TOWARDS NEAR-REAL-TIME-INTRUSION DETECTION FOR IoT DEVICES USING SUPERVISED LEARNING AND APACHE SPARK

[1]VODNALA SREEJA, [2]Dr.B. SATHEESH KUMAR

[1] M. Tech Scholar, [2]Professor, Department of CSE,

JNTUH UNIVERSITY COLLEGE OF ENGINEERING, JAGTIAL, T.S., INDIA..

ABSTRACT

Attack and anomaly detection in the infrastructures for the Internet of Things (IoT) are growing problems. Threats and attacks against IoT infrastructure are rising proportionally as its use expands across all industries. This study compares the application performances and training/application timelines of multiple machine learning methods for detecting cyberattacks (namely SYN-DOS attacks) on Internet of Things (IoT) systems. As a quick and versatile engine for big data processing, Apache Spark's MLlib module contains supervised machine learning techniques. We use a training set of up to 2 million instances to demonstrate the implementation details and the performance of those algorithms on open-source datasets. We choose a cloud environment, highlighting the value of scalability and use flexibility. The outcomes demonstrate that all of the Spark algorithms employed produce very strong identification accuracy (>99%). Random Forest, in particular, obtains an overall accuracy of 1. We also describe a remarkably brief training period (23.22 sec for Decision Tree with 2 million rows). Additionally, utilising Apache Spark in the Cloud, the trials demonstrate a very fast application time (0.13 seconds for more than 600,000 instances of Random Forest). Furthermore, using high- or low-level programming languages, it is quite simple to implement the explicit model produced by Random Forest. The application of an explicit Random Forest model, implemented directly on the IoT device, along with a second level analysis (training), performed in the Cloud, is suggested as a hybrid approach for the detection of SYN-DOS cyber-attacks on IoT devices in light of the results obtained, both in terms of computation times and identification performance.

Keywords: cyber-attacks, Intrusion, Apache spark, IoT.

I. Introduction

Network intrusions are attacks that happen when inbound network packets carry out damaging actions, like Denial of Service (DoS) attacks or even make attempts to break into machines. A DoS attack is a technique for preventing the intended users from accessing a computer's resources. There are many other types of attacks, such as Flood, Land, and Ping Of Death (POD) attacks. Unexpected results while processing various user requests, slow system performance, unexpected system crashes, changes to kernel data structures, and unusually slow network performance are all signs of intrusions (opening files or opening websites). Computer systems have been secured utilising methods for preventing invasions such as authenticating people (e.g., using biometrics or passwords), preventing programming errors, and information security (e.g., encryption). As systems get more complex, vulnerabilities that can be taken advantage of owing to poor design, programming errors, or other "socially designed" penetration techniques will always exist, making security measures alone insufficient.

The use of Internet of Things (IoT) devices has significantly increased in recent years. By 2020, Gartner predicts that there will be 26 billion IoT devices worldwide [1, 2], and a Statista study predicts that there will be 75.44 billion [3].

The use of these devices is expanding steadily across a variety of applications, including mobile health, the Internet of Vehicles, smart homes, industrial control, and environmental monitoring, broadening the definition of mobile communications from interpersonal to clever interconnection between things and people as well as between things themselves [4]. Think of the smartwatch, IP camera, and smart TV as examples of how these gadgets are becoming a greater and more common part of billions of people's daily lives. These gadgets can unlock doors, monitor homes, and even record heartbeats by using sensors and actuators to interact with people. But these gadgets nearly always have an Internet connection. They are hence vulnerable to cyberattacks.

This means that while IoT devices increase human productivity and improve people's lives on the whole, they also expand the potential attack surfaces for cybercriminals [1]. According to a Hewlett Packard survey, 70% of the most popular IoT devices have significant security flaws [5]. IoT devices are vulnerable because there is no transport encryption, there are unsafe Web interfaces, there is insufficient software protection, and there is not enough authorisation. Each device has, on average, 25 vulnerabilities that pose a threat to the home network [1]. We introduce the datasets that will be utilised, the Apache Spark framework, and the machine learning-based Spark library MLLIB for IoT. We go over the datasets used, the chosen cloud environment, the measured parameters, and the experimental findings.

II. LITERATURE SURVEY

The use of deep machine learning and deep learning techniques in intrusion detection systems is a growing area of research and development. The job will be described in the following manner:

In order to increase the level of precision the machine learning system could provide, the study [5] developed the hybrid machine learning technology (decision tree and support Vector machine algorithms). The Decision Tree technique can be used to classify different attacks of different types. The support vector machine (SVM) algorithm is employed to classify typical data. The NSL-KDD Dataset was used to build the model. 96.4 percent of the time, the method is accurate. Researchers in [6] employed the support vector machines (SVM) and a method called as the gene algorithm (GA) to detect intrusion packets. SVM and GA are used in conjunction with specific features. Researchers utilise SVM to solve classification and regression problems. KDD Cup 1999 was used as the source of the study's data. The detection's precision was 97.3%.

Using the Recurrent Neural Networks method and the NSL-KDD data set, researchers from [7] developed an algorithm to identify networks. The research paper's findings are segmented into binary classifications, which have an accuracy rate of 83.28 percent, and multiclass classifications, which have an accuracy rate of 81.29 percent. The convolutional neural network is used by the method presented in [8] to detect network intrusions. The datasets were created using the KDD Cup 1999 dataset, and the model's accuracy was evaluated using test data that had been two-dimensionalized. The method has a detection rate of 97.7%. Using the KDD Cup 1999 dataset, researchers in [9] applied Artificial Neural Network for network intrusion detection. The system used min/max algorithms for normalised data as well as Principal Component Analysis (PCA) to reduce the number of attributes during preprocessing.

The article explores the architecture of artificial neural networks. In this study, the loss functions are mean squared error (MSE), Levenberg-Marquardt (LM) backpropagation, and feed forward neural network (FFNN). The accuracy of the system was 97.97%. The deep neural network used in the proposed system was built using NSL-KDD data. They recommended processing data using label-encoders and min-max normalisation, and building the deep learning layers using auto-encoders. Five separate categories were used to build this model. The dos assault has the best detection rate out of the five categories. attaining 89.8% and 97.7% for the probe and.

III. PROPOSED METHODOLOGY

The suggested intrusion detection system uses deep neural networks, an algorithm that applies anomaly detection techniques without accessing any data from its payload, to identify attacks. This system is created employing datasets through a sequence of operations to guarantee that there is no privacy violation. The most crucial component in the development of machine learning algorithms to spot suspicious threats and train them to recognise them is datasets. The results of this study, however, reveal that many researchers continue to use the KDDCup99 and NSL-KDD datasets, which have been heavily criticised for being inadequate and out of date for the current network infrastructure. This dataset is almost two decades old, having been created in 1999. The network infrastructure landscape is evolving as a result of the rapid expansion of information technology, including the cloud, social media, and the Internet of Things. These changes are a result of the threats and attacks themselves evolving in nature.

As a result, many research findings that demonstrate great accuracy are regarded as being overblown because the data used is not indicative of the current threat or the infrastructure.

This year's Third International Knowledge Discovery and Data Mining Tools Competition used the well-known KDDCup99 dataset. The 41 attributes used to describe each connection are (38 continuous or discrete numerical attributes, as well as three symbolic attributes). Each is categorised as either normal or a specific type of attack. These assaults can be divided into one of four groups: Probe DoS, U2R, and R2L. The following is a list of these groups. Before beginning the main attack, a probe attack gathers information about the system being attacked. DoS (pronunciation: "dos"): Disruption of Service (DoS) The type of attack may prevent legitimate requests from accessing network resources by consuming all available bandwidth or overloading computing power.

People who Root (U2R) in this specific situation In this case, the attacker has access to an account that the user frequently uses to log in. The attacker is able to gain access to the system's source by taking advantage of security flaws.

Local to remote (R2L) The attacker in this instance lacks access to any accounts on distant systems. Instead, they impersonate a user of the machine to exploit flaws for local access by sending an email message to the distant computer through a network.

The KDDCup99 dataset was upgraded into the NSL-KDD dataset, which was created in 2009. NSLKDD is an effort to enhance the KDDCup99 dataset by deleting duplicate records, incorrect instance counts, and other attack classifications [12]. However, it has replaced the dataset's primary limitation.

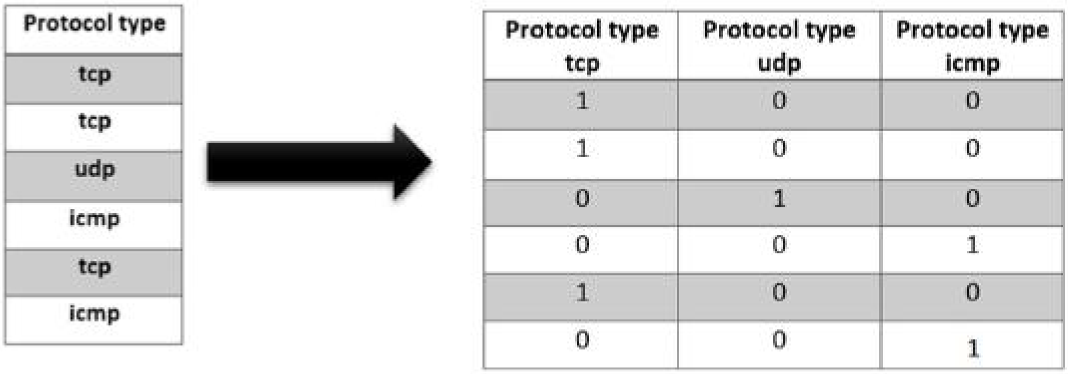
KDDCup99 has drawbacks. The first is that the data was created in 1999 using the user-friendly Solaris operating system, which was used to capture a variety of data. But the operating systems that aren't even close to Solaris have some significant distinctions from one another. Solaris has nearly little market share in the present era of Ubuntu, Windows, and MAC.

TCPdump, the traffic collector used for the KDD datasets, is very likely to become overloaded and stop transmitting packets as a result of a high volume of traffic. The assault distributions in the datasets are also unclear in certain places. According to an examination of attacks, Probe isn't considered an attack unless the quantity of repeats surpasses a set threshold and label discrepancy is noticed.

The emergence of new technologies like cloud computing is the third factor. The Internet of Things, cloud computing, and social media have all significantly altered the network architecture. New risks could emerge as a result of these changes.

.

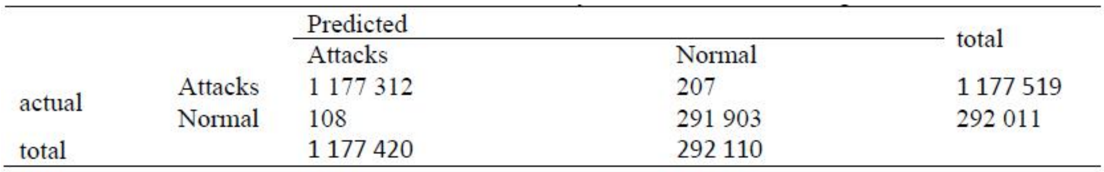
IV. DISCUSSION:



Protocol type
tcp
tcp
udp
icmp
tcp
icmp

Protocol type tcp	Protocol type udp	Protocol type icmp
1	0	0
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1

Fig 1: One hot encoder for the model



		Predicted		total
		Attacks	Normal	
actual	Attacks	1 177 312	207	1 177 519
	Normal	108	291 903	292 011
total		1 177 420	292 110	

Fig 2: Active and passive cases

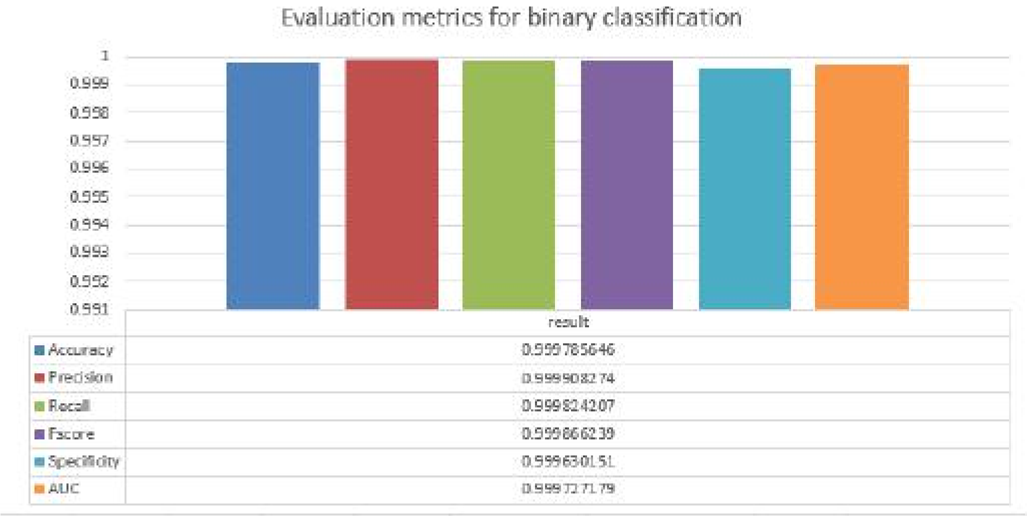


Fig 3: Evaluation metrics for the results

		Predicted					total
		DOS	Probe	R2L	U2R	normal	
	DOS	1 165 359	1	0	0	13	1 165 373
	Probe	5	12293	1	0	100	12 399
actual	R2L	1	0	269	0	79	349
	U2R	0	0	0	0	9	9
	normal	42	7	50	0	291 391	291 490
total		1 165 407	12 301	320	0	291 592	

Fig 4: Classification on multi class with confusion matrix

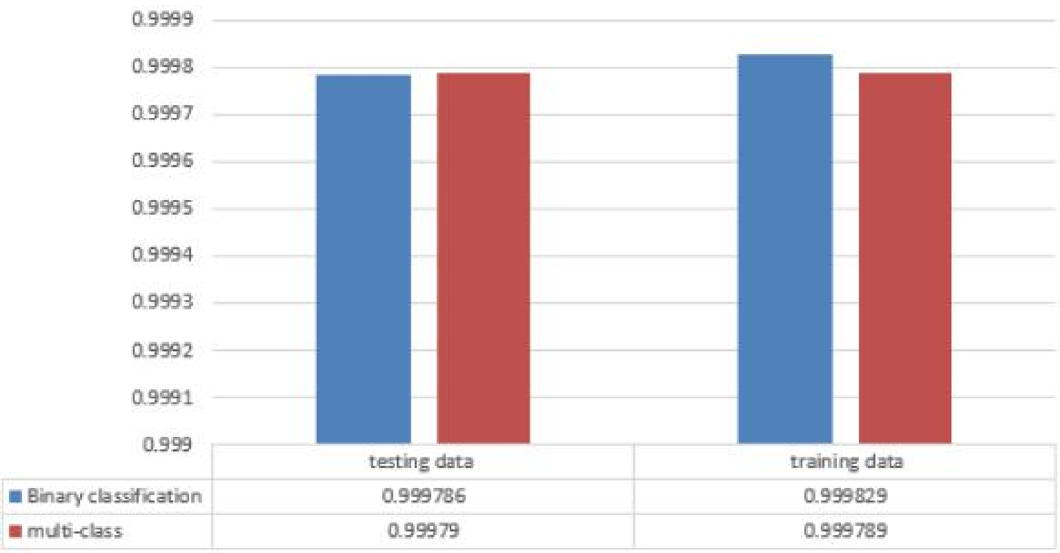


Fig 5: Accuracy of multi-class and binary classification

Researchers in IDS are paying close attention to soft computing techniques. This is because this approach is straightforward to use and frequently produces superior outcomes to a single program. The best solution is a carefully balanced combination of numerous algorithms. The classification of IDS, which can be helpful in defining the type of intrusions, is the main area of research attention. However, it could make it difficult to spot suspicious incursions that contain fresh or updated intrusion assaults. It is advised that a clustering method be used in the near future to further development in order to generate an improved IDS. Despite being more than 20 years old, the KDDCup99 dataset and its variation, the NSL-KDD dataset, are two of the most widely used datasets. Because intrusion attacks are always evolving with the newest technology and user habits, the frequent change in the data could cause IDS growth to stagnate. IDS will ultimately cease to exist as a general cyber security tool as a result of this.

V. CONCLUSION

The two types of models (multi-class and binary classification) are proposed in this study. Then, instead of using machine learning signatures or rules, we suggested using deep learning techniques in these models to detect network attacks. In this study, we've shown multi-class classification, which was found by looking at KDD cup 99 and other data sets. We have demonstrated that supervised, DNN models are capable of decoding and classifying assaults with high accuracy (99.98%). This was done by looking at network packets and connection parameters that don't contain payload packet information. Additionally, 99.99 percent of dos attacks were detected with great precision.

VI. REFERENCES

[1] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, “Network anomaly detection: methods, systems and tools,” Ieee communications surveys & tutorials, vol. 16, no. 1, pp. 303–336, 2013.

[2] K. Finnerty, S. Fullick, H. Motha, J. N. Shah, M. Button, and V. Wang, “Cyber security breaches survey 2019,” 2019.

[3] Z. E. Huma, S. Latif, J. Ahmad, Z. Idrees, A. Ibrar, Z. Zou, F. Alqahtani, and F. Baothman, "A hybrid deep random neural network for cyberattack detection in the industrial internet of things," IEEE Access, vol. 9, pp. 55 595–55 605, 2021.

[4] D. E. Denning, "An intrusion-detection model," IEEE Transactions on software engineering, no. 2, pp. 222–232, 1987.

[5] G. G. Liu, "Intrusion detection systems," in Applied Mechanics and Materials, vol. 596. Trans Tech Publ, 2014, pp. 852–855.

[6] C. Chio and D. Freeman, Machine Learning and Security: Protecting Systems with Data and Algorithms. " O'Reilly Media, Inc.", 2018.

[7] A. Ali, S. Shaukat, M. Tayyab, M. A. Khan, J. S. Khan, J. Ahmad et al., "Network intrusion detection leveraging machine learning and feature selection," in 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET). IEEE, 2020, pp. 49–53.

[8] S. Shaukat, A. Ali, A. Batool, F. Alqahtani, J. S. Khan, J. Ahmad et al., "Intrusion detection and attack classification leveraging machine learning technique," in 2020 14th International Conference on Innovations in Information Technology (IIT). IEEE, 2020, pp. 198–202.

[9] M. A. khan, M. A. Khan, S. Latif, A. A. Shah, M. U. Rehman, W. Boulila, M. Driss, and J. Ahmad, "Voting classifier-based intrusion detection for iot networks," in 2021 2nd International Conference of Advance Computing and Informatics (ICACIN). Springer, 2021.

[10] L. N. Tidjon, M. Frappier, and A. Mammar, "Intrusion detection systems: A cross-domain overview," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3639–3681, 2019.

[11] A. Shenfield, D. Day, and A. Ayesh, "Intelligent intrusion detection systems using artificial neural networks," ICT Express, vol. 4, no. 2, pp. 95–99, 2018.

[12] M. Rege and R. B. K. Mbah, "Machine learning for cyber defense and attack," DATA ANALYTICS 2018, p. 83, 2018.

[13] S. Latif, Z. Zou, Z. Idrees, and J. Ahmad, “A novel attack detection scheme for the industrial internet of things using a lightweight random neural network,” IEEE Access, vol. 8, pp. 89 337–89 350, 2020.

[14] K. Keshari, Top 10 Applications of Machine Learning: Machine Leaning Applications in Daily Life, 2020.

[15] R. Sober, Data Breach Response Times: Trends and Tips, 2020.

[16] A. Shafique, J. Ahmed, W. Boulila, H. Ghandorh, J. Ahmad, and M. U. Rehman, “Detecting the security level of various cryptosystems using machine learning models,” algorithms, vol. 1, p. 5, 2021.[1] S. Sarkar, S. Chatterjee and S. Misra, "Assessment of the Suitability of Fog Computing in the Context of Internet of Things," IEEE Transactions on Cloud Computing, vol. 6, no. 1, pp. 46-59, 2018.

[2] M. Al-Kasassbeh and M. Adda, "Network fault detection with Wiener filter-based agent," Journal of Network and Computer Applications, vol. 32, no. 4, pp. 824-833, 2009.

[3] M. Ge, X. Fu, N. Syed, Z. Baig, G. Teo and A. Robles-Kelly, "Deep Learning-Based Intrusion Detection for IoT Networks," in 2019 IEEE 24th Pacific Rim International Symposium on Dependable Computing (PRDC), 2019.

[4] F. Hussain, S. A. Hassan, R. Hussain and E. Hossain, "Machine Learning for Resource Management in Cellular and IoT Networks: Potentials, Current Solutions, and Open Challenges," IEEE Communications Surveys and Tutorials, vol. 22, no. 2, pp. 1251-1275, 2020.

[5] M. Al-Kasassbeh, "Network Intrusion Detection with Wiener Filter-based Agent," , 2011.

[6] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas and Y. Zhou, "Understanding the mirai botnet," in SEC'17 Proceedings of the 26th USENIX Conference on Security Symposium, 2017.

[7] A. S. Tarawneh and A. B. Hassanat, "DeepKnuckle: Deep Learning for Finger Knuckle Print Recognition," Electronics, vol. 11, no. 4, 2022.

[8] M. Al-Kasassbeh, M. Almseidin, K. Alrfou and S. Kovacs, "Detection of IoT-botnet attacks using fuzzy rule interpolation," Journal of Intelligent and Fuzzy Systems, vol. 39, no. 1, pp. 421-431, 2020.

[9] M. Alkasassbeh, G. Al-Naymat, A. Hassanat and M. Almseidin, "Detecting Distributed Denial of Service Attacks Using Data Mining Techniques," International Journal of Advanced Computer Science and Applications, vol. 7, no. 1, 2016.

[10] A. Abuzuraiq, M. Alkasassbeh and A. Mohammad, "Intelligent methods for accurately detecting phishing websites," in 11th International Conference on Information and Communication Systems (ICICS), 2020.

[11] D. Denning, "An Intrusion-Detection Model," IEEE Transactions on Software Engineering, vol. 13, no. 2, pp. 222-232, 1987.

[12] M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," in 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), 2017.

[13] M. M. Rathore, A. Ahmad and A. Paul, "Real time intrusion detection system for ultra-high-speed big data environments," The Journal of Supercomputing, vol. 72, no. 9, pp. 3489-3510, 2016.

[14] M. Alkasassbeh and M. Almseidin, "Machine Learning Methods for Network Intrusion Detection.," arXiv preprint arXiv:1809.02610, 2018.

[15] M. Almseidin, J. Al-Sawwa and M. Alkasassbeh, "Anomaly-based Intrusion Detection System Using Fuzzy Logic," in 2021 International Conference on Information Technology (ICIT), 2021.

[16] R. Vishwakarma and A. K. Jain, "A survey of DDoS attacking techniques and defence mechanisms in the IoT network," Telecommunication Systems, vol. 73, no. 1, pp. 3-25, 2020.

[17] A. Azmoodeh, A. Dehghantanha and K.-K. R. Choo, "Robust Malware Detection for Internet of (Battlefield) Things Devices Using Deep Eigenspace Learning," IEEE Transactions on Sustainable Computing, vol. 4, no. 1, pp. 88-95, 2019.

[18] Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher and Y. Elovici, "N-BaIoT— Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders," IEEE Pervasive Computing, vol. 17, no. 3, pp. 12-22, 2018.

[19] B. Roy and H. Cheung, "A Deep Learning Approach for Intrusion Detection in Internet of Things using BiDirectional Long Short-Term Memory Recurrent Neural Network," in 2018 28th International Telecommunication Networks and Applications Conference (ITNAC), 2018.

[20] A. Dawoud, S. Shahristani and C. Raun, "Deep learning and software-defined networks: Towards secure IoT architecture," Internet of Things, pp. 82-89, 2018.

[21] Y. Zhou, M. Han, L. Liu, J. S. He and Y. Wang, "Deep learning approach for cyberattack detection," in IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2018.[22] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman and A. Alazab, "A novel ensemble of hybrid intrusion detection system for detecting internet of things attacks," Electronics, vol. 8, no. 11, p. 1210, 2019.

[23] O. Ibitoye, O. Shafiq and A. Matrawy, "Analyzing Adversarial Attacks against Deep Learning for Intrusion Detection in IoT Networks," in 2019 IEEE Global Communications Conference (GLOBECOM), 2019.

[24] Z. A. Baig, S. Sanguanpong, S. N. Firdous, V. N. Vo, T. G. Nguyen and C. So-In, "Averaged dependence estimators for DoS attack detection in IoT networks," Future Generation Computer Systems, vol. 102, pp. 198-209, 2020.

[25] M. A. Ferrag, L. A. Maglaras, S. Moschoyiannis and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," Journal of Information Security and Applications, vol. 50, p. 102419, 2020.

[26] Y. N. Soe, Y. Feng, P. I. Santosa, R. Hartanto and K. Sakurai, "Towards a lightweight detection system for cyber attacks in the IoT environment using corresponding features," Electronics, vol. 9, no. 1, p. 144, 2020.