). In first stage it performs feature selection using Bat Algorithm and in the second stage it performs flow classification using Random Forest. It gives the system self-learning capacity. Zhang et al.[4] proposed Deep Belief using genetic algorithm that optimizes the number of neuron and hidden layer. Prabavathy et al.[8] introduces online sequential extreme learning machine for detecting novel attack that can be used for real time environment where data is generating at high velocity. It shows IDS placed at fog layer provides better security than cloud layer. In [9], authors uses consine similarity of the vectors for checking message in rate of packets in the switch port of SD-IoT environment. It uses a threshold value at the switch to detect DDoS attack effectively. Sana and her team [10] proposed a supervised ML based support vector machine(SVM) to detect unnecessary data injection in the IoT network. This paper considered only one feature i.e rate of transmitting packet, that helps in maintaining a lightweight system. Zarpelao et al.[1] present a survey paper of attack detection for IoT and summarise placement strategies, detection approach, detected threats by different papers. Security threats and requirements, issues, challenges and countermeasures are listed in [11–13].

As we can see that there is lot of research conducted using Artificial Intelligence tools. As Machine Learning (ML), Deep Learning (DL) and Artificial Neural Network (ANN) are the sub parts of Artificial Intelligence (AI). Apart from AI, many other research have been performed in the field of intrusion detection that is not only limited to the IoT network. Many more are designed for Wireless Sensor Network (WSN). Some are limited to some specific types of attacks only. Most of the cyber-attack like zero day attack has gained serious attentions for research. Raza et al.[14] gives a method for real time routing attack (zero day) like forwarding attack, spoofed information attack.Rathore et al.[15] work on NSL-KDD data set using semi-supervised integration of Fuzzy C-Means and Extreme Learning Machine that helps in quick learning and perform better detection rate. Attack detection is carried out at the fog layer that satisfy resources problem of IoT devices. In Fu et al.[16] introduced an finite automata based detection system in which Fu describes three attack scenarios keeping the heterogeneous nature of IoT network. It maintains an Event database to analyze the event for detecting intrusions. Sedjelmaci et al.[17] uses Game Theory for intrusion detection. It combines anomaly and signature based approaches in very intelligent way.Signature based detection is carried out for all packet but it activates anomaly detection only when any new attack/signature is predicted to occur. This reduces the computational overhead and improves detection time. In [18], Abhishek et al. proposes a ensembler method by including four classifier to classify intrusions. But it takes high computational time.

It is easy to provide security in wired channel but it is difficult to fully secure wireless channel as it is more vulnerable to attack and placement of IDS is also a difficult task due to infrastructural difference[19]. Attackers easily exploit vulnerable spot due to open channel and deploy different intrusions. Heterogeneous type of data can smartly

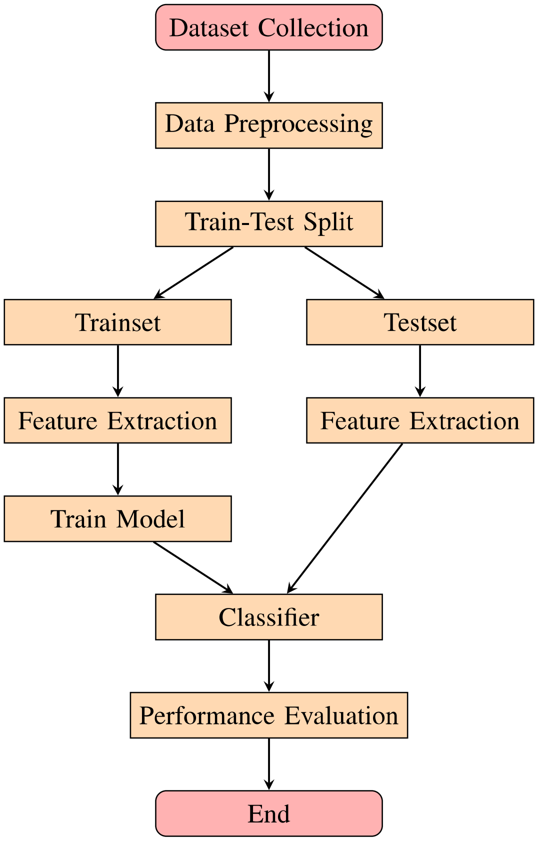

Fig. 2: Proposed Workflow

bypass classical IDS.

Undoubtedly, previous referenced paper shows concerned over security in IoT and provides value-able issues, challenges, methods. but does not show concern over countermeasure or mitigation techniques and does not provide any common approaches for any type of dataset.

III. Proposed System Implementation

This section includes workflow of proposed method, dataset collection, data preprocessing, balanced splitting of dataset and then implementation of Random Forest classifier. The workflow of proposed implementation is presented in Figure 2.

A. Data Collection:

Most widely used dataset is KDDCUP99[20] and [21]NSL KDD[7, 22, 23]. NSL-KDD dataset is extracted from KDDCUP99 dataset. Proposed method works on both the dataset effectively. KDD CUP 99 is extension of DARPA dataset. Data is very much redundant and duplicated. But NSL KDD dataset eliminates the redundant duplicated records from KDD CUP 99 dataset. New type of attacks are also incorporated to enhance the richness in the dataset.

B. Feature Description:

There are total 41 common features in both the dataset and 1 label class that contain categories of attack and normal class. It is of numeric and object type. KDD cup contains 494020 X 41 data and in NSL KDD contains 125973 X 42 (Train Set) and 22544 X 41 (Test Set). On the basis of features, it is classified as three sub categories- (1) Ten features are related to basic TCP connection. (2) Twelve features related to content features (3) Rest contains features within a two-second time window.

TABLE I: Attack Classification of KDD CUP 99 Dataset

DoS	**User to Root**	**Probe**	**Remote to Local**
back	buffer_overflow	Satan	multihop
teardrop	perl	ipsweep	guess_password
neptune	rootkit	portsweep	ftp_write
land	loadmodule	nmap	imap
smurf			phf
pod			warezmaster
			spy
			warezclient

C. Data Classification:

Label is broadly classified as five different categories. One labeled as normal class and four as attack classes namely Denial of Service, U2R, R2L and Probe. These attack classes are further classified into different attack types that are listed in Table I (KDDCUP 99) and in Table II (NSL KDD).

1) **Denial of Services (DoS):** It is a cyber-attack in which attacker make server very busy by sending superfluous request to server so that legitimate user cannot access server or denies request of legitimate user.
2) **Probe:** In this attacker attempts to exploit vulnerabilities in a network. So that they can easily enter and manipulate network activity.
3) **User to Root (U2R) :** In this lower layer attackers try to access root privileges in order to gain root access. Unauthorized user violates security restriction to access top authority privilege using escalation technique.
4) **Remote to Local (R2L):** It is very much similar to User to Root attack. In this, unauthorized access from server or remote area to the local user takes place. Firstly, attacker tries to exploit targeted system vulnerability and then tries to access user privilege.

D. Data Preprocessing Phase:

For data preprocessing pandas and numpy from sklearn based libraries tools of machine learning are used.This is most important phase of any machine learning project. It gives the clear understanding of the problem and makes an organized dataset. Since data is highly unbalanced and of two different types, therefore there is a need of preprocessing data so that data can be classified into train and test set evenly on the basis of label of attack and encoding of data in terms of numeric value can be done.

TABLE II: Attack Classification of NSL KDD Dataset

DoS	Probe	User to Root	Remote to Local
apache2	portsweep	buffer_overflow	ftp_write
back	nmap	loadmodule	guess_password
neptune	satan	perl	imap
pod	ipsweep	rootkit	multihop
smurf	saint	ps	phf
processtable	mscan	xterm	spy
udpstrom		sqlattack	warezclient
mailbomb		worm	warezmaster
land		snmpguess	spy
teardrop			sendmail
			snmpgetattack
			xlock
			httptunnel
			named
			xsnoop

1) **Data Encoding:** There are total 41 features in which three Features namely services, protocol_type and flag are of object type and machine learning works on numeric/vector form of data. That's why we need to encode these fields in terms of numeric value with the help of encoder. This process is done on both the dataset.

2) **Balanced Splitting:** For KDD CUP 99, records are split into 80% training and 20% testing set. Splitting of the data records is done on each label of attack rather than splitting whole data. This makes it balanced splitting. After that, all labels are merged into two train and test set. Distribution of Data of KDD CUP 99 and NSL KDD are shown in the Table III and Table IV respectively
For NSL KDD, initially it is divided into train and test set but splitting is not balanced. Therefore, firstly we merged both dataset and then splitting the data in accordance with their label fields in the ratio of 4:1. After that again merged into two sets namely Trainset and Testset.

E. Feature Extraction:

Feature extraction is the process of selecting important features among 41 features that are present in the dataset. These feature sets will help the classifier to learn the behaviour and pattern of records on every attack and normal packet type. Improper feature selection will lead to performance depletion of any classifier. With the help of feature classifier will predict the behaviour and pattern of new packets. So, in this paper, only 10 features are selected for training and testing. This will result in fast learning and predicting the attack with better accuracy.

TABLE III: Data Distribution of KDD CUP dataset

class	Training Set		Testing Set	
	Data	Percentage	Data	Percentage
Normal	77,822	19.6	19,455	19.6
DoS	3,13,167	79.2	78,291	79.2
Probe	3,287	19.6	820	19.6
R2L	897	0.2	229	0.2
U2R	37	0.09	15	0.09
Total	3,95,210		98,810	

TABLE IV: Data Distribution of NSL KDD dataset

class	Training Set		Testing Set	
	Data	Percentage	Data	Percentage
Normal	61643	51.8	15,410	51.8
DoS	42,710	35.9	10,675	35.9
Probe	11,264	9.4	2,813	9.4
R2L	2,842	2.3	707	2.3
U2R	362	0.03	90	0.03
Total	1,18,821		29,695	

F. Train Model:

For training the model, Random Forest classifier is used. Random forest is a collection of large number of Decision Tree classifier. It is a ensembler classification technique. Several decision trees are constructed on the basis of training set and using majority voting final class is predicted as shown in the Figure 3. Hence, It generate more stable and accurate prediction.Thus makes system's performance better in terms of accuracy, recall, precision and false alarm rate.

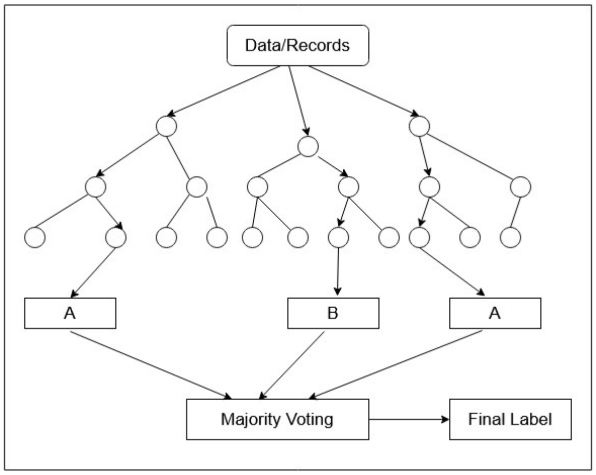

Fig. 3: Random Forest Classifier

IV. Perfomance Evaluation

In this section, numerical and statistical measurement is represented to strengthen evaluation procedure. This section covers the complete table of data distribution used in our evaluation, accuracy in terms of different classifier and finally proposes accuracy comparison with other available approaches.

The performance is measured in terms of accuracy, precision, recall and f1-score. Calculation of these terms are given in the matrix.

		Actual	
		Positive	**Negative**
Predicted	**Positive**	TP	FP
	Negative	FN	TN

1) **True Positive:** Predicted as malicious and packet is malicious.
2) **True Negative:** Predicted as not malicious and packet isn't actually malicious.
3) **False Positive:** Predicted as malicious but actually it is not a malicious packet.
4) **False Negative:** Predicted as not malicious but packet is actually a malicious.

Accuracy, Precision, Recall and F1 score are calculated by using True (True positive and negative) and False (positive and negative) rate variables. Measurement of performance by using different machine learning classifier namely proposed (Random Forest), KNN (K-Nearest Neighbour), Naive Bayes, Decision Tree and Logistic Regression is shown in the Table V. Result is extracted by using only ten features out of 41 features. As these features are applicable for both datasets. Proposed method is having highest accuracy i.e 99.9 %. Further proposed method is again compared with existing available method for intrusion detection in the field of Internet of Things in the Table VI. And the result shows that this method has highest accuracy with least false alarm rate. Figure 4 and Figure 5 show the comparison with different approaches based on accuracy on KDD CUP and NSL KDD dataset. Since this method gives 98.1% accuracy on NSL KDD dataset and the highest accuracy is 98.82% in [4]. But proposed method has less false alarm rate as compared to referenced paper. This approach reduces the computation time of system by selecting minimal feature set. It takes about 1.78 seconds to train the system and 0.28 second to predict the intrusion.

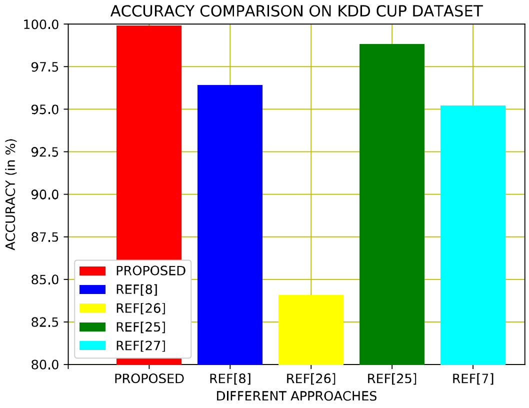

Fig. 4: Performance Measurement on KDDCUP dataset

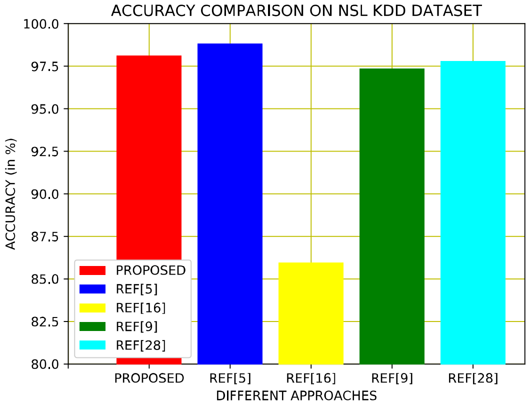

Fig. 5: Performance Measurement on NSL KDD dataset

V. Conclusions and Future Work

This paper is based on the supervised machine learning method that uses Random Classifier to train and predict intrusions according to trained set in the IoT environment. In this approach, feature is selected manually after analysing different attacks and their characteristics that depend upon the feature and extracted minimal features. After that, trains the system with most suitable classifier that gives better result in terms of accuracy and false alarm rate. It takes lesser time in learning and predicting. The effectiveness of this method can be verified by comparing with the existing approaches. But, it is implemented on synthesised generated network traffic dataset so, it may not suitable for real network traffic.

In the future work, this method will be implemented on real network traffic and measuring the performance. In future predicting intruder's next action can be done to protect IoT environment proactively. This method is only limited to detection approach, we can also add mitigation and prevention measure for enhance its effectiveness.

References

[1] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in internet of things," *Journal of Network and Computer Applications*, vol. 84, pp. 25–37, 2017.

[2] L. Goasduff, "Gartner says 5.8 billion enterprise and automotive iot endpoints will be in use in 2020."

[3] P. Sethi and S. R. Sarangi, "Internet of things: Architectures, protocols, and applications," *Journal of Electrical and Computer Engineering*, pp. 1–25, 2017.

TABLE V: Performance Measurement using different classifier

Classifier	**KDD CUP 99**				**NSL KDD**			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Proposed	**99.9**	**99.9**	**99.9**	**99.9**	**98.1**	**98.2**	**98.1**	**98.0**
KNN	99.8	99.8	99.8	99.8	96.8	96.7	96.8	96.5
Naive Bayes	91.9	95.8	91.9	93.0	36.3	52.0	36.3	21.9
Decision Tree	99.9	99.9	99.9	99.9	98.0	98.1	98.0	97.9
Logistic Regression	97.1	96.3	97.1	96.4	80.7	70.3	80.7	75.0

TABLE VI: Performance Comparison related to different paper

KDD CUP 99		**NSL KDD**	
References	**Accuracy**	**References**	**Accuracy**
Proposed	**99.9**	**Proposed**	**98.1**
Reference No.[7]	96.42	**Reference No.[4]**	**98.82**
Reference No.[24]	98.82	**Reference No.[15]**	85.95
Reference No.[25]	84.06	**Reference No.[8]**	97.36
Reference No.[26]	95.21	**Reference No.[27]**	97.80

[4] Y. Zhang, P. Li, and X. Wang, “Intrusion detection for iot based on improved genetic algorithm and deep belief network,” *IEEE Access*, vol. 7, pp. 31711–31722, 2019.

[5] S. Suganthi and D. Usha, “A survey of intrusion detection system in iot devices,” *International Journal of Advanced Research*, vol. 6, pp. 23–30, 2018.

[6] R. Berthier and W. H. Sanders, “Specification-based intrusion detection foradvanced metering infrastructures,”

[7] J. Li, Z. Zhao, R. Li, and H. Zhang, “Ai-based two-stage intrusion detection for software defined iot networks,” *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2093–2102, 2018.

[8] S. Prabavathy, K. Sundarakantham, and S. M. Shalinie, “Design of cognitive fog computing for intrusion detection in internet of things,” *Journal of Communications and Networks*, vol. 20, pp. 291–298, June 2018.

[9] D. Yin, L. Zhang, and K. Yang, “A ddos attack detection and mitigation with software-defined internet of things framework,” *IEEE Access*, vol. 6, pp. 24694–24705, 2018.

[10] S. U. Jan, S. Ahmed, V. Shakhov, and I. Koo, “Toward a lightweight intrusion detection system for the internet of things,” *IEEE Access*, vol. 7, pp. 42450–42471, 2019.

[11] H. Suo, J. Wan, C. Zou, and J. Liu, “Security in the internet of things: a review,” in *2012 international conference on computer science and electronics engineering*, vol. 3, pp. 648–651, IEEE, 2012.

[12] E. Leloglu, “A review of security concerns in internet of things,” *Journal of Computer and Communications*, vol. 5, no. 1, pp. 121–136, 2016.

[13] S. Madakam, R. Ramaswamy, and S. Tripathi, “Internet of things (iot): A literature review,” *Journal of Computer and Communications*, vol. 3, no. 05, p. 164, 2015.

[14] S. Raza, L. Wallgren, and T. Voigt, “Svelte: Real-time intrusion detection in the internet of things,” *Ad hoc networks*, vol. 11, no. 8, pp. 2661–2674, 2013.

[15] S. Rathore and J. H. Park, “Semi-supervised learning based distributed attack detection framework for iot,” *Applied Soft Computing*, vol. 72, pp. 79–89, 2018.

[16] Y. Fu, Z. Yan, J. Cao, O. Koné, and X. Cao, “An automata based intrusion detection method for internet of things,” *Mobile Information Systems*, vol. 2017, 2017.

[17] H. Sedjelmaci, S. M. Senouci, and M. Al-Bahri, “A lightweight anomaly detection technique for low-resource iot devices: A game-theoretic methodology,” in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2016.

[18] A. Verma and V. Ranga, “Elnids: Ensemble learning based network intrusion detection system for rpl based internet of things,” in *2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, pp. 1–6, April 2019.

[19] A. Mishra, K. Nadkarni, and A. Patcha, “Intrusion detection in wireless ad hoc networks,” *IEEE wireless communications*, vol. 11, no. 1, pp. 48–60, 2004.

[20] W. L. A. P. Salvatore J. Stolfo, Wei Fan and P. K. Chan, “Intrusion detection learning.”

[21] K. Mahesh, “Dataset for intrusion detection system.”

[22] A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” *IEEE Communications Surveys Tutorials*, vol. 18, pp. 1153–1176, Secondquarter 2016.

[23] D. D. Protić, “Review of kdd cup’99, nsl-kdd and kyoto 2006+ datasets,” *Vojnotehnički glasnik*, vol. 66, no. 3, pp. 580–596, 2018.

[24] J. Li, Z. Zhao, and R. Li, “Machine learning-based ids for software-defined 5g network,” *IET Networks*, vol. 7, no. 2, pp. 53–60, 2018.

[25] W. Alhakami, A. ALharbi, S. Bourouis, R. Alroobaea, and N. Bouguila, “Network anomaly intrusion detection using a nonparametric bayesian approach and feature selection,” *IEEE Access*, vol. 7, pp. 52181–52190, 2019.

[26] J. Li, Z. Zhao, and R. Li, “A machine learning based

intrusion detection system for software defined 5g network," *arXiv preprint arXiv:1708.04571*, 2017.

[27] A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for internet of things," *Future Generation Computer Systems*, vol. 82, pp. 761–768, 2018.