International Journal of Innovative Research in Engineering
Volume 4, Issue 2 (March-April 2023), PP: 383-387.
www.theijire.com ***ISSN No: 2582-8746***

Machine Learning Algorithms for IoT Security

Mani Arora[1], Sukhwinder Kaur[2], Anureet Kaur[3]
[1,2,3] P.G., Department of Computer Science and Applications, Khalsa College, Amritsar, India.

How to cite this paper:
Mani Arora[1], Sukhwinder Kaur[2], Anureet Kaur[3], "Machine Learning Algorithms for IoT Security", IJIRE-V4I02-383-387.

https://www.doi.org/10.59256/ijire2023040205

Abstract: *Worldwide Internet of Things has been rapidly evolved within smart devices that interact with each other via machine to machine communications. It is revolutionizing the IT market and will have deep, economic and social impact on our lives. The interconnecting devices in IoT networks usually become targets for cyber-attacks. In this regard, far-reaching efforts have been made to address the security and privacy issues in IoT networks primarily through symmetric and asymmetric cryptographic approaches. Due to resource constraints, heterogeneity, massive real time data generated by the IoT devices the traditional cryptographic algorithm become insufficient to encompass the entire security to the IoT devices. Therefore, Machine Learning (ML) and Deep Learning (DL) techniques, which are able to provide embedded intelligence in the IoT devices and networks, are leveraged to cope with different security problems. In this paper, we systematically review the security algorithms, attack parameters, and the current security solutions for the IoT networks.*

Key Word: *Internet of Things (IoT), IoT Applications, Security, Attacks, Privacy, Machine Learning, Deep Learning.*

I.INTRODUCTION

IoT is considered as an interconnected and distributed network of embedded systems communicating through wired or wireless communication technologies [1]. These devices are deployed in a wide variety of consumer applications at offices, buildings, home, smart healthcare systems. The cross cutting and large scale nature of IoT systems with various components lead to threat of data security. It gives rise to new challenges in area of data security. Also the IoT uses enabling technologies such as Software-Defined Networking (SDN), Cloud Computing (CC), and fog computing, also increases the landscape of threats for the attackers.

Data generated by the IoT devices is enormous and therefore, outmoded data collection, storage, and processing techniques may not work at this gauge. Several studies have been done which focus on security of popular consumer and Industrial IoT devices and highlighting the associated security susceptibilities. Therefore, to harness the value of the IoT-generated data, new mechanisms are needed. In this context, Machine Learning (ML) is considered to be one of the most suitable computational paradigms to provide embedded intelligence in the IoT devices [3]. ML can help machines and smart devices to infer useful knowledge from the device-or human- generated data. ML techniques have been used in tasks such as classification, regression and density estimation. Variety of applications such as computer vision, fraud detection, bio-informatics, malware detection, authentication, and speech recognition use ML algorithms and techniques. In this paper, however, we focus on the applications of ML in providing security and privacy services to the IoT networks.

II.SECURITY CHALLENGES IN IOT DISPOSITION

The two main factors in the commercial comprehension of the IoT services and applications are security and privacy. It is highlighted in various studies that the flooding of Iot devices over the next decade would cause serious security issues, lack of regulations and cyber threats. The IoT domain have been extensively researched in domains like communication security, data security, privacy, architectural security, identity management, malware analysis, and so on [2]. Most of the security challenges are complex and the solutions cannot be distinct.

III.MACHINE LEARNING ALGORITHMS USED IN IOT SECURITY

Machine learning refers to smart methods used to augment performance criteria using sample data or past experience(s) through learning. ML algorithms through learning build models of activities using mathematical techniques on huge data sets act as a basis for making future predictions based on the newly input data. Traditional approaches are widely used for different domains of IoT including applications, services, security etc.

Some of the security related real-world applications of ML are as follows:

Face recognition for forensics: pose, lighting, occlusion (glasses, beard), make-up, hair style, etc. Character recognition for security encryption: different handwriting styles.

Malicious code identification: identifying malicious code in applications and software.

Basic Machine Learning Algorithms are classified into four categories: supervised, unsupervised,semi-supervised, and reinforcement learning algorithms.

- Supervised Learning: Supervised learning is performed when specific targets are defined to reach from certain set of inputs. For this type of learning, the data is first labelled followed by training with labelled data (having inputs and desired

outputs). It tries to identify automatically rules from available datasets and define various classes, and finally predict the belonging of elements (objects, individuals, and criteria) to a given class.

- Unsupervised Learning: In unsupervised learning, the environment only provides inputs without desired targets. It does not require labelled data and can investigate similarity among unlabelled data and classify the data into different groups. Supervised learning and unsupervised techniques mainly focus on data analysis problems while reinforcement learning is preferred for comparison and decision-making problems.

- Semi-supervised Learning: In the previous two types, either there are no labels for all the observation in the dataset or labels are present for all the observations. Semi-supervised learning falls in between these two. In many practical situations, the cost to label is quite high, since it requires skilled human experts to do that. So, in the absence of labels in the majority of the observations but present in few, semi-supervised algorithms are the best candidates for the model building.

- Reinforcement Learning: In Reinforcement Learning (RL), no specific outcomes are defined, and the representative learns from feedback after interacting with the environment. It performs some actions and makes decisions on the basis of the rewardobtained.

Some of the most common machine learning algorithms used for IoT security is as follows:

Machine Learning Algorithms Used In Iot Security

Machine Learning Algorithm	Description
Naïve Bayes	It is the classification algorithm used with binary and multi-class Environment. It is named as ―Naive‖, as over-simplified assumptions are made for the calculation of probabilities for specific hypothesis. All the characteristics are assumed to be conditionally independent instead of calculating the actual values [5].
K-Nearest Neighbour	It is simple and effective supervised learning algorithm and is used for connecting new data points to the existing similar points by searching through the available dataset. The model is trained and grouped according to some criteria and incoming data is checked for similarity within K neighbours [6].
K-Means Algorithm	The most commonly used well know technique is K-means clustering algorithm belonging to the unsupervised category of ML family. K-Means clustering is used to classify or group devices based on attributes or parameters, into K number of groups, where K is a positive integer number and its value has to be known for the algorithm to work [7].
Random Forest and Decision Tree (DT)	It is a supervised learning method. It defines a model by implementing certain rules inferring from the data features. Afterwards, this model is used to predict the value of new targeted variable. Decision tree is used in classification and as well as regression problems. Essentially, these trees are used to split dataset into several branches based on certain rules [8].
Support Vector Machines (SVM)	SVM is a supervised ML algorithm with low computational complex-ity, used for classification and regression. It has the ability to work with binary as well as with multi-class environments [9], [10]. It classifies input data into n dimensional space and draws n 1 hyperplane to divide the entire data points into groups.
Recurrent Neural Networks (RNN)	This is a supervised learning algorithm used to develop a cascaded chain of decision units for solving the complex problems . It essentially constructs network with certain number of inputs to trigger outputs. Various types of neural networks have been proposed in the literature, e.g. Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) [11], [12], [13].
Principal Component Analysis	It is an unsupervised ML algorithm and is multivariate technique for data compression. It performs dimensionality reduction in large data sets and extracts useful information in the form of set comprised of orthogonal variables known as ―principal components‖. These

	components are organized in an increasing order of variance where first component is associated with highest variance of the data and it continues to the last. The least variance components having least information can be discarded [14].
Q-Learning	It is used for resource scheduling in spectrum management in addition to security in IoT. Q-learning belongs to reinforcement learning (RL) class of the ML. In RL, an agent learns by trial and error that how its actions effect the environment. It estimates the reward after each action and moves to the new state accordingly [7]. It will get reward for good action and penalty for bad actions.
Deep Learning	It is essentially a feed forward Neural Network (NN) in which each neuron is connected to another layer and no connection exists within the layer. The term deep learning refers to multiple layers holding multiple levels of perception such that each layer receives input from the previous layer and feeds the result to following layer [15].

IV.LIMITATIONS IN APPLYING MACHINE LEARNING IN IOT NETWORKS

IoT traffic is usually characterized by its sheer volume, miscellany, variable speed, and ambiguity. Most of the traditional ML techniques are not inherently effectual and ascendable enough to manage IoT data and thus need considerable modifications [4]. Moreover, characteristic qualms exist in IoT data and are difficult to expunge this inherent randomness. In the following, we discuss some of the common limitations of using ML techniques in IoT networks.

1) Processing power and energy: The complexity issues associated with machine learning algorithms are memory, computational, and sample complexity. Also, conventional ML approaches lack scalability and are only limited to low-dimensional problems. IoT devices are small and typically have energy constraints with limited processing power. Therefore, direct application of conventional ML techniques is not suitable in resource-constrained environments. On the other hand, smart IoT devices require real-time data processing for real time applications, while traditional ML techniques are not designed to handle constant streams of data in a real- time. In the wake of such limitations, it is imperative to merge the existing streaming solutions with ML algorithms; however, it will increase the overall complexity of an algorithm.

In addition to this, ML-based networks are developed with presumptuous that the entire data set is available for processing during training phase. However, this is not true for the IoT data. This phenomenon gives rise to various challenges when traditional techniques have to handle an unprecedented volume of data. Also, predictive ability of an algorithm decreases with the increase in the dimensionality of data [16]. The preceding discussion is, at par, applicable for the security-related functions in the IoT where real-time data is processed for possible attack vectors such as intrusion and so on.

2) Data management and analytics: Wireless data can be generated from different sources including networked information systems, and sensing and communication devices [17]. Data is the most important component for IoT systems where efficient analysis must be performed to obtain meaningful information from the data; however, massive data management is a serious challenge in IoT from every application standpoint. The data generated in IoT networks is diverse in nature with different types, formats and semantics, thus exhibiting syntactic and semantic heterogeneity. Syntactic heterogeneity refers to diversity in the data types, file formats, encoding schemes, and data models. While semantic heterogeneity refers to differences in the meanings and interpretations of the data. Such heterogeneity leads to problems in terms of efficient and unified generalization, specifically in case of big data and various datasets with different attributes.

Machine learning assumes that the statistical properties across the entire dataset remain the same, and requires pre-processing and cleaning of the data before fitting within a specific model. However, that is not the case in real-world where data from various sources have different formatting and representations. Furthermore, there might also be differences among different parts of the same dataset. This situation causes difficulties for machine learning algorithms because the algorithms are usually not designed to handle semantically and syntactically diverse data. This phenomenon advocates for efficient solutions to the heterogeneity problem.

V.FUTURE RESEARCH CHALLENGES

In this section, we discuss the challenges faced by and future research opportunities in ML and DL techniques for the security of IoT networks.

The heterogeneity of the IoT network enables the production of sheer amount of data with very high frequency from different domains. This huge amount of data from varying data leads to various issues related to data collection, security, dependency and unavailability of appropriate and enough data sets. More in-depth investigation is needed to come up with big data techniques to cater with the volume and heterogeneity of IoT-generated data.

1) **Data collection**: Data collection from every domain is not straight forward. For instance, data collected from vehicular sensors is used for the organization and safeguarding of the vehicles and also for the efficient traffic management. Similarly data generated from smart home machines and body sensors contain personal information that would easily jeopardize user privacy. Similarly, medical data of the patience collected by the service providers have similar challenges. All the aforementioned data is subject to processing with ML and DL algorithms. However, unbalanced negative and

positive data and false positive will have catastrophic consequences on our lives. Therefore, the ML and DL techniques must be matured enough to be used commercially in such subtle areas. One way would be to develop ML and DL methods that do not solely rely on the previous data and sharply learn from the patterns and able to decide in a way that minimize the side effects of the possible outcomes. This phenomenon is more important in situations when ML and DL are used for security solutions where false positives and true negative will have dire consequences on the networks. One solution could be the context-aware ML and DL solutions. In this regard, more research is needed.

2) **Proximity effect**: Proximity can play a pivotal role in the data collection. The fact that the Internet is walled-off, adds to the challenges of collecting data in IoT networks. Usually the Internet in such domains has limited access from the outside to mitigate different kinds of attacks; however, it adversely affects the utility of the applications. This phenomenon is also related to the statute indirectly where the data collection policies must be defined for cross- platform, cross-network, and cross-organization. The storage of these data is also subject to in- depth exploration. For instance, where and how to store the medical records of the patients, whether in public or private cloud, who should have access to such data, how to apply ML and DL algorithms to such data and what level of privacy should be preserved by the ML and DL algorithms.

3) **Data dependency**: DL algorithms are data dependent and are required to learn gradually from data, and work best if high superiority data is available. Similar to human brain that needs lots of experience to learn first and afterwards is able to make conclusions; it requires huge amount of data to make powerful notions. Similarly, DL algorithm can predict about company's stocks after thoroughly learning about historic rise and fall in company's stocks. If enough data is not available, the DL system will fail to make deep forecasts. This is due to the fact that DL algorithm works in a methodical way; that is, first learn about the domain and afterwards solve the problem. During the training phase, DL algorithm starts from scratch and needs a huge number of data/parameters to tune/play around.

4) **Unavailability of training datasets**: Efficient use of ML and DL solutions need profound datasets that are currently missing. Furthermore, the rules and policies required for defining the learning approaches still need to be explored. Additionally, reliable datasets from real physical environment are required to analyse and compare the recital of various DL and RL algorithms. To date, efforts have been made to cope with this challenge, but more research is needed in this direction.

VI.CONCLUSION

IoT security and privacy are of paramount importance and play a crucial role in the commercialization of the IoT technology. Traditional security and privacy solutions suffer froma number of issues that are related to the dynamic nature of the IoT networks. ML and more specifically DL and DRL techniques can be used to enable the IoT devices to adapt to their dynamic environment. These learning techniques can support self-organizing operation and also optimize the overall system performance by learning and processing statistical information from the environment (e.g. human users and IoT devices). These learning techniques are inherently distributed and do not require centralized communication between device and controller. However, the datasets needed for ML and DL algorithms are still scarce, which makes benchmarking the efficiency of the ML- and DL-based security solutions a difficult task. In this paper, we have considered the role of ML and DL in the IoT from security and privacy perspective. We have discussed the security and privacy challenges in IoT, attack vectors, and security requirements. We have described different ML and DL techniques and their applicationsto IoT security. We have also shed light on the limitations of the traditional ML mechanisms. Then we have discussed the existing security solutions and outlined the open challenges and future research directions. In order to mitigate some of the shortcomings of machine learning approaches to IoT security, the theoretical foundations of DL and DRL will need to be strengthened so that the performances of the DL and DRL models can be quantified based on certain parameters such as computational complexity, learning efficiency, parameter tuning strategies, and data driven topological self-organization. Furthermore, new hybrid learning strategies and novel data visualization techniques will be required for intuitive and efficient data interpretation.

References

1. *O. Novo, N. Beijar, and M. Ocak (2015), –Capillary Networks - Bridging the Cellular and loT Worlds ,‖ IEEE World Forum on Internet of Things (WF-IoT), vol. 1, pp. 571–578.*
2. *2 .J. Granjal, E. Monteiro, and J. S. Silva (2015), –Security for the internet of things: A survey of existing protocols and open research issues,‖ IEEE Communications Surveys Tutorials, vol. 17, pp. 1294–1312.*
3. *M. at. El (2018), –Machine Learning for Internet of Things Data Analysis:A Survey ,‖ Journal of Digital Communications and Networks, Elsevier, vol. 1, pp. 1–56.*
4. *J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng (2016), –A survey of machine learning for big data processing,‖ EURASIP Journal of Advance Signal Process.*
5. *L. Deng, D. Li, X. Yao, D. Cox, and H. Wang (2018), –Mobile network intrusion detection for iot system based on transfer learning algorithm,‖ Cluster Computing.*
6. *R. Doshi, N. Apthorpe, and N. Feamster (2018), –Machine learning ddos detection for consumer internet of things devices,‖ in 2018 IEEE Security and Privacy Workshops (SPW), pp. 29–35.*
7. *F. Hussain, A. Anpalagan, A. S. Khwaja, and M. Naeem (2015), –Resource Allocation and Congestion Control in Clustered M2M Communication using Q-Learning,‖ Wiley Transactions on Emerging Telecommunica-tions Technologies.*
8. *M. S. Alam and S. T. Vuong (2013), –Random forest classification for detecting android malware,‖ in 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing,*

pp. 663–669.

9. *W. Zhou and B. Yu (2018), ―A cloud-assisted malware detection and suppres-sion framework for wireless multimedia system in iot based on dynamic differential game,‖ China Communications, vol. 15, pp. 209–223.*
10. *H.-S. Ham, H.-H. Kim, M.-S. Kim, and M.-J. Choi (2018), ―Linear svm-based android malware*
11. *H. HaddadPajouh, A. Dehghantanha, R. Khayami, and K.-K. R. Choo (2018), ―A deep recurrent neural network based approach for internet of things malware threat hunting,‖Future Generation Computer Systems, vol. 85,pp 88 – 96.*
12. *E. B. Karbab, M. Debbabi, A. Derhab, and D. Mouheb (2018), ―Maldozer: Automatic framework for android malware detection using deep learn-ing,‖ Digital Investigation, vol. 24, pp. S48 – S59.*
13. *J. Su, D. V. Vargas, S. Prasad, D. Sgandurra, Y. Feng, and K. Sakurai (2018), ―Lightweight classification of iot malware based on image recognition,‖ CoRR, vol. abs/1802.03714.*
14. *N. An, A. Duff, G. Naik, M. Faloutsos, S. Weber, and S. Mancoridis (2017), ―Behavioral anomaly detection of malware on home routers,‖ in 2017 12th International Conference on Malicious and Unwanted Software (MALWARE), pp. 47–54.*
15. *A. Azmoodeh, A. Dehghantanha, and K. R. Choo (2018), ―Robust malware de-tection for internet of (battlefield) things devices using deep eigenspace learning,‖ IEEE Transactions on Sustainable Computing, pp. 1–1.*
16. *A. LHeureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz (2017), ―Machine Learning With Big Data: Challenges and Approaches,‖ IEEE Access, vol. 5, pp. 7776 – 7797.*
17. *T. E. Bogale, X. Wang, and L. B. Le (2018), ―Machine Intelligence Techniques for Next- Generation Context-Aware Wireless Networks,‖ Arxiv, vol. 19,1–10..*