International Journal of Internet, Broadcasting and Communication Vol.13 No.3 42-49 (2021)
http://dx.doi.org/10.7236/IJIBC.2021.13.3.42

IJIBC 21-3-6

A Study on Security Event Detection in ESM Using Big Data and Deep Learning

Hye-Min Lee *, Sang-Joon Lee **

**MS student., Interdisciplinary Program of Information Security, Chonnam National University, Korea*
E-mail: leehyemin14@naver.com
***Professor., Interdisciplinary Program of Digital Future Convergence Service, Chonnam National University, Korea*
E-mail: s-lee@chonnam.ac.kr

Abstract

As cyber attacks become more intelligent, there is difficulty in detecting advanced attacks in various fields such as industry, defense, and medical care. IPS (Intrusion Prevention System), etc., but the need for centralized integrated management of each security system is increasing. In this paper, we collect big data for intrusion detection and build an intrusion detection platform using deep learning and CNN (Convolutional Neural Networks). In this paper, we design an intelligent big data platform that collects data by observing and analyzing user visit logs and linking with big data.

We want to collect big data for intrusion detection and build an intrusion detection platform based on CNN model. In this study, we evaluated the performance of the Intrusion Detection System (IDS) using the KDD99 dataset developed by DARPA in 1998, and the actual attack categories were tested with KDD99's DoS, U2R, and R2L using four probing methods.

Keywords: *Intrusion Detection Systems, Enterprise Security Management, Convolutional Neural Network Intrusion Prevention System.*

1. Introduction

Currently, due to development of network technology using 5G communication network along with IT, internet of things and cloud computing technology have appeared and entering a hyper-connected society and continuous exposure to personal information leakage, cyber terrorism and cybercrimes it is becoming [1]. To protect this, various security systems such as intrusion detection system (IDS), virtual private network (VPN), and intrusion prevention system (IPS) have appeared along with firewalls, but the need for centralized and integrated management of each security system it is a recognized situation. In this case, the proposed method

Manuscript Received: May. 26, 2021 / Revised: June. 2, 2021 / Accepted: June. 7, 2021
Corresponding Author: s-lee@jnu.ac.kr
Tel: +82-62-530-1447, Fax: +82-62-530-1449
Professor., Interdisciplinary Program of Digital Future Convergence Service, Chonnam National University, Korea

is the Integrated Management System (IMS) [2].

Several works studied about IDS [3-4] propose IDS based on Software Defined Network (SDN). The proposed IDS detects DDoS attacks and informs to SDN controller [5]. study about experimental performance of Snort-based IDS (S-IDS) in network [6]. suggest Distributed Denial-of-Service (DRDoS) detection and defense model based on Deep Forest model (DDDF). In particular, they focus on attacks in Internet of Things (IoT) devices and big data environment. In addition, there are several studies about anomaly detection schemes for Industrial Wireless Sensor Networks (IWSNs) based on machine learning [7-8]. suggest Hierarchical Intrusion Detection System (HIDS) based on statistical preprocessing and NN classification [9]. show that Hidden Naive Bayes (HNB), which is one of the data mining models, can be used in IDS [10]. suggest analysis about threat of IoT based on Artificial Neural Network (ANN) to detect DoS/DDoS attacks.

Enterprise System Management(ESM) is composed of IPS, intrusion detection and prevention systems (IDPS), virus blocking system, and vulnerability diagnosis system to enable comprehensive intrusion response. The advantage of ESM is that it can reduce redundant investment and waste of resources because companies do not use various types of security solutions, and can establish security strategies by integrating information and communication systems through mutual communication between solutions [2].

This paper aims to collect big data for intrusion detection and detect intrusion using convolutional neural networks (CNN). The final purpose of the thesis is to collect relevant data through observation and analysis of the user's visit log. In order to observe and detect the communication data of illegal intruders, intuitive UI is designed to automatically collect information to enable real-time feedback, and to synthesize and analyze data collected through repeaters and can use a structured database in the form of big data.

2. Design of intrusion detection system

In this paper, the KDD99 data set developed by DARPA [11] is used to evaluate the performance of IDS [3], and the actual injection attack for each category follows the five types of KDD99 shown in Table 1. In addition, the training and test sets are used independently of each other because for each experiment we split the training and test samples exclusively 90:10.

The KDD99 data set is extensively used in machine learning research (MLR) and IDS. The KDD data set created a military network environment for the Air Force by the MIT lincoln lab and then generated various attacks and TCP/IP data to simulate [12].

As shown in Table 1, each record of data has 41 network parameters, and all data belongs to one of five types of attacks (Normal, DoS, U2R, R2L, Probeing). It consists of 494,021 connection records as described in Table 1. It can be easily calculated that the Normal, DoS, PRB, R2L, and U2R connection ratios on the dataset are 19.69%, 79.24%, 0.831341%, 0.23%, and 0.01% on the KDD Cup 1999 dataset is 10%.

Table 1. 10% version of the KDD Cup 1999 dataset distributions Classification and proportions

No	Classification	Quantity	proportions
1	NORMAL	97,278	19.69%
2	DoS	391,458	79.23%
3	U2R	52	0.01%
4	R2L	1,126	0.22%
5	PROBE	4,107	0.83%
	Total	494,021	100%

As shown in Table 2, DoS represents most of the dataset and the rest of the categories represents other general connections representing less than 1% of the training dataset. Therefore, neural network models over-training on two types of data that take a very long time to repeat learning, at the same time, we treat trivial types as noise due to negligible proportions in our training dataset. To overcome this problem, we are sampled 10% versions of the KDD Cup 1999 dataset with comparable values in various categories. The proposed system is trained using both the original and sample datasets.

Table 2. 5 types of attack and 10% version of the KDD99

Classification	Category	Quantity	proportions	Classification	Category	Quantity	proportions
Normal	Normal	97,278	19.69%	R2L	ftp_write	8	0.001%
DOS	Back	2,203	0.44%		Guess_passwd	53	0.01%
	Land	21	0.004%		imap	12	0.002%
	Neptune	107,201	21.69%		multihop	7	0.001%
	Pod	264	0.05%		phf	4	0.0008%
	Smurf	280,790	56.83%		spy	2	0.0004%
	Teardrop	979	0.19%		warezclient	1,020	0.2%
Probing	Ipsweep	1,247	0.25%	U2L	warezmaster	20	0.004%
	Nmap	231	0.04%		Buffer_overflow	30	0.006%
	Portsweep	1,040	0.21%		loadmodule	9	0.001%
	Satan	1,589	0.32%		perl	3	0.0006%
					rootkit	10	0.002%

In Figure 1, we tested with only three high-value records: normal, Neptune, and Smurf model.

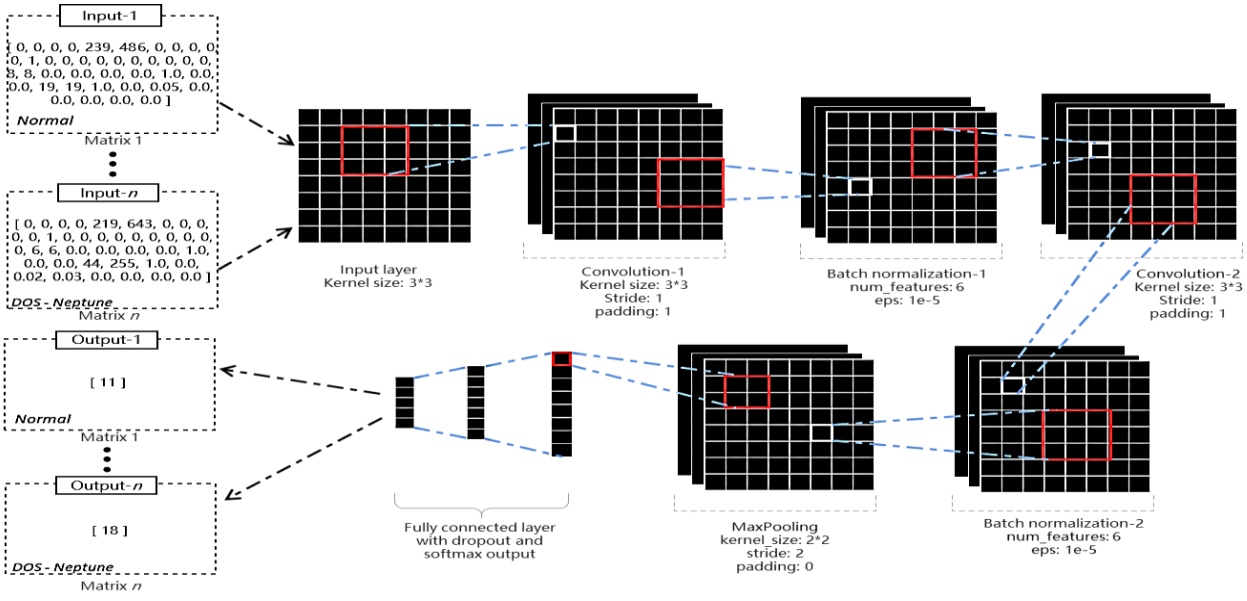

Figure 1. Structure of a mixed normal, Neptune and Smurf model

3. Structure of Intrusion Detection System

As shown in Figure 2, the system of this paper is written as four functional modules: big data collection and storage, data preprocessing and request, API module, and data visualization. The big data collection and storage module in Figure 2 fetches and analyzes the user's behavior data from the warehouse and makes it easy to call.

Transmits the processed data in Json Format. The CNN-based intrusion detection module analyzes the data, and when the analysis is completed, it assists the network security manager through the data visualization function.

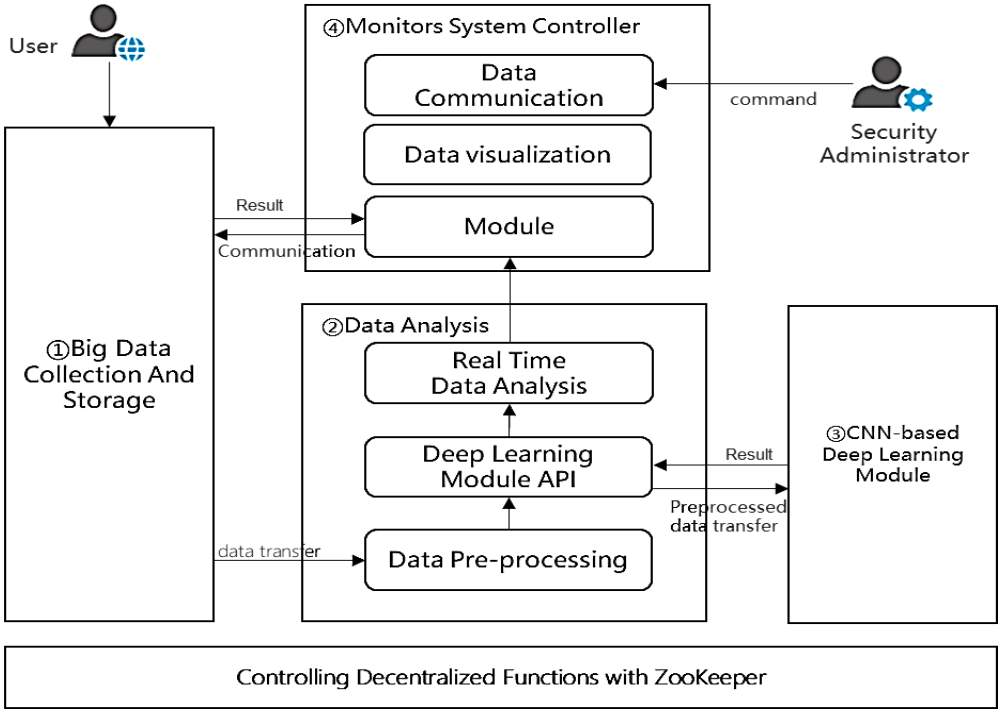

Figure 2. Structure of Intrusion Detection System

3.1 CNN-based intrusion detection module

After calling the API of the Deep Learning Module (DLM) for the analyzed data, first extract the data in the package and then encode it into numbers. To make it easier to input into the CNN neural network, the encoded data is arranged in an 8*8 matrix, and insufficient data is filled with zeros. The processed data matrix is input to the CNN Module of Figure 3 and analyzed. Finally, request module analysis data decoding.

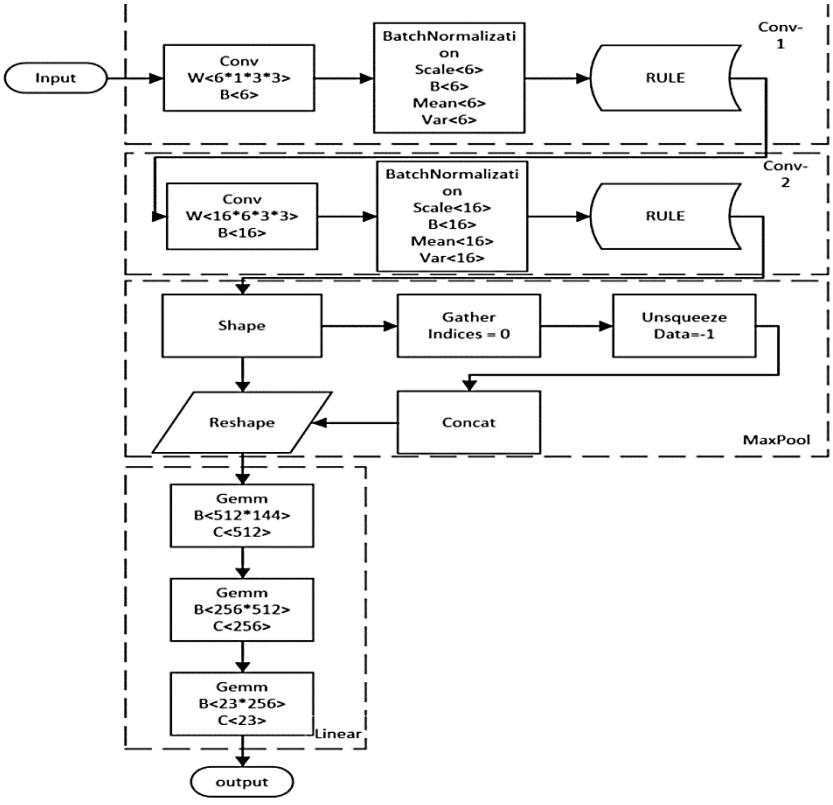

Figure 3. CNN model process

4. Implement and Result

The Data Storage and Data Pre-processing process as follows. First, log accessory's behavior using Log 4j, a data logging tool. The logged value indicates the log format after being stored in HDFS, and is stored in the database as log storage time, Network Input Output NIO, visit destination, USER ID, visitor name, and access time. Extract the log stored in HDFS, as shown in Table 3, mobile phone number, usage amount, visitor IP address, number of visits today, number of past visits, visitor name, version, visit method (Windows Phone (WP), Android, IOS, WEB Site) in the form of save in HDFS.

Table 3. Extract required data from log data

Phone Number	Amount Consumed	IP Address	Number of Visits Today	Historical visits	Visitor Name	Average Visit Time	Version	Visit Method
069125395	0	171.9.161.17	7	71	User5	5404	1.1.2	WP
069125400	0	171.14.10.110	8	81	User5	5984	1.1.2	IOS
069125405	3000	36.59.72.57	8	87	User7	5258	1.1.0	WP
069125410	7000	61.237.79.233	3	33	User3	5304	1.1.3	Android
069125415	2000	171.11.146.63	6	60	User6	5001	1.1.1	WP

In this paper, the recorded result values are shown as shown in Table 4 and Figure 4 using 'wandb', an API that records loss and accuracy during neural network training. When the loss and accuracy are checked for

each step during neural network training, one-steps were measured to be about2.838. 20steps has a record of 1.069, 50 steps have a record of 0.819,100 steps has a record of 0.315, 200 steps have a record of 0.1555, 300 steps have a record of 0.127and 1000 steps has a record of about 0.069. The good robustness of the network appears as the loss value changes according to training and testing. In the case of validation accuracy, step 1 continuously rises within 0.571, step 20 is 0.575, step 50 is 0.612, 100 steps are 0.9137, 200 steps are 0.970, 300 steps are 0.92, and finally, stable values show high accuracy at around 0.985 at 1000 step.

Table 4. Results of Learning

Division	Train loss	Val loss	Train Accuracy	Val Accuracy
1 step	2.838	3.012	0.472	0.571
20 step	1.069	1.076	0.578	0.575
50 step	0.819	0.844	0.618	0.612
100 step	0.315	0.329	0.915	0.913
200 step	0.155	0.172	0.971	0.970
300 step	0.127	0.130	0.983	0.982
1000 step	0.065	0.066	0.985	0.985

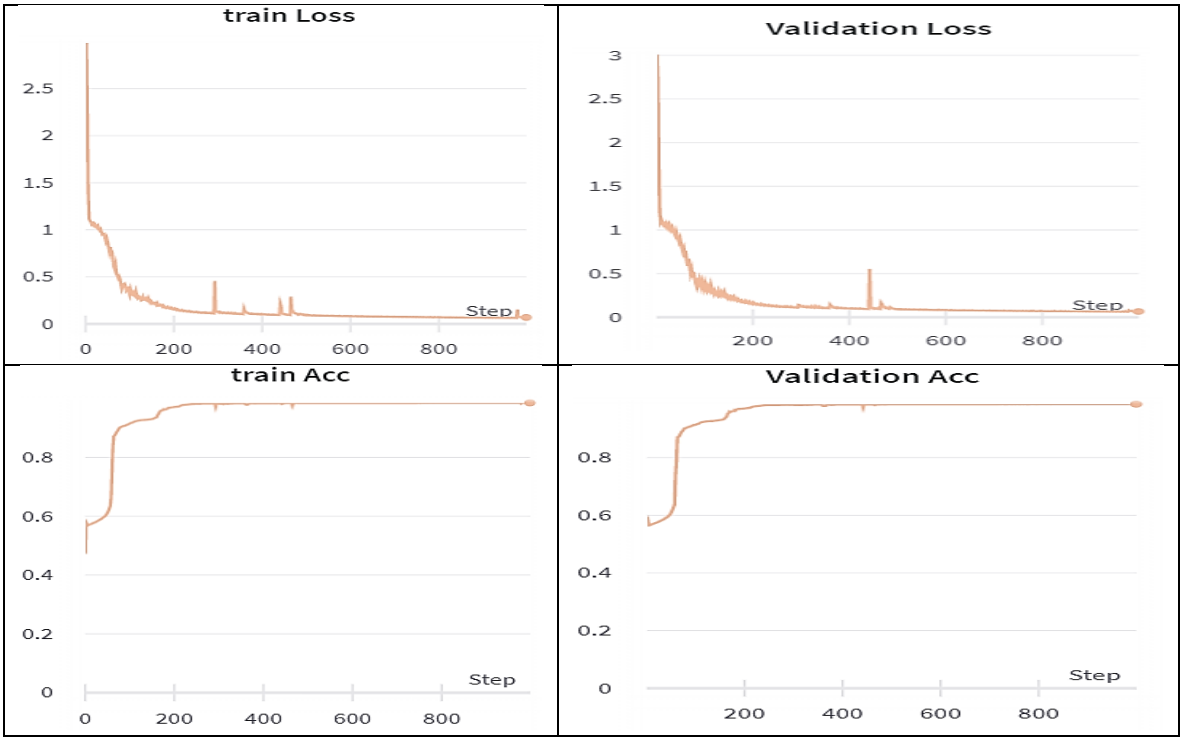

Figure 4. Graph of Learning Results

In this paper, a unit test for the intrusion detection module function is performed. As in Table 5, depending on the result, the module can perform an accurate test on one test data with a delay time of 0.02~0.04s. It can be seen that this shows the high processing speed of the system.

Table 5. Unit test result for intrusion detection module function

No	Real_label	Predict_label	Test_Data	Accuracy	Time(sec)
1	Normal	Normal	[0 b'tcp' b'http' b'SF' 239 486 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 8 8 0.0 0.0 0.0 0.0 1.0 0.0 0.0 19 19 1.0 0.0 0.05 0.0 0.0 0.0 0.0 0.0]	97.5%	0.0278
2	DOS	Surmf	[0 b'icmp' b'ecr_i' b'SF' 1032 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 511 511 0.0 0.0 0.0 0.0 1.0 0.0 0.0 255 255 1.0 0.0 1.0 0.0 0.0 0.0 0.0 0.0]	93.4%	0.0406
3	DOS	Neptune	[0 b'tcp' b'telnet' b'RSTO' 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 239 6 0.0 0.0 1.0 1.0 0.03 0.07 0.0 255 6 0.02 0.07 0.0 0.0 0.0 0.0 1.0 1.0]	83.7%	0.0482

As shown in Table 4 and Figure 4, the CNN model used in this paper has fast convergence and good robustness on KD99 dataset. As shown in Table 5, according to the result of unit test on intrusion detection module function, we can see that the system has a faster processing speed while ensuring high accuracy. The system design described above is suitable for ESM development and design.

5. Conclusion

ESM consists of IPS, IDPS, Virus Blocking System (VBS), and Vulnerability Diagnosis System (VDS) to enable comprehensive intrusion response.

This paper aims to collect big data through observation and analysis of user's visit log for EMS intrusion detection and use CNN to detect intrusion. After encoding, CNN arranges the data in an 8*8 matrix. At this time, on average, there are 61 packet data features, so the missing three are supplemented with zero. After processing, the data matrix was analyzed using CNN.

In this study, CNN and softmax function were applied to multiple classification in the intrusion detection field. In the case of the kdd-99 dataset, the Rectified Linear Unit activation function (ReLU) was used for optimization and the accuracy of intrusion detection was effectively improved.

As a result of checking the loss and accuracy for each step during neural network training as an experiment in this paper, 20 steps were measured to be about 0.181, 100 steps were 0.128 and 1000 steps were about 0.065. As for the accuracy of the train value, step 1 continuously increased within 0.6%, step 20 showed 0.951%, 100 steps were 0.977%, and finally, 1000 step showed high accuracy to 0.985%.

Acknowledgement

This work was supported by an Institute of Information & communications Technology Planning & evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-01343).

References

[1] J. K. Bae, “A Study on the Establishment of Enterprise Security Management System Based on Artificial Intelligence and BigData Analysis,” Logos Management Review, Val.18, No.1, pp.151-166, 2020.

[2] Ensxoddl, Tistory. Submission of manuscript. https://ensxoddl.tistory.com/193.
[3] Anwar, Shahid, Mohamad Zain, Jasni and Zolkipli, Mohamad and Inayat, Zakira and Khan, Suleman and Anthony Jnr, Bokolo and Chang, Victor. "From intrusion detection to an intrusion response system: Fundamentals, requirements, and future directions" Algorithms, Val. 10, No. 30, 2017. Doi:10.3390/a10020039
[4] Jing-xin, Wang, Zhi-ying, Wang and Kui, Dai, "A Network Intrusion Detection System Based on the Artificial Neural Networks," Association for Computing machinery, Val.04, No.5, pp.166-170, 2004. https://doi.org/10.1145/1046290.1046324
[5] Manso. P, Moura. J and Serrão. C, "SDN-Based Intrusion Detection System for Early Detection and Mitigation of DDoS Attacks. Information," Vol. 10, No. 3, 106, 2019. https://doi.org/10.3390/info10030106
[6] Karim. I, Vien. Q. -T, Le. T. A and Mapp G. A, "Comparative Experimental Design and Performance Analysis of Snort-Based Intrusion Detection System in Practical Computer Networks," Computers. Vol. 6, No. 1, 6, 2017. https://doi.org/10.3390/computers6010006
[7] R. Xu, J. Cheng, F. Wang, X. Tang and J. Xu, "A DRDoS Detection and Defense Method Based on Deep Forest in the Big Data Environment," Symmetry, vol.11, No.1, pp.78, 2019. https://doi.org/10.3390/sym11010078
[8] Ramotsoela, Daniel and Abu-Mahfouz, Adnan and Hancke, Gerhard, "A Survey of Anomaly Detection in Industrial Wireless Sensor Networks with Critical Water System Infrastructure as a Case Study," Sensors, Vol. 18, No. 8, pp.2491, 2018. https://doi.org/10.3390/s18082491
[9] Zheng Zhang, Jun Li, C. N. Manikopoulos, Jay Jorgenson, Jose Ucles, "HIDE: A hierarchical network intrusion detection system using statistical preprocessing and neural network classification," IEEE Workshop on Information Assurance and Security, pp.5-6, 2001.6
[10] Levent Koc, Thomas A. Mazzuchi, Shahram Sarkani, "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier," Expert Systems with Applications, Val.39, No.18, pp.13492-13500, 2012. https://doi.org/10.1016/j.eswa.2012.07.009
[11] kdd.ics.uci, Submission of manuscript. https://kdd.ics.uci.edu/databases/kddcup99/task.html
[12] Paliwal, S. Gupta, R, "Denial-of-Service, Probing & Remote to User (R2L) Attack Detection using Genetic Algorithm," International Journal of Computer Applications, Val.60, pp.57-62, 2012.
[13] W. J. Kang, " An extended Access Control with Uncertain Context," International Journal of Computer Applications, International Journal of Advanced Smart Convergence, Vol. 7, No.4, pp. 66-74, 2018.