

A Distributed Intelligent Intrusion Detection System based on Parallel Machine Learning and Big Data Analysis

Faten Louati^{1,4} ^a, Farah Barika Ktata^{2,4} and Ikram Amous Ben Amor^{3,4}

¹Faculty of Economics and Management of Sfax, Tunisia

²Higher Institute of Applied Sciences and Technology of Sousse, Tunisia

³National School of Electronics and Telecommunications of Sfax, Tunisia

⁴Multimedia, Information Systems and Advanced Computing Laboratory (MIRACL), Tunisia

Keywords: Intrusion Detection, Big Data, Reinforcement Learning, Multi Agent System.

Abstract: Networking security continue to be a serious challenge for all domains because of the increasing number of attacks launched every day due to the advent of connected devices and the emergence of the Internet. Hence, Intrusion detection system comes into focus, especially with the inception of big data challenges. In this paper, we propose a distributed and parallel intrusion detection system suitable for big data environments using machine learning-based multi agent system and big data analysis.

1 INTRODUCTION

Machine Learning (ML) techniques are investigated for malicious purposes as well as for good purposes. In fact, Hackers spares no effort to benefit from the novel technologies and intelligent ML algorithms to launch novel attacks. For this reason, researchers are facing a very serious challenge, they should find solutions to secure networks from zero day attacks and win the war. The largest amount of proposed solutions is based on Intrusion Detection Systems (IDS). An IDS is a kind of software that monitors, analyzes networking traffics and send an alert automatically once a malicious activity is detected (Louati and Ktata, 2020). Two major techniques are used in the intrusion detection field, namely signature-based technique (misuse detection) and anomaly-based detection (behavioral detection). Signature-based technique is based on a set of rules (signatures) stored in a database and present the attacks known by the system. The technique consists of comparing the activities of the network with the stored patterns, if any similarity is detected, then an alarm is sent automatically to the administrator. The main advantage of this technique is that it reduces considerably the false alarm rate. However, it fails to detect new attacks, even a little deviation in known attacks can deceive it. However, hackers quickly learned to use a variety of techniques


to modify their attacks to avoid detection. Thus, signature database should be frequently updated which is a hard task. Anomaly-based detection in turn consists of comparing the activities in the network with the normal behavior. Hence, the alarm is triggered when an abnormal event is detected. Therefore, this technique can efficiently detect known, as well as unknown attacks. However, it generates a high false alarm rate, which is the main drawback of this technique.

Most of the existing and well-used IDSs are signature-based (Kukielka and Kotulski, 2008), so they usually fail to detect unseen attacks, also they are not able to handle big data requirements. To tackle this problem, we propose an intelligent and distributed IDS using multi agent system based on parallel ML algorithms.

The remaining part of the paper is organized in the following way: In section 2 we discuss some previous works. The description of the proposed IDS and our contributions are introduced in section 3. Finally, section 4 concludes the entire work and describes our future plans.

2 RELATED WORKS

Several previous works are investigated to build an intrusion detection mechanism for networks. Recently,

^a  <https://orcid.org/0000-0002-8582-6092>

with the emergence of ML, many researches benefit from it. In instance, (Diro and Chilamkurti, 2017) presented distributed and parallel IDS for IoT using DL. Experimental results on the NSL KDD dataset show that DL provides greater accuracy compared to other ML algorithms, however, it consumes larger time in training.

The proposed intrusion detection framework in (Rathore and Park, 2018) is based on semi-supervised learning. To this end, the authors combine two algorithms, namely, Extreme Learning Machine (ELM), in which the neural network is composed of a single hidden layer, and semi supervised Fuzzy c-means (SFCM) which is based on the concept of different types of prior knowledge to enhance clustering performance. In this paper labels are considered as the prior knowledge.

(Pajouh et al., 2018) used deep recurrent neural network (RNN) to detect malware by analyzing application's operation codes. The evaluation of the proposed model achieved an accuracy of 98%.

The work in (Mahmudul et al., 2019) compared different ML algorithms, namely, logistic Regression, support vector machine (SVM), decision tree, random forest and artificial neural network and the results show that random forest is the most efficient technique in detect attacks for IoT networks with 99.4% of accuracy rate.

(Zhang et al., 2019) proposed an intrusion detection scheme using genetic algorithm (GA) and deep believe network (DBN). Evaluation on NSL KDD dataset provided more than 99% of detection rate.

Multi agent system (MAS) approach has been also exploited in detection task. In instance, (Achbarou et al., 2018) proposed a distributed IDS (DIDS) for cloud computing that combines both misuse detection and anomaly detection techniques. The proposed model is based on a number of agents. The authors performed a comparison between their model and similar systems that not used MAS and they concluded that IDS with MAS worked better in term of efficiency and detection time.

Other works used MAS approach in intrusion detection such as (Al-Yaseen et al., 2016) in which authors used two classifiers with MAS namely SVM and ELM, and (Mehmood et al., 2018) which used naive Bayes algorithm.

There are some existing works that used reinforcement learning (RL) algorithm with MAS technology in intrusion detection task for IoT such as (Servin and Kudenko, 2007), however, there is a lack of communication between agents which is an essential requirement to create MAS.

(Arel et al., 2010) provided an IDS based on DRL

for cloud computing.

(Nie et al., 2021) introduce an IDS for green IoT (composed of a number of devices connected to application) based on DRL, the model performs good results (more than 99% of prediction rate).

(Sethi et al., 2020), (Gu et al., 2020) and (Arel et al., 2010) propose an IDS for cloud computing and IoT based on DRL.

Ather works investigated zero shot learning (ZSL) such as, (Zhang et al., 2020) who used ZSL based on autoencoder. This method gives 88.3% of accuracy rate with NSL KDD dataset.

(Zerhoudi et al., 2020) proposed an IDS using zero shot recognition via graph embedding. The results show an accuracy rate = 88.3% in NSL KDD dataset.

Although those cited works performed very good results, their proposed solutions are not suitable for new environments based on big data. Hence, they are not able to deal with big data challenges.

Few papers focus on Intrusion detection systems in big data frameworks (Hassan et al., 2020), such as the work of (Terzi et al., 2017) which detects network anomaly from Netflow data using clustering algorithm. The work was tested on the CTU-i3 dataset and performs 96% of accuracy rate. However, it causes high false alarm rate (FAR).

In addition, (Hassan et al., 2020) used conventional neural network (CNN) and weight dropped long short-term memory (WDLSTM) network in big data context. Experiments are performed on UNSW-NB15 and show good results (e.g. accuracy=97.17%).

3 METHODOLOGY

3.1 Concepts

3.1.1 Big Data

The term big data dates back to 2005, it is typically defined with three words: volume, velocity and variety (the famous 3Vs of big data):

- Volume: Size of the data; Big data refers to a huge amount of data.
- Velocity: The speed of the data; Big data characterized by its speed of movement.
- Variety: Diversity of data; Varied Big data are coming from diverse sources, it could be structured, non structured or semi structured.

3.1.2 Multi Agent System

Multi agent system is a collection of entities called agents that are communicating with each other. According to the definition of FIPA (Foundation for the Intelligent Physical Agent), an agent is a kind of entity with autonomy, activity, mobility, re-activity, sociality, intelligence and other features (FIPA, 2002) (Wooldridge, 2009). (Oprea, 2004) defines the multi-agent system equation as it "states that in a multi-agent system a task is solved by agents that communicate among them."

3.2 Proposed Solution

Our solution consists of merging machine learning algorithms with Big data analysis techniques to create an intrusion detection system able to perform detection in new big data environments. Big data means a huge amount of data and that what new machine learning algorithms need to make their models.

Also benefiting from multi agent system is a promising idea as multi agent systems help to improve reliability and availability and ensure information collecting, sharing and processing in distributed networks.

Figure 1 depicts the structure of our solution: The model is composed of a combination of multi agent system, machine learning algorithm and big data analytic. Firstly, the network traffic is collected by Agent Sniffer then sent to the Agent preprocessor to be processed. After that, the data are sent to Hadoop framework to be analyzed in parallel manner. This framework in turn is composed of a name node which consists of an Agent decision maker and a number of data nodes (at least three nodes). The agent decision maker runs reinforcement learning (RL) algorithm -which is sub-filed of machine learning- in the name node and detect intrusions.

RL is about an agent interacting with the environment, learning an optimal policy, by trial and error (and not by simple recognition of data that may have been provided to it which is the case of traditional ML techniques): the program repeats one or more experiences a large number of times in order to learn to take good choices and avoid bad choices in a given situation. This learning method allows agents to understand the environment in which they operate, to build their own decision-making mechanism, and to act with more rationality. Moreover, this learning technique could ultimately make agents even smarter by allowing them to react to unexpected situations. See (Li, 2017) for deep details.

Due to the dynamic changes of networks, new

kinds of attacks are emerging every day. DRL could be a good choice to deal with this challenge as it enable the agent to learn with a rational manner by optimizing its choices from previous experiences simulated by himself.

All the agents in the MAS ensure communications between them.

Merging Big data analysis, recent machine learning algorithms and multi agent system seems promising solution since it performs 1) parallels of the processing of large volume of data by means Hadoop framework, 2) Excellent classification of the data using efficient machine learning classifiers and 3) Enhance time consumption and performance using multi agent system.

3.3 Experiments

This paper presents a work in progress, the implementation of this framework is not yet achieved. So far, we build an IDS based on Deep Reinforcement learning agent, This agent is expected to be executed in the name node.

RL agent is trained on NSL KDD which is a very large dataset so it well represents the context of big data. NSL KDD (NSL,) is an enhanced version of KDD 99 benchmark dataset (Ring et al., 2019) and is well known and wide used by researchers in intrusion detection field. Figure2 describes the dataset which is composed of three sub-datasets KDDTrain+20% composed of 25192 samples, KDDTrain+ composed of 125973 samples and KDDTest+ composed of 22544 samples. The dataset is about 42 feature; the last one presents whether the record is normal or one of the four known networking attacks namely, Denial of Service (DoS) Probe, User to Root (U2R) and Remote to Local (R2L).

Firstly, the dataset was preprocessed. The preprocessing phase is composed of three steps 1) Numericalization; as there are symbolic and categorical features in the dataset 2) Removing attributes with missing data 3) Data scaling because the data have varying ranges.

The RL agent is based on deep Q network(DQN) algorithm which the neural network is composed of 1 input layer (122 neurons, i.e the number of features of NSL KDD data set after being preprocessed), three hidden layers composed of 80, 50 and 20 neurons respectively and an output layer composed of 5 neurons which refer to the five classes (normal, DoS, U2R, Probe and R2L). All the layers are fully connected and use Relu as activation function. Tables 2 and 1 describe the parameters of the model which performs good results (see Table 3).

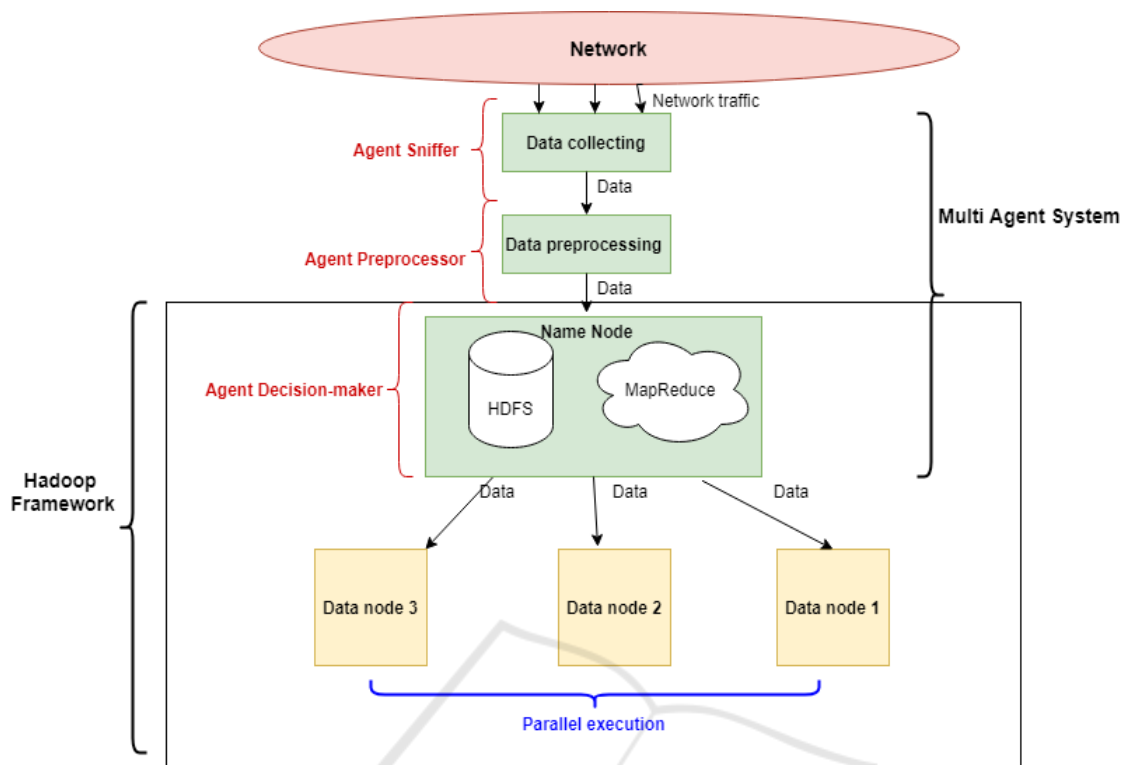


Figure 1: The structure of the proposed solution.

Dataset	Number of Records:					
	Total	Normal	DoS	Probe	U2R	R2L
KDDTrain+20%	25192	13449 (53%)	9234 (37%)	2289 (9.16%)	11 (0.04%)	209 (0.8%)
KDDTrain+	125973	67343 (53%)	45927 (37%)	11656 (9.11%)	52 (0.04%)	995 (0.85%)
KDDTest+	22544	9711 (43%)	7458 (33%)	2421 (11%)	200 (0.9%)	2654 (12.1%)

Figure 2: Description of NSL KDD dataset (Saporit, 2019).

Table 1: Reinforcement learning components.

Algorithm	Environment	States	Actions	Rewards
DQN	Network	Features of the dataset	attack types + normal	1 if good classification 0 if not

The objective of this paper is to share our solution and explain our idea. The next step consists of :

- 1) Improving the results
- 2) Implementing the whole proposed scheme, i.e. Hadoop and multi agent system

- 3) Performing experimentation with the dataset as well as real network traffic
- 4) Performing near-real time detection.

Table 2: Reinforcement learning parameters.

Policy	Epsilon greedy
Optimizer	Adam
batch size	1
minibatch size	500
decay rate	0.99
gamma	0.001

Table 3: Experimental results.

Accuracy	75.65%
Precision	79.51%
F1 score	72.58%
Recall	75.65%

4 CONCLUSION

In this paper, we presented a short description of our proposed solution for security problem in networks. We introduced firstly the general context of this research, then we listed a number of existing solution in literature. Next we described our solution and our contributions. This work is an initial proposal; The next steps are implementation and evaluation of our proposed model using conventional metrics to show its efficiency in the detection of zero-day attacks in real time.

REFERENCES

Nsl kdd. <https://www.unb.ca/cic/datasets/nsl.html>.

Achbarou, O., Kiram, M., Bourkhouk, O., and Elbouanani, S. (2018). *A Multi-agent System-Based Distributed Intrusion Detection System for a Cloud Computing: MEDI 2018 International Workshops, DETECT, MEDI4SG, IWCFS, REMEDY, Marrakesh, Morocco, October 24–26, 2018, Proceedings*, pages 98–107.

Al-Yaseen, A. P. D. W., Othman, Z., and Ahmad Nazri, M. Z. (2016). Real-time multi-agent system for an adaptive intrusion detection system. *Pattern Recognition Letters*, 85.

Arel, I., Liu, C., Urbanik, T., and Kohls, A. (2010). Reinforcement learning-based multi-agent system for network traffic signal control. *Intelligent Transport Systems, IET*, 4:128 – 135.

Diro, A. and Chilamkurti, N. (2017). Distributed attack detection scheme using deep learning approach for internet of things. *Future Generation Computer Systems*.

FIPA (2002). Fipa abstract architecture specification. Available from: <http://www.fipa.org/specs/fipa00001/SC00001L.html>.

Gu, T., Abhishek, A., Fu, H., Zhang, H., Basu, D., and Mohapatra, P. (2020). Towards learning-automation iot attack detection through reinforcement learning. In *2020 IEEE 21st International Symposium on "A*

World of Wireless, Mobile and Multimedia Networks" (WoWMoM), pages 88–97.

Hassan, M. M., Gumaai, A. H., Alsanad, A., Alrubaian, M., and Fortino, G. (2020). A hybrid deep learning model for efficient intrusion detection in big data environment. *Inf. Sci.*, 513:386–396.

Kukielka, P. and Kotulski, Z. (2008). Analysis of different architectures of neural networks for application in intrusion detection systems. *2008 International Multiconference on Computer Science and Information Technology*, pages 807–811.

Li, Y. (2017). Deep reinforcement learning: An overview.

Louati, F. and Ktata, F. (2020). A deep learning-based multi-agent system for intrusion detection. *SN Applied Sciences*, 2.

Mahmudul, H., Islam, M, Z., and M.M.A., H. (2019). Attack and anomaly detection in iot sensors in iot sites using machine learning approaches. 7.

Mehmood, A., Mukherjee, M., Ahmed, S. H., Song, H., and Malik, K. (2018). Nbc-maids: Naïve bayesian classification technique in multi-agent system-enriched ids for securing iot against ddos attacks. *The Journal of Supercomputing*, 74.

Nie, L., Sun, W., Wang, S., Ning, Z., Rodrigues, J. J. P. C., Wu, Y., and Li, S. (2021). Intrusion detection in green internet of things: A deep deterministic policy gradient-based algorithm. *IEEE Transactions on Green Communications and Networking*, 5(2):778–788.

Oprea, M. (2004). Applications of multi-agent systems. In *IFIP Congress Tutorials*.

Pajouh, H. H., Dehghantanha, A., Khayami, R., and Choo, K.-K. R. (2018). A deep recurrent neural network based approach for internet of things malware threat hunting. *Future Gener. Comput. Syst.*, 85:88–96.

Rathore, S. and Park, J. (2018). Semi-supervised learning based distributed attack detection framework for iot. *Appl. Soft Comput.*, 72:79–89.

Ring, M., Wunderlich, S., Scheuring, D., Landes, D., and Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Comput. Secur.*, 86:147–167.

Saporit, G. (2019). A deeper dive into the nsl-kdd data set. <https://towardsdatascience.com/a-deeper-dive-into-the-nsl-kdd-data-set-15c753364657>.

Servin, A. and Kudenko, D. (2007). Multi-agent reinforcement learning for intrusion detection. In *Adaptive Agents and Multi-Agents Systems*.

Sethi, K., Kumar, R., Prajapati, N., and Bera, P. (2020). Deep reinforcement learning based intrusion detection system for cloud infrastructure. In *2020 International Conference on COMMunication Systems NETWORKS (COMSNETS)*, pages 1–6.

Terzi, D. S., Terzi, R., and Sagioglu, S. (2017). Big data analytics for network anomaly detection from netflow data. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 592–597.

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems*. Wiley Publishing, 2nd edition.

Zerhoubi, S., Granitzer, M., and Garchery, M. (2020). Improving intrusion detection systems using zero-shot recognition via graph embeddings. In *2020 IEEE 44th*

Annual Computers, Software, and Applications Conference (COMPSAC), pages 790–797.

Zhang, Y., Li, P., and Wang, X. (2019). Intrusion detection for iot based on improved genetic algorithm and deep belief network. *IEEE Access*, 7:31711–31722.

Zhang, Z., Liu, Q., Qiu, S., Zhou, S., and Zhang, C. (2020). Unknown attack detection based on zero-shot learning. *IEEE Access*, 8:193981–193991.

