# Effective and efficient network anomaly detection system using machine learning algorithm

**Mukrimah Nawir, Amiza Amir, Naimah Yaakob, Ong Bi Lynn**
Embedded, Networks, and Advanced Computing Research Cluster (ENAC),
School of Computer and Communication Engineering (SCCE), Universiti Malaysia Perlis (UniMAP), Malaysia

**Article Info**

***Article history:***

Received Oct 11, 2018
Revised Nov 16, 2018
Accepted Dec 19, 2018

***Keywords:***

Anomaly detection
Averaged One Dependence Estimator (AODE)
Machine learning
UNSW-NB15

## ABSTRACT

Network anomaly detection system enables to monitor computer network that behaves differently from the network protocol and it is many implemented in various domains. Yet, the problem arises where different application domains have different defining anomalies in their environment. These make a difficulty to choose the best algorithms that suit and fulfill the requirements of certain domains and it is not straightforward. Additionally, the issue of centralization that cause fatal destruction of network system when powerful malicious code injects in the system. Therefore, in this paper we want to conduct experiment using supervised Machine Learning (ML) for network anomaly detection system that low communication cost and network bandwidth minimized by using UNSW-NB15 dataset to compare their performance in term of their accuracy (effective) and processing time (efficient) for a classifier to build a model. Supervised machine learning taking account the important features by labelling it from the datasets. The best machine learning algorithm for network dataset is AODE with a comparable accuracy is 97.26% and time taken approximately 7 seconds. Also, distributed algorithm solves the issue of centralization with the accuracy and processing time still a considerable compared to a centralized algorithm even though a little drop of the accuracy and a bit longer time needed.

***Corresponding Author:***

Mukrimah Nawir,
School of Computer and Communication Engineering,
Embedded, Network, and Advanced Computing,
Universiti Malaysia Perlis (UniMAP), 02600, Pauh, Perlis, Malaysia.
Email: mukrimahnawir@yahoo.com

## 1. INTRODUCTION

Computer network required an intelligent system that can monitored and analyzed the activities or behaviors in the network system, intrusion detection system (IDS). This paper focused on anomaly detection where monitor the system by classify the network traffics that representing in a network labelled dataset (UNSW-NB15) either it is a normal data or anomalous data. In simple word, anomaly detection defined as a keen interest in uncovering known or even unknown anomalous complex patterns of various malicious attacks in the network protocol. This cause damage to the functionalities of computer. For example, when the computer behaves in unusual way means that computer had been hacked [1]. Problem in choosing algorithm that classify the data instances are not a straightforward where different applications domains have different viewpoint about anomalies [2]. For instance, in medical domain a small deviation considered anomalies meanwhile not in business environment. Machine learning is one of the helpful tools for anomaly detection because it is automatically learnt the information or behaviors of the system and easily recognizes the complex patterns [3]. Then, it will act based on the data given and decides an intelligent solution [4]. It is primarily enumerating the effective and efficient of the network performance within computer network system. It can be model either in inherent (supervised) or unambiguous (unsupervised) way to classify the patterns or for new information that coming into a network can adopt the network protocol and automatically behave according to present network. These gives various advantages to the implementation of ML in system like high detection rates, false alarm rates, reasonable computation, and low communication [5].

According to 2017 Global State of Information Security Survey, the issue regarding the security attacks in internet of Things (IoT) paradigm is getting more important nowadays [6]. The advanced technologies tend to be harmed by attackers that have intentions towards them. Since, the data is might be in different geographical location might cause the difficulty to monitor these new technological paradigms that involve safety-critical where it will force the operation to be shut down before completely done the classification when the system suddenly been compromise by malicious attacks that arise the issue of centralization [7]. Therefore, the distributed algorithm is required for the availability of collected decision made by whole nodes present in the computer network. In other word, if one of devices (node) leave the network system does not affect or endanger the whole computer network system. The subsequent sections of this paper are organized as follow: the existing works (Section II) where previous researchers conducted the experiment for anomaly detection using machine learning to detect anomalous in various domains. After the literature review of related topics, the paper explains briefly the construction or setup of the experiments in Section III. In Section IV, it shows the experimental result upon different supervised machine learning algorithm based on classification rate and time taken. Lastly, the paper concluded the overall works and future work provided in Section V.

## 2. EXISTING WORKS:ANOMALY DETECTION USING MACHINE LEARNING ALGORITHM

There are numerous numbers of works on anomaly detection using machine learning approaches either using a supervised as well unsupervised strategy that had been employed in various domains. With or without information within the application, the machine learning can achieve a high accuracy of traffic classifier using Bayesian Neural Network by labelling their instances during training stage [8]. For an unsupervised learning, Balagani et al. [9] conducted k-means algorithm to improve the accuracy of classification for anomaly detection. Furthermore, an incremental learning classifier is appropriate to learn the knowledge for a continuous or streaming data. Authors [10] face difficulties to classify a large dataset. Hence, they implemented three supervised learning approaches for binary classification of web data (normal or malware) and believe that it is relevant to be applied as well as in anomaly detection system as they produce a relatively high level of accuracy, sensitivity, and specificity. Table 1 shows some of classification method that using a supervised learning for anomaly detection that employed within various domains. Also, the contributions of each work are provided.

Table 1. Classification using supervised machine learning algorithm for anomaly detection within various domains

| Ref | Algorithm | Contributions | Domain |
|---|---|---|---|
| [11] | Classification | Random forest algorithm more accurate to be implemented in medical wireless sensor network with additive regression methods for anomaly detection | Medical wireless sensor networks |
| [12] | Neural network | This project scalable even in high dimension data involve and simple calculation required | Large system |
| [13] | Self-Organization Maps (SOM) | SOM algorithm applicable to be implemented for large scale virtual machine that only need small amount of time to process the data | Virtual machine (cloud platform) |
| [14] | Classification | In anomly detection for cloud, Random Forest more precise and high recall value because this algorithm uses a large group of decision trees | Cloud-based |
| [15] | Decision tree | Regression tres refine the better detection of anomalis data due to it is high alert alarm can be produced. By considering available data, it can determine if there is phising or spam in a suystem | Web spam |

## 3. EXPERIMENT SETUP

The experiments conducted using an Intel ® Xeon (R) CPU E3-1270 v5 @ 3.60GHz x 8, 16GB RAM and written in JAVA language. We used WEKA version 3.8 and Eclipse, to measure the network performances of various supervised machine learning algorithms (Naïve Bayes, Averaged One Dependence Estimator, Multi-Layer Perceptron, Radial Basis Perceptron Network (RBFN), and J48 Trees) with ten-fold cross validation that validate, prove, and get the best accuracy and fast processing algorithm for anomaly detection of the given network dataset. Table 2 presents the properties of network labelled dataset used (UNSW-NB15) [16, 17]. The experiment conducted in this work is only a binary classification by using the feature of label that consist only normal or anomalous data. Multi-classification is not being done in this paper. Additionally, the number of instances is 257 673 data with 44 features. The features of UNSW-NB15 includes the binary, numeric, and nominal types.

Table 2. Properties of network labelled dataset (UNSW-NB15)

| UNSW-NB15 dataset | | | |
|---|---|---|---|
| Class | Binary | Number of Instances | 257 673 instances |
| Attributes Types | Binary, Numeric, and Nominal | Number of Attributes | 44 features |
| Task | Classification | Area | Computer |

### 3.1. Experiment 1: evaluate the best ML algorithm (centralized algorithm) for anomaly detection over UNSW-NB15 dataset

Presently, experiment 1 used WEKA version 3.8 tool to employed machine learning algorithms for anomaly detection over UNSW-NB15 dataset. There are five classification algorithms used in this experiment includes Naïve Bayes (NB), Averaged One Dependence Estimator (AODE), Radial Basis Function Network (RBFN), Multi-Layer Perceptron (MLP), and J48 trees. There are four stages involved (preparation of dataset, training, validation, and testing) [18, 19]. First, by loading the network dataset (UNSW-NB15) that need to classify the data instances. Once dataset is ready, the second stage (training) being proceed. Chose the algorithms (as mentioned earlier) that used in this experiment. The parameters of algorithms used set to default as in WEKA. Third, tenfold cross validation. Ten-fold cross validation high in accuracy even though the scarce data where no wasting information data. Last, the testing stage by collect the performance measures (accuracy and time taken).

### 3.2. Experiment 2: Distributed Algorithm for Network Anomaly Detection System

For experiment 2, the distributed algorithm was design as well as centralized algorithm for anomaly detection by writing a code in JAVA language using eclipse. This is to extend the experiment 1 by design a distributed AODE algorithm only to overcome the problem of centralization (a fatal destruction if the system malfunction occurs). The package of ML algorithms from WEKA imported to eclipse. Begin the experiment by load the dataset and class path of network labelled dataset. Then, the network size or number of nodes initiated. The number of nodes set to 20, 40, 60, 80, and 100 nodes. In the context of ML approaches, training and testing is compulsory to be applied. For distributed algorithm, the decision making by randomly select available node in the network system to aggerate the collected result to be measured the prediction result.

## 4. RESULTS

In this section, we evaluate the results obtained from experiment 1 and experiment 2 performed to test the effectiveness (accuracy) and efficiency (processing time) of distributed AODE-based anomaly detection system by comparing with several supervised ML algorithms. The result of accuracy computed based on the following formula in (1) for evaluating intelligent algorithms:

$$Accuracy\ (\%) = \frac{TP+TN}{N}\ X\ 100\% \qquad (1)$$

where; TP=Instances correctly predicted as attacks
TN=Instances correctly predicted as normal
N=Total number of instances that equal to 257,673 instances

### 4.1. Classification Rate (Accuracy)

Figure 1 showed that J48 trees is the highest percent of accuracy with 98.71% for anomaly detection over UNSW-NB15 dataset followed by AODE algorithm that small different only 1.45 percent. The percentage of accuracy of AODE is 97.26%. Among the chosen algorithms used in this paper, classification rate of NB algorithm is the worst where only 76.12%. Whilst, the features of given dataset are dependent one to another feature and this not fulfilled the assumption of NB algorithm where the features are independent. Different applications domains have different viewpoint about anomalies and this cause the different ML algorithm might well suited for anomaly detection. Accordingly, AODE is higher in accuracy compare to NB algorithm where it is alleviated the independent features of network data by assume the features dependency and comprehend the averaging the classifiers that monitor the network traffic [20]. Another two algorithms with accuracy 84.41% and 89.76% for RBFN and MLP respectively.

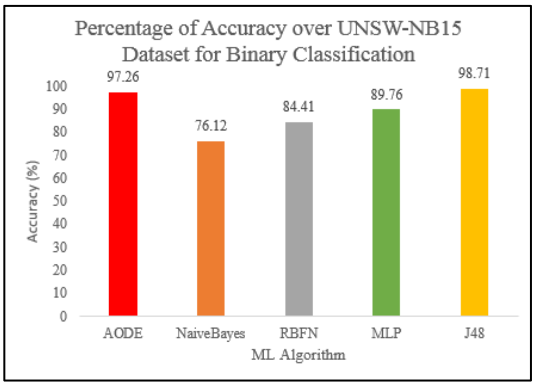

Figure 1. Percentage of accuracy over UNSW-NB15 dataset

### 4.2. Time

Time is very important and necessary when dealing a build network anomaly detection system and Table 3 is tabulated the processing time to build ML algorithms for classification purpose. From the Table 3, the fast algorithm to classify the data instances is NB that takes only about 1.22 seconds due to their virtue independency features. Yet, this algorithm is worst effective to detect the data either normal or anomalous data as stated in previous section. Meanwhile, the AODE algorithm recorded that the time taken for them to classify the data instances is approximately 7 seconds which is consider a fast processing. Although, it is dependent features still give a short time to determine the class of data instances of the given dataset. The system that takes longer time that might possibly take hours or maybe days (MLP that takes more than 7 hours) to finish the classification cause the wastage of man power to wait in order to know the prediction result. Also, this cause the prediction might be failed to be collected if there is an unexpected disaster such as the computer shut down suddenly.

Table 3. Processing time to build ML algorithms for classification

| ML algorithm | Time Taken |
|---|---|
| AODE | 6.98 seconds |
| Naïve Bayes | 1.22 seconds |
| RBFN | 7.50 seconds |
| MLP | 7 hours and 58 minutes and 38.72 seconds |
| J48 | 31.95 seconds |

### 4.3. Comparison

In the final experiment of this paper is solve the issue centralization by designing a distributed algorithm and this case we only built a distributed AODE algorithm which is an effective and efficient for anomaly detection that proved in experiment 1. Even though, the number of nodes varies there is no changes of percentage of accuracy for centralized algorithm due to there is only one node (act as server or center site) that make a prediction. As in followed bar chart the increase number of nodes the accuracy is similar with 97.26%. Differ for distributed algorithm, where as can be seen in Figure 2 the different values of percentage accuracy when different number of nodes is set in network system for anomaly detection of network labelled UNSW-NB15 dataset. Although, the results of distributed algorithm degrade the classification rate, but it still considers high where the recorded result shown that the accuracy for 20, 40, 60, 80, and 100 nodes in the range above 95% to 96% (96.59%, 96.30%, 96.15%, 95.98%, 95.86% respectively).

In term of time taken, suppose the distributed algorithm fast to classify the data instances because whole the present nodes in the network have a same level capability and share their prediction data to make a final prediction. But, in this conducted experiment the time needed for distributed AODE algorithm showed an increasingly when the number of nodes large. This is because the design of our distributed AODE algorithm is in batch manner where the collected data group (during training stage) first before a prediction made and the data instances of network labelled UNSW-NB15 dataset is a streaming data. Therefore, the future work needs to be done by design a distributed online algorithm for network anomaly detection to improve the performance (efficiency). Figure 3 revealed that with the large-scale network (scalability issue) the shorter time is produced using a centralized AODE algorithm for network anomaly detection system.

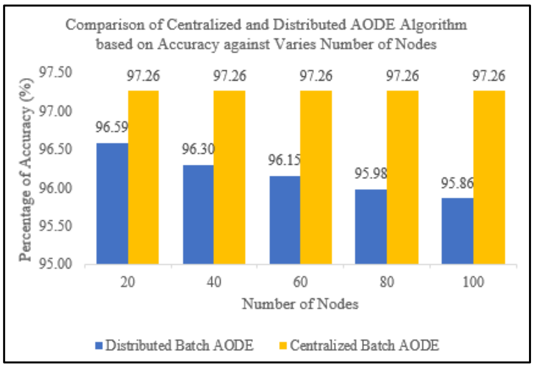

Figure 2. Comparison of centralized and distributed aode algorithm based on accuracy against number of nodes

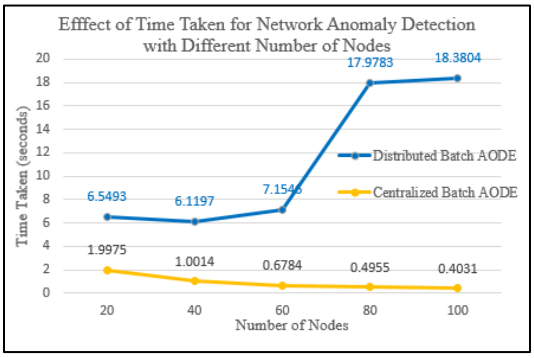

Figure 3. Effect of processing time for network anomaly detection with different number of nodes

## 5. CONCLUSION AND FUTURE WORKS

As a conclusion, machine learning approach is more effective, efficient, relevant, and best performing for classification that determine either the data instances are normal or anomalous data within any applications or environments. AODE algorithm is the most outperformed based on effectiveness and efficiency compared to another four ML algorithms that been used in this present work when classify the UNSW-NB15 datasets. Moreover, this proved that distributed AODE algorithm overcome the issue of centralization when the finding showed the considerably result although a little drop of accuracy and a bit longer time needed. Since, the performance of distributed batch ML algorithm is surprisingly taking longer. Therefore, the future work is to improve it by designing a distributed algorithm using online learning instead of batch learning that take time during training stage.

## ACKNOWLEDGEMENTS

The research reported in this paper is supported by Research Acculturation Grant Scheme (RAGS), Grant Number: 9018 00080. The authors would also like to express gratitude to the Malaysian Ministry of Higher Education (MOHE) and University of Malaysia Perlis (UniMAP) for the sponsor, financial support, and facilities provided.

## REFERENCES

[1] M. Ahmed, A. Naser Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” *J. Netw. Comput. Appl.*, vol. 60, pp.19–31, 2016.
[2] V. Chandola, “Anomaly Detection: A Survey,” no. September, pp. 1–72, 2009.
[3] Engen, Vegard. "Machine learning for network-based intrusion detection: an investigation into discrepancies in findings with the KDD cup'99 data set and multi-objective evolution of neural network classifier ensembles from imbalanced data." PhD diss., Bournemouth University, 2010.
[4] S. Ben-David and S. Shalev-Shwartz, Understanding Machine Learning: From Theory to Algorithms. 2014.
[5] T. Ahmed, B. Oreshkin, and M. Coates, “Machine Learning Approaches to Network Anomaly Detection,” Proc. 2nd USENIX Work. Tackling Comput. Syst. Probl. with Mach. Learn. Tech. USENIX Assoc., pp. 1–15, 2013.
[6] “2017 Global State of Information Security Survey – IDG Enterprise.”
[7] M. Nawir, A. Amir, N. Yaakob and O. B. Lynn, "Internet of Things (IoT): Taxonomy of security attacks," *2016 3rd International Conference on Electronic Design (ICED)*, Phuket, 2016, pp. 321-326.
[8] P. Singhal, R. Mathur, and H. Vyas, “*State of the Art Review of Network Traffic Classification based on Machine Learning Approach*,” IJCA Proc. Int. Conf. Recent Trends Eng. Technol. 2013, vol. ICRTET, no. 1, pp. 12–16, 2013.
[9] K. S. Balagani, V. V. Phoha and G. K. Kuchimanchi, "A Divergence-measure Based Classification Method for Detecting Anomalies in Network Traffic," *2007 IEEE International Conference on Networking, Sensing and Control*, London, 2007, pp. 374-379.
[10] M. Kruczkowski and E. Niewiadomska-Szynkiewicz, “Comparative study of supervised learning methods for malware analysis,” *J.Telecommun. Inf. Technol.*, vol. 2014, no.4, pp. 24–33, 2014.
[11] G. Pachauri and S. Sharma, “*Anomaly Detection in Medical Wireless Sensor Networks using Machine Learning Algorithms*,” Procedia Comput. Sci., vol. 70, pp. 325–333, 2015.
[12] J. Murphree, "Machine learning anomaly detection in large systems," *2016 IEEE AUTOTESTCON*, Anaheim, CA, 2016, pp. 1-9.

[13] J. Liu, Jun Liu, Shuyu Chen, Zhen Zhou, and Tianshu Wu, “An Anomaly Detection Algorithm of Cloud Platform Based on Self-Organizing Maps,” *Mathematical Problems in Engineering*, vol. 2016, Article ID 3570305, 9 pages, 2016.

[14] A. Gulenko, M. Wallschläger, F. Schmidt, O. Kao and F. Liu, "Evaluating machine learning algorithms for anomaly detection in clouds," *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2016, pp. 2716-2721.

[15] Reddy Jidiga, Goverdhan & Sammulal, P & Publication, IAEME. (2013). Machine Learning Approach To Anomaly Detection In Cyber Security With A Case Study Of Spamming Attack. *International Journal of Computer Engineering and Technology*. 0976-6375. 4. 113-122.

[16] N. Moustafa, J. Slay, *UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set).* 2015 Military Communications and Information Systems Conference (MilCIS), pp. 16, 2015. https://doi.org/10.1109/MilCIS.2015.7348942.

[17] N. Moustafa, J. Slay,The UNSW-NB15 data set description, 2 March 2016, Accessed date: sept, 2017 from https://www.unsw.adfa.edu.au/australian-centre-for-cybersecurity/ cybersecurity/ADFA-NB15-Datasets/.

[18] R. R. Bouckaert et al., “WEKA Manual for Version 3-7-8,” pp.1–327, 2013.

[19] D. Bhattacharyya and J. Kalita, Network anomaly detection: A machine learning perspective. 2013.

[20] Z. A. Baig, A. S. Shaheen and R. AbdelAal, "An AODE-based intrusion detection system for computer networks," *2011 World Congress on Internet Security (WorldCIS-2011)*, London, 2011, pp. 28-35.

## BIOGRAPHIES OF AUTHORS

Mukrimah Nawir received B.Eng (Hons) degrees in Computer Engineering from University Malaysia Perlis (UniMAP), Perlis, Malaysia in 2015. Currently, she is studet Master of Sciences (by research) in Computer Engineering. Her research on distributed classification and machine learning.

Amiza Amir is a senior lecturer at School of Computer and Communication Engineering at Universiti Malaysia Perlis (UniMAP). She received her Ph.D in Information Technology on Distributed Artificial Intelligence, from Monash University, Australia in 2015. Her current research interest includes machine learning, distributed system, meta heuristic optimization, and software-defined network (SDN). She teaches courses in Data Analytics and Artificial intelligent.

Naimah Yaakob is a senior lecturer at School of Computer and Communication Engineering, University Malaysia Perlis, Perlis, Malaysia (UniMAP). She received B.Sc. and M.Sc. degrees in Computer and Information Engineering from the International Islamic University (IIUM), Kuala Lumpur. Malaysia in 2004 and 2008 respectively. She also received the Ph.D. degree in 2014 from the School of Computer Science and Information Technology, RMIT University, Melbourne, Austalia. Her research interests include Wireless Sensor Networks, Wireless Body Sensor Networks, Vehicular Ad-Hoc Network, congestion control in distributed networking systems, and pervasive communication in healthcare environment.

Ong Bi Lynn is a senior lecturer at School of Computer and Communication Engineering, University Malaysia Perlis, Perlis, Malaysia (UniMAP). She received B.Eng (Hons) Electrical & Electronics, Universiti Malaysia Sabah (UMS, 2001). Also, received her Master in Business Administration (MBA) in 2003 and Ph.D. in Computer Network (2008) at Universiti Utara Malaysia.