Information Security Technology for Computer Networks through Classification of Cyber-Attacks using Soft Computing Algorithms

Jason A. Villaluna [1] and Febus Reidj G. Cruz [2]
1 School of Graduate Studies, Mapua University, Intramuros, Manila, Philippines
2 School of Electrical Electronics and Computer Engineering, Mapua University, Intramuros, Manila, Philippines

***Abstract*—The Internet is the global platform which revolutionized the computer and communications domain. Although it becomes one of the most useful tools in people's lives, the presence of cyber-attacks that can cause damage, modification, and theft of vital data and information over this platform has increased. Utilization of soft-computing based on the behavior of the network may detect new or modified old attacks. An information security system is developed for the recognition the network infrastructure's behavior. This is limited to Normal, DoS, Probe, U2R, and R2L. The packets on the network are processed in MATLAB and analyze using Fuzzy Logic, Artificial Neural Network, and Fuzzy-Neural Network. Different tests are done with different datasets of varied parameters. The best model for each algorithm, which is rendered from the tests, is used for the information security system. The cyber-attacks were identified within a short period: 51.64us for Fuzzy Logic, 1.34us for Artificial Neural Network, and 14.23us for the Fuzzy Neural Network. The detection rate and accuracy of the three algorithms are 94.84%, 98.51%, 98.60% and 89.74%, 96.09%, 96.19% respectively. The Fuzzy Neural Network has the best performance which used the advantage of Fuzzy Logic and Artificial Neural Network.**

Keywords—Aritificial neural network; cyber-attacks; internet; fuzzy logic; fuzzy neural network

I. INTRODUCTION

Information dissemination, interaction, and collaboration between individuals and computers without considering the geographic location is a major challenge. With the rise of technology and the Internet, a global platform which revolutionized the computer and communications domain, this issue is already addressed. The Internet becomes a worldwide tool that has an enormous impact on education, health, wealth, government, and business. According to Statista, there are approximately 3.5 billion Internet users with 7.5 percent growth rate which is 47.3 percent of the world population as of 2016. This large growth rate of Internet users is associated with the fast implementation of Internet of Things-a system of interconnected computing devices, machines, sensors, microelectromechanical systems, and other electronic objects that have the capability to transfer data over a network without requiring human interaction. This system is running in the environment either at offices, homes, hospitals, schools, universities, banks, etch. According to Business Wire as of 2014, most technology and services revenue connected to it will grow at a rate of 8.8 percent compounded annually from 2012-2017. Indeed, it makes the way of living of everyone for faster, efficient, and productive way.

Although Internet has become one of the most beneficial tools in people's lives, the presence of threats in the form of cyber-attacks that can cause damage, modification, and theft of vital data and information over this platform has increased. Such cyber-attacks are Denial of Service (DoS), making memory and resources of the network or computer too demanding to disrupt the normal functions; Probing, gaining access to the network or computer and collect information or find known vulnerabilities; User-to-Root (U2R), exploiting vulnerabilities to the local user and gaining access to super-user (root) privileges; Remote-to-Local (R2L), accessing a remote machine to gain access to the system by password guessing; and others [1]. Due to the possibility of deploying these on the network, the operation of smart grids, smart homes, smart vehicles, cloud, and other components of Internet-of-Things will be compromised. This becomes detrimental since it will not only harm the infrastructure at large but will eventually have a ripple effect on the individual's security. Therefore, protecting the confidentiality, integrity, and availability (CIA) of vital data and information from the other users connected to the Internet is the main challenge. Failure to protect the CIA of data and information can lead to disclosure to unauthorized individuals. Manipulation, modification, and unavailability of the data and information will be possible. It may lead to fraud, identity theft, sabotage, potential loss of privacy, data, and money, and other higher form of crimes.

Since the reliance on the stringent rules set is not sufficient to detect attacks, the development of information security system that would competently detect and identify the cyber-attacks over the network infrastructure can help to counteract these and prevent the further effects. Utilization of soft computing based on the behavior of the network may detect new or modified old attacks. Moreover, it is efficient in terms of speed for providing identification of the detected attacks. To ensure the security, most network engineers, information

978-1-5386-0912-5/17/$31.00 © 2017 IEEE

security engineers, and network security engineers use many systems which monitor and analyze the activity of the network. Firewall and Intrusion Detection Systems (IDS) are network systems used for securely monitoring the network from the attacks and abnormal behavior of it. Some approaches have been implemented such as set of policies for the Firewall, rule-based systems and soft computing for the IDS. There is a possible approach implied by the study "Anomaly Network Intrusion Detection System Based on Distributed Time-Delay Neural Network". Soft computing e.g., Fuzzy Logic, Artificial Neural Network, Probabilistic Reasoning, and Genetic Algorithms are set of processing and optimization techniques which are lenient to imprecision and uncertainty (Ibrahim 2010). Application of soft computing in analyzing the Internet traffic would allow the determination of abnormalities in network activity. This can be developed by creating a system that will be used to detect and identify the cyber-attacks on the Internet. The unsupervised classification of cyber-attacks from analysis of Internet packets is an active subject of the study. Black hat hackers are continuously propagating new attacks and modifying old attacks over the Internet. Novel attacks cannot be detected by systems based on rules, policies, and signatures. With this incorrect assessment of the network activity, unrealized attacks can occur. Real-time detection and identification of old and novel attacks are essential to provide fast counter measure and prevent further effects on the network infrastructure. However, many algorithms have not addressed their reliability with the amount of time to respond real-time, low detection rate, and high false alarm. A system which can perform real-time, high detection rate, and low false alarm is necessary to effectively detect and counter the attacks. The proponent of the study sought answers to the following questions: The study aims to answer the question of how to develop a soft computing-based information security system that would identify cyber-attacks over the Internet. Specifically, it seeks to answer the questions of how to develop a system that would detect cyber-attacks which can perform real-time, high detection rate, and low false alarm; how to utilize the soft computing algorithms in packet analysis to classify the cyber-attacks; and how to recognize the best algorithm for the information security system among Fuzzy Logic, Artificial Neural Network, and Fuzzy-Neural Network.

This study aims to develop a soft computing-based information security system using packet analysis. Also, this study aims to meet the following goals needed for the development of a cyber-attack detection and identification program: to detect cyber-attacks in a short period, with high detection rate and low false alarm; to identify different cyber-attacks such as DOS, Probe, U2R, and R2L; and to compare Fuzzy Logic, Artificial Neural Network, and Fuzzy-Neural Network algorithms for the information security system.

With the rise of the said attacks, the completion of the study will undeniably help different sectors which are using the Internet as their major platform. Such beneficiaries are businesses, companies, banks, hospitals, houses, offices, etc. It may also help to protect the privacy and identity of the general population that may become a victim of the cyber-attacks. It is an immensely advantageous development for the world's dynamic technology.

The study is focused on the detection and identification of cyber-attacks over the Internet. The packet analysis is focused on the recognition of behavior of the network infrastructure which is limited to Normal, DoS, Probe, U2R, and R2L. The KDD99 and KDD-NSL Dataset from the Third International Knowledge Discovery and Data Mining Tools Competition held by DARPA is used. The development of information security system will concentrate on the network layer of the Open System Interconnection (OSI) Model, a framework which standardized the communication functions of the computing and telecommunication system without regard to internal underlying technology for interoperability. The packets on the network is processed in MATLAB to analyze using the Fuzzy Logic, Artificial Neural Network, and Fuzzy-Neural Network algorithms. Fuzzy logic has more tolerance for imprecision of data; Neural networks have more tolerance for noise. The Fuzzy-Neural Network is intended to use the advantage of both fuzzy logic and neural networks. It can approximate any nonlinear function to any prescribed accuracy. It is a machine learning algorithm to maximize the detection accuracy and minimize the complexity of computation in real time. Its performance is experimentally compared with other algorithms and shows better convergence speed. This confirms its applicability in learning large-sized neural networks of real-life applications [2]. The testing and validation of the results are simulated in the MATLAB.

II. Review of Related Literature

This section covers the background theories, principles, and studies useful for the development of the idea for the information security system. Also, some technical terminologies from previous and present projects developed, giving focus on cyber-attacks detection and identification.

A. Internet and Cyber- Attacks

The Internet has become a significant part of everyone's life. It is used at home, office, schools, hospitals, stores, everywhere. It's a tool to track on business, to keep updated with news, and to communicate with everyone. Advancement of life has become more pronounce, but it comes with a threat to privacy, identity, personal resources, valuable data, and information

To combat Cyber-attack, a sensitive issue in the world of Internet security because of the rise of breaches, governments and business organizations are doing its best to provide different types of tools and techniques to secure their data and private information, and to keep their business running [3].

Denial of Service (DoS) attacks the computer system or network or website by suspending temporarily or permanently the function and availability of the network by making the memory and computing resources too demanding, so that legitimate users access to these resources are denied [1].

Probing is a malicious program that accesses computer files or information about the remote victim [1].

User to Root (U2R) gains unauthorized root/administrator privileges which has local access to victim system. Some vulnerability in the victim machine are exploited by buffer overflow attacks [1].

Remote to Local (R2L) accesses a normal user account on the system by exploting some vulnerability. The attack intrudes in to a remote machine and gains local access of the victim machine [1].

B. Dataset

A standard set of data which contains different intrusions simulated in a network environment.

KDD99 Dataset is a preprocessed DARPA dataset submitted to Knowledge Discovery and Data Mining (KDD) yearly competition. It is easier to use for machine learning algorithm than the original DARPA dataset therefore most researches are using this dataset. The characteristics of KDD99 are described in [4].

NSL-KDD Dataset was introduced to lessen the deficiencies of the KDD99 Dataset. It has been produced by removing redundant and duplicate instances, also by decreasing size of dataset [4]. The size of the training and testing dataset are enough to run the experiments on the complete set without the need to randomly select a small portion. Also, evaluation results of different research work will be consistent and comparable [5]. The advantages of NSL-KDD dataset over the original KDD dataset were identified in [6].

C. Soft Computing

Soft computing e.g., Fuzzy Logic, Artificial Neural Network, Probabilistic Reasoning, and Genetic Algorithms are set of processing and optimization techniques which are lenient to imprecision and uncertainty (Ibrahim 2010). Learning algorithms used in data mining-based applications are categorized as supervised and unsupervised determine by way of learning and classifying of data. In supervised learning, classification of data is learned from labeled datasets. It can be used to create an intrusion classifier in IDS. Some of the algorithms based on supervised learning are decision tree, support vector machines, prototype-based models, distance-based models, Bayesian networks, neural networks, k-means, boosting, and bagging. While in unsupervised learning, learning is applied in unlabeled datasets to create detection model. It can be used to create an IDS operating on a core hypothesis for anomaly or outlier detection problem. Some of the algorithms based on unsupervised learning are density-based model, cluster analysis, self-organizing map, neural-networks, one-class support vector machines [7].

D. Other Algorithms

Support Vector Machine (SVM) is an algorithm used for classification and regression analysis of statistical data. It is a supervised learning which uses non-linear kernels for remodeling the training data in higher dimension space. When non-linear kernel is applied on the training data set that is linearly non-separable it is possible to create a linearly separable data set in higher dimension space. Support Vector Machine characterized the training data set to separate the groups of data by a hyper plane as wide as possible. Prediction of testing data is based on its projection in the dimension space and belong to the group on which side of hyper plane it falls [8].

Genetic Algorithm is heuristic approach used to optimize the combinatorial state using a set of parameters. It helps in simulating the population evolution process by applying crossover mutation operators with a fitness function. Set of rules are developed by calculation to get the fitness. Rules having higher fitness value will be used for another calculation. Numerous calculations go through the same procedure to produce a solution that is acceptable. This approach used to obtain classification rules for quantitative and distinct features of network data. Rules are developed in training phase by using evolutionary concept from genetic algorithm then rules are used to classify data in testing [9].

III. METHODOLOGY

This section presents the methods and procedures that were used in the development and implementation of the study. Also, the program algorithms and the design flow process in making the whole system are included.

A. Conceptualization of Design

The NSL-KDD and KDD99 Dataset which are Internet packets derived from the competition held by Defense Advanced Research Projects Agency (DARPA) served as the input of the system, as shown in Figure 1. Each packet contained features such as protocols used by the connections, login attempts, service ports, network services, etc. The Internet packets were processed using MATLAB program and undergone several phases to achieve desired data. Detection classifier classified the type of attack and alerts the system.

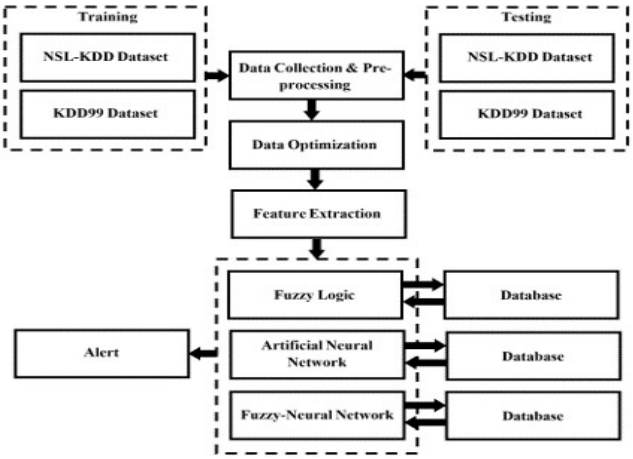

Fig. 1. System block diagram.

MATLAB, as Integrated Development Environment (IDE) for packet analysis and processing, included the normalization, elimination of redundant data, extractions of features, Fuzzy logic database, Artificial Neural Network database, and Fuzzy-Neural Network database. Fuzzy logic, Artificial Neural Network and Fuzzy-Neural Network were used as models of the system. Fuzzy logic had tolerance for imprecision of data, Neural networks had tolerance for noise, and Fuzzy-Neural Network had advantages of both models.

B. Software Development

The information security technology in Figure 2 was created and tested using MATLAB. Data were adapted to Fuzzy Logic, Artificial Neural Network, and Fuzzy-Neural Network algorithms which required normalization. Only the most relevant data was selected for further processes. Factorial Multiple Correspondence Analysis was used for data selection. It was based on the calculation of GINI indexes where values were directly proportional to the relevance of the data. Relevant attributes were established on maximum information gain with ratio over 0.6 for most of features [10].

NSL-KDD and KDD99 dataset were composed of huge files that contained redundant records of internet packets. This caused the algorithms to be biased towards the frequent data and prevented the learning of the infrequent data. Repeated record in the dataset was removed for optimum detection.

Enhancement of learning parameters was associated with different factors, such as number of epochs, number of membership functions, number of hidden layers, types of membership functions, and types of training functions. The classifier model was trained with training dataset labelled 0 for normal activity, 1 for DoS, 2 for Probing, 3 for R2L, and 4 for U2R. The last phase that followed the trained model with final parameter values was the test classifier. It verified the competence of the predicted values of the test model and the value in the trained model.

KDD99/NSL-KDD dataset for the network comprised of sufficient data examples. Training dataset was composed of 1011 Internet packets: 516 instances of normal behavior; 380 instances of DoS attack; 91 instances of probing attack; 11 instances of U2R attack; and 13 instances of R2L attack. The testing dataset for simulation was composed of 25192 Internet packets: 13449 instances of normal behavior; 9234 instances of DoS attack; 2289 instances of probing attack; 11 instances of U2R attack; and 209 instances of R2L attack [1][4]. Training and performance parameters were initialized after creating the models. Since the algorithms used iterative learning, weights and biases were arbitrarily initialized and the packets were presented to the network one at a time. At least one of the training parameters satisfied the model to consider the data as correctly classified. This process was repeated once the training number was reached. The learning algorithm allowed the model to improve performance by adjusting the weights to predict the next set of data correctly. The training stopped when the confusion value was below 1e-2.

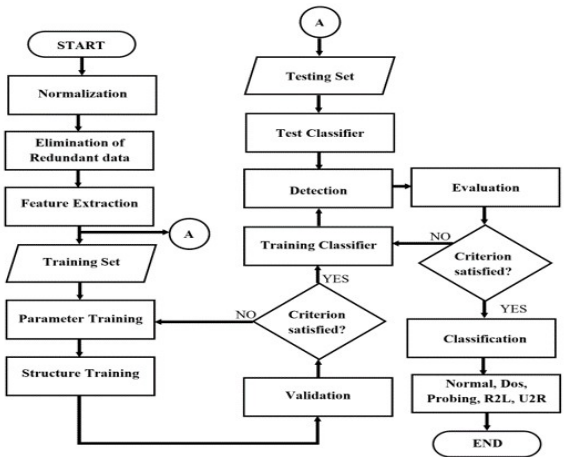

Fig. 2. Information Security System

IV. Results And Discussion

A. Cross Validation

The training data was used as test data to the trained models of the Fuzzy Logic, Artificial Neural Network, and Fuzzy Neural Network. The data in Figure 3 were the average of the results from different sets of simulation test for the trained models of Fuzzy Logic, Artificial Neural Network, and Fuzzy Neural Network. Based on the results, the percentage of validated results on each attack were plotted against the results for each model to compare the performance. For a normal behavior of a network, the Artificial Neural network had the highest accuracy. Whereas the results for the DoS, Probe, U2R, and R2L, the Fuzzy Neural Network had the highest accuracy. The Artificial Neural Network was more susceptible to be biased towards the frequent records in the dataset since the normal behavior is 51.04 percent of the training data. The Fuzzy Logic had better handling of frequent records than the former model. Moreover, the Fuzzy Neural Network was consistent for detection on all the behaviors of the network.

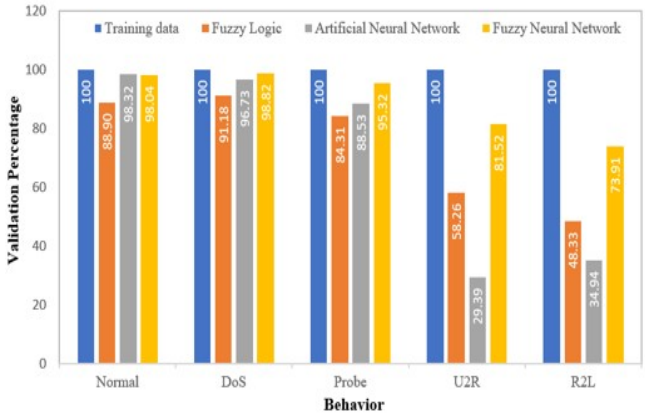

Fig. 3. Cross validation results

B. Comparison of Algorithms

The following data were the results from the best model for each algorithm rendered from the former different tests done. The Figure 4, Figure 5, and Figure 6 show the classification results of the Fuzzy Logic, Artificial Neural Network, and Fuzzy Neural Network respectively. Both Artificial Neural Network and Fuzzy Neural Network provided good overall accuracy. However, the Fuzzy Neural Network classified the U2R and R2L attacks better than Artificial Neural Network which was the least number of packets in the dataset. As observed, the Artificial Neural Network was biased towards the most frequent number of packets in the dataset.

The Attack Detection Rate (ADR) and the F-Measure in Table 1 were calculated using Equation 1, Equation 2, and Equation 3. The closer the value of the F-Measure to 1, the better the result.

$$\text{Attack Detection Rate} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$F - \text{Measure} = 2\,\frac{\text{Precision(Attack Detection Rate)}}{\text{Precision+Attack Detection Rate}} \quad (3)$$

The results in Table 1, both Fuzzy Logic and Fuzzy Neural Network had more linear trained data than the Artificial Neural Network as indicated by the R Value. The predicted values for the validation were closer with the target values of Fuzzy Logic and Fuzzy Neural Network than with the Artificial Neural Network. The fastest algorithm in training the dataset was Fuzzy Neural Network whereas the Artificial Neural Network was the fastest in detecting the behavior of the network. Both the Artificial Neural Network and Fuzzy Neural Network performed better in classifying the behavior of the network than with the Fuzzy Logic.

TABLE I COMPARISON OF ALGORITHMS

Results	Fuzzy Logic	Artificial Neural Network	Fuzzy Neural Network
R Value	0.999999999	0.925208582	0.999999999
RMSE	7.48858E-06	0.498266034	9.30728E-06
Train Time	640.6771245	213.2465686	54.80185094
Validation Time	6.42411E-05	2.10E-05	1.26412E-05
Test Time	5.16351E-05	1.34E-06	1.4226E-05
ADR	0.94841279	0.985061886	0.985975637
F-Measure	0.74542018	0.862748093	0.893571527
Validation Accuracy	100	97.13155292	100
Test Accuracy	89.75468403	96.09399809	96.18926643

The unit of time is in seconds.

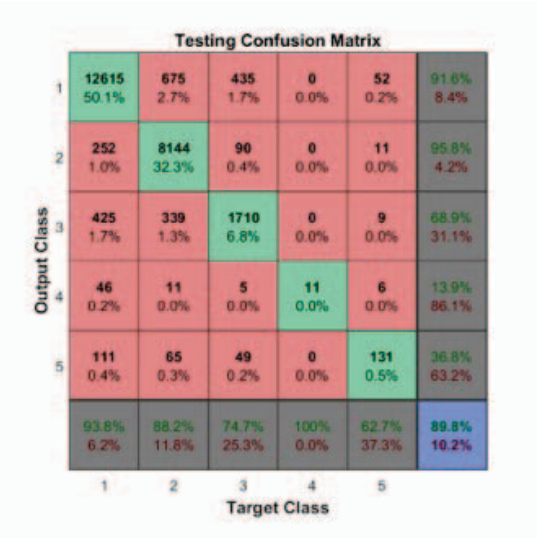

Fig. 4. Confusion matrix for Fuzzy Logic

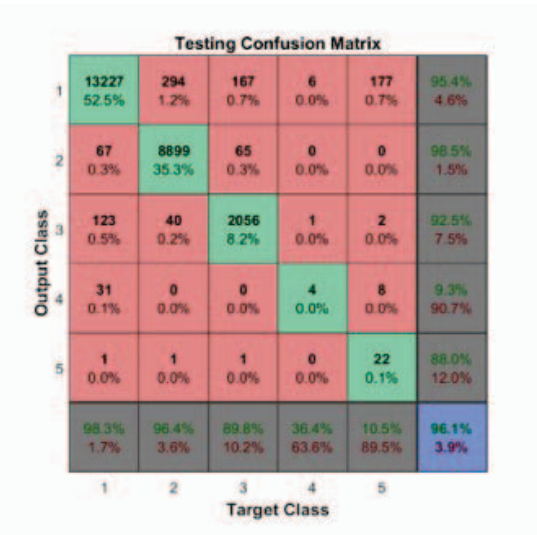

Fig. 5. Confusion matrix for Artificial Neural Network

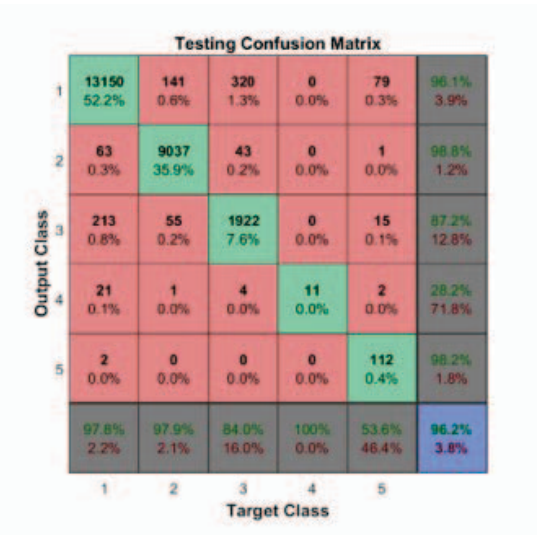

Fig. 6. Confusion matrix for Fuzzy Neural Network

V. Conclusion

The development of the information security system based on three different soft computing algorithms—i.e., Fuzzy Logic, Artificial Neural Network and Fuzzy Neural Network which determines the behavior of the network based on the data packets was successfully implemented using the MATLAB. The cyber-attacks such as DoS, Probe, U2R, and R2L was identified within a short period: 51.64us for Fuzzy Logic, 1.34us for Artificial Neural Network, and 14.23us for the Fuzzy Neural Network. The attack detection rate for the three algorithms is 94.84%, 98.51%, and 98.60% respectively. Also, the accuracy of each algorithm is 89.74%, 96.09%, and 96.19%. Accordingly, the information security system detected the cyber-attacks in a short period of time with high detection rate and low false alarm.

The three algorithms, Fuzzy Logic, Artificial Neural Network, and Fuzzy Neural Network had good performance for the information security system. However, the Fuzzy Neural Network had the best performance among the three. It used the advantages of both Fuzzy Logic and Artificial Neural Network. Based on the tests done, the Artificial Neural Network had a good performance but was vulnerable to be biased towards the most frequent data in the training which was the advantage of Fuzzy Logic. Also, across all the tests the Fuzzy Neural Network was very consistent in the results even when the different parameters were varied. The Fuzzy Neural Network had the least complexity of the calculation based on the results of varied sizes of the models.

Acknowledgement

The authors are deeply grateful to the Engineering Research and Development for Technology (ERDT) of the Department of Science and Technology (DOST), through Mapua University for the MS scholarship of Jason A. Villaluna.

References

[1] U. S. R. Erothi and S. Rodda, "Class imbalance problem in the network Intrusion Detection Systems," ICEEOT, 2685-2688, 2016.

[2] A. Biran and M. Breiner, "Control," in MATLAB 5 for Engineers, England: Prentice Hall, 1999.

[3] H. Al-Mohannadi et al., "Cyber-Attack Modeling Analysis Techniques: An Overview," 4th Intl. Conf. on the Future IoT and Cloud Workshops, 69-76, 2016.

[4] H. Erdem and A. Özgür, "A Review of KDD99 Dataset Usage in Intrusion Detection and Machine Learning Between 2010 and 2015," PeerJ Preprints, 1-22, Apr. 2016.

[5] A. E. Ghorbani, W. Lu, and M. Tavallaee, "A Detailed Analysis of the KDD CUP 99 Data Set," 2nd IEEE Symp. On Comp. Intel. For Sec. and Def. Apps., 1-6, 2009.

[6] A. E. Ghorbani et al., "Toward Developing a Systematic Approach to Generate Benchmark Datasets for Intrusion Detection," *Comp. & Sec*., 357–374, 2012.

[7] E. Bahri and N. Harbi, "Real Detection Intrusion using Supervised and Unsupervised Learning," *Intl. Conf. of SoCPaR*, 321-326, 2013.

[8] M.A. Manzoor, and Y. Morgan, "Real-time Support Vector Machine Based Network Intrusion Detection System Using Apache Storm," 2016 IEEE 7th Annual IEMCON, 1-5, 2016.

[9] Y. Danane, and T. Parvat, "Intrusion Detection System using fuzzy genetic algorithm," 2015 ICPC, 1-5, 2015.

[10] M. B. Ahmed, F. Jemili, and M. Zaghdoud, "Intrusion Detection Based on Hybrid Propagation in Bayesian Networks," *IEEE Intl. Conf. on Intel. And Sec. Informatics*, 137-142, 2009.