ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, No 5, May 2013

Detection Model for Daniel-of-Service Attacks using Random Forest and k-Nearest Neighbors

Phyu Thi Htun and Kyaw Thet Khaing

Abstract*— Because of the increasing of networks' speed and the amount of network traffic, it is essential that IDSs need to be lightweight to cope with it before classification. Feature selection has been successfully used to increase classification accuracy and reduce the false positive for classification of attacks in Intrusion Detection System. In this paper we explore feature selection and classification methods for Denial-of-Service (DoS) attacks detection since they are the most threatening intrusions these days using with Random Forests (RDF) and k-Nearest Neighbor for feature selection and classification respectively. The experimental results presented in this paper show that by estimating the most important features of our data set using RF-KNN. The purpose of this paper is to study the best features selection algorithm, Random Forest in building an IDS that is computationally efficient and effective and the best classification algorithm k -Nearest Neighbors that have been widely used for IDS. Experimental results prove that the proposed method can get the high accuracy in detection those known and unknown attacks by using WEKA tool.

***Index Terms*— Intrusion Detection System, DoS, Random Forest, and k-Nearest Neighbor**

I. INTRODUCTION

An intrusion detection system is an important component of the network architecture in an organization which attempts to protect the network computers against various kinds of possible attacks. Computer systems are exposed to increasing number of such security threats. To overcome these threats, a network intrusion detection system has to adopt the network security policies to detect and react against these threats as quickly as possible. Intrusion detection techniques fall under two categories i.e., misuse detection and anomaly detection.

In misuse detection, the IDS analyzes the information it gathers and compares it to large databases of attack signatures, while in anomaly detection, any deviation from the established profiles of normal activities is treated as an attack. Anomaly detection can detect novel attacks but has a high false positive rate, whereas a misuse detection system cannot detect new attacks. Many intrusion detection techniques today are rule-based [17], which has an intrinsic limitation of low detection rate for new attacks. Therefore, to overcome the limitations of the rule-based network intrusion detection techniques, several data mining techniques have been employed to find models that are better understandable by the data owner [18],[19]. There is a wide variety of network traffic, and the data required for detecting network intrusion is composed of various features. All the features in the network traffic are not necessarily required for intrusion detection. So we need to extract those features which have higher intrusion detection tendency. For that reason, different statistical techniques have been used to reduce the feature space.

There are four main categories of attacks found in the literature, namely: 1) DoS (denial-of-service) attacks, 2) R2L (Remote to Local) attacks, 3) U2R (User to Root) attacks, and 4) PROBE attacks. For establishing a connection, an attacker may follow the same steps e.g., establishing a connection from source IP to the target IP and sending data to the attack target [6]. In KDD99Cup dataset [7] different attacks have different connections, as some of the attacks have few network connections such as U2R and R2L whereas others may have hundreds of network connections such as DoS and Probe. There are different feature values for normal and attack connections in the packet header, and the packet contents can be used as signatures for intrusion detection.

In this paper we propose and implement a hybrid classifier based on Random Forests (RDF) and k-Nearest Neighbor (KNN) algorithm for the classification of DoS attacks in a network. Random Forests is an ensemble classification and regression approach which is unbeatable in accuracy among current data mining algorithms. Random forests algorithm has also been used in applications like prediction [20], probability estimation [4], and pattern analysis. After reducing the features we need to classify the records between other records and DOS attacks. We further optimize the selected features with the help of RF algorithm and compare our results with several different classification techniques. We evaluate our proposed approach on KDD'99Cup dataset. Experimental results show that by using the proposed approach, average detection rate is increased and at the same time average false positive rate is also decreased when compared with other algorithms.

The rest of the paper is organized as follows. Section 2 presents the related research using corresponding machine learning Algorithms. In section 3 described the KDD 99 CUP intrusion detection datasets which including DoS attacks data. Section 4 introduces about the proposed system for DoS detection. The usefulness of Random Forest and k-NN, presented in Section 4, Section 5 describes the experimental results obtained, using the Random Forest for feature selection and examined with k-NN for performance

Manuscript received May, 2013.

***Phyu Thi Htun**, Faculty of Information and communication Technology, University of Technology ,Yatanarpon Cyber City, Myanmar., Yatanarpon Cyber City, Myanmar, Phone/ Mobile No: +959402560402*

***Kyaw Thet Khaing**, Computer Hardware Department, University of Computer Studies, Yangon,Yangon, Myanmar,*

evaluation analysis. Section 6 explains the conclusion and further extension to our research by using out coming results.

II. Related works

Park et al. [22] proposed Correlation-Based Hybrid Feature Selection (CBHFS) approach. GA was used to generate subsets of features from given feature set. CBHFS took full feature set as input and returned the optimal subset of feature after being evaluated by Correlation-Based Feature Selection (CFS) [23] and SVM. Each chromosome represented a feature vector. Merit of each chromosome was calculated by CFS. The chromosome having highest Merit represented the best feature subset in population. This subset was then evaluated by SVM classification algorithm. Then, this procedure iterated with a new population of chromosomes, which is generated through performing genetic operations. The algorithm stopped if better subset was not found in next generation or when maximum number of generation was reached. All these methods yield good improvements but they are also fairly complex and computationally expensive as Kim et al.'s approach [21]. In other words, these two hybrid approaches [21][15] sometimes showed a little degradation on detection rates with more computations rather than the nave filter methods, did not provide the variable importance of features and were complicated to implement.

Chen et al. [24] conducted wrapper-based feature selection algorithm aiming at modeling lightweight IDSs. They used Modified Random Mutation Hill Climbing (MRMHC) as search strategy to specify a candidate subset for evaluation. Then, they adopted SVM as wrapper approach to obtain the optimum feature subset. Their MRMHC method can be enhanced in terms of speed compared to Kim et al. [21] and Park et al. [22]. They also used MRMHC to obtain the optimal parameters for kernels in SVM but this approach is still complex to implement.

Zhang et al. [12] and Kim et al. [21] performed feature selection and parameter optimizations based on RF for lightweight DoS attacks detection. Feature importance ranking was performed according to the result of variable importance values. Then, irrelevant features are eliminated and only important features are selected. Zhang et al. [12] only cut off 3 features and optimized only mtry value. On the other hand, Kim et al. [21] eliminated much more irrelevant features and optimized both of mtry and ntree, two parameters of RDF.

Moreover, one of the main problems of existing approaches is that IDSs provide only binary detection results: intrusion (attack) or normal. That is a main cause of high false rates and inaccurate detection rates in IDSs. If some attack or normal data are belonging to boundary, they may be classified wrong. To cope with it, Lee et al. [25] proposed Quantitative Intrusion Intensity Assessment (QIIA). It provides intrusion (or normal) quantitative intensity value. It is capable of representing how an instance of audit data is proximal to intrusion (DoS attacks) or normal in a numerical value such as "0.95" proximity to intrusion. It can be interpreted as the instance has a probability of 0.95 to be classified as an intrusion. This approach is very novel and refreshing paradigm. It can overcome the drawback of current binary detection and classify intrusions in more detail. For example, DoS attacks can be classified as Smurf, Neptune, Teardrop, etc.

III. Datasets description

Since 1999, KDD'99 [8] has been the most widely used data set for the evaluation of anomaly detection methods. This data set is built based on the data captured in DARPA'98 IDS evaluation program [5]. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic. The two weeks of test data have around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. There can caterogrized into four types: Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing Attack.

Table I showed the four categories and their corresponding attacks on each categories.

TABLE I. Classification of Attacks on KDD Dataset

Classification of Attacks	**Attack Name**
DoS	smurf, land, pod, teardrop, neptune, back
R2L	ftp_write, guess_passwd, imap, multihop, phf, spy, warezmaster, warezclient
U2R	perl, buffer_overflow, rootkit, loadmodule
Probe	ipsweep, nmap, satan, portsweep

TABLE II. Number of connection in each attack type

Datasets	Normal	DoS	U2R	R2L	Probe	Total
Train+	67343	45927	993	54	11656	125973
Train+20 Percent	13449	9234	206	12	2289	25190
Test+	9711	7458	2421	533	2421	22544
Test-21	2152	4342	2421	533	2402	11850

It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data which make the task more realistic. Some intrusion experts believe that most novel attacks are variants of known attacks and the signature of known attacks can be sufficient to catch novel variants.

TABLE III. NUMBER OF DOS ATTACK CONNECTIONS IN EACH ATTACK TYPE

Datasets	back	neptune	smurf	Pod	teardrop	land
Train+	956	41214	8646	201	892	18
Train+20 Percent	196	8282	529	38	188	1
Test+	359	4657	665	41	12	7
Test-21	359	1579	627	41	12	7

The KDD CUP shared 4 dataset file, Train+, Train+_20Percent,Test+ and Test-21. The first two files represent for training datasets and contain the general attacks. The rest two files represent for testing datasets and contain not only general attacks but also the unknown (novel) attacks. The connection for each attack type is shown in Table 2.

According the number of connection in Table 2, the number of connection of DoS are the highest. Also, we can say that those DoS are the most threatening intrusions in these days. According to these reasons, we must think an importance fact to detect the DoS attacks as precise as we can. The number of DoS attacks connection according in each DoS attacks are shown in Table 3.

IV. FEATURE SELECTION FOR IDSS

With the growth of the bandwidth and interconnectivity of computer networks, DoS attacks became a major intrusion to the Internet infrastructure integrity and one of the most devastating attacks possible to throw across the network. Unfortunately, automated tools make these attacks increasingly easier to execute and then they are becoming more sophisticated. A growing number of DoS attacks impose a significant threat on the availability of network services since DoS attacks attempt to overwhelm victim machines and it causes legitimate users to prevent from accessing their computing resources.

Daniel-of-Service (DoS) Attacks

Some DoS attacks target the bandwidth capabilities of computer systems while others target the machines' computational state. To foil the DoS attacks, a number of countermeasures are needed and detecting is the first step of countermeasures. Many approaches based on very well-known detection (classification) algorithms, such as SVM, RDF, and so on, have been proposed and focused on the achievable accuracy (detection rates).

Since a simple DoS attack is normally that large amounts of traffic are generated and sent to a target machine, DoS attacks detection system should keep up with the large amounts of traffic. Thus, DoS attacks detection should be lightweight. Unlike previous detection approaches, there are many approaches which adopted parameters optimization and feature selection together.

Selecting Features with RDF

The random forests[2] is an ensemble of unpruned classification or regression trees. Random forest generates many classification trees. Each tree is constructed by a different bootstrap sample from the original data using a tree classification algorithm. After the forest is formed, a new object that needs to be classified is put down each of the tree in the forest for classification. Each tree gives a vote that indicates the tree's decision about the class of the object. The forest chooses the class with the most votes for the object.

The main features of the random forests algorithm are listed as follows:

- It is unsurpassable in accuracy among the current data mining algorithms.
- It runs efficiently on large data sets with many features.
- It can give the estimates of what features are important.
- It has no nominal data problem and does not over-fit.
- It can handle unbalanced data sets.

Information Gain

In this method, the important features are calculated over multiple RDF iterations, the least important features being removed after each. The objective of using Random Forest is to reduce the impurity or uncertainty in data as much as possible .A subset of data is pure if all instances belong to the same class. The heuristic is to choose the attribute with the maximum Information Gain or Gain Ratio based on information theory.

Entropy is a measure of the uncertainty associated with a random variable. Given a set of examples D is possible to compute the original entropy of the dataset such as:

$$H[D] = -\sum_{j=1}^{|C|} P(c_j)\log_2 P(c_j)$$

where C is the set of desired class.

If we make attribute Ai, with v values, the root of the current tree, this will partition D into v subsets D1,D2….Dj . The expected entropy if Ai is used as the current root.

$$H_{A_i}[D] = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} H[D_j]$$

Information gained by selecting attribute Ai to branch or to partition the data is given by the difference of prior entropy and the entropy of selected branch.

$$gain(D, A_i) = H[D] - H_{A_i}[D]$$

We can choose the attribute with the highest gain to branch/split the current tree.

Classification with k-NN

k-NN classification is an easy to understand and easy to implement classification technique[13]. Despite its simplicity, it can perform well in many situations. k-NN is particularly well suited for multi-modal classes as well as applications in which an object can have many class labels. For example, for the assignment of functions to genes based on expression profiles, some researchers found that k-NN outperformed SVM, which is a much more sophisticated classification scheme [1].

The 1-Nearest Neighbor (1NN) classifier is an important pattern recognizing method based on representative points [14]. In the 1NN algorithm, whole train samples are taken as representative points and the distances from the test samples to each representative point are computed. The test samples have the same class label as the representative point nearest to them. The k-NN is an extension of 1NN, which determines the

test samples through finding the k nearest neighbors.

Proposed System

In this section, we describe the methods employed in the proposed system, and illustrate how to apply these methods to detect DoS attacks with high detection rate with high true positive rate, low false positive rate for network intrusion detection.

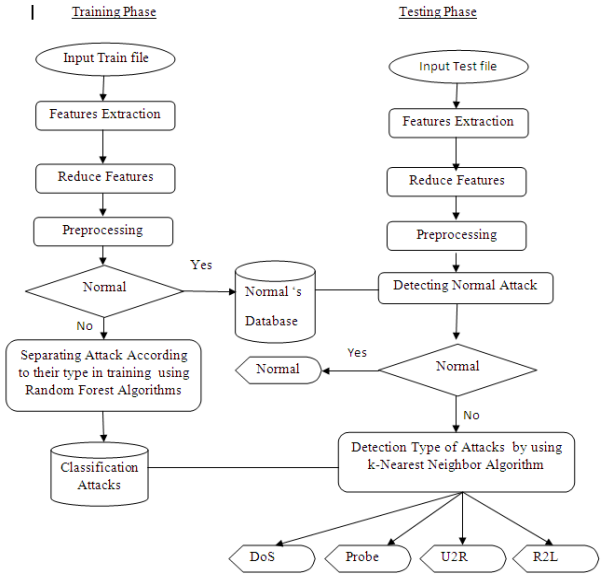

Figure 1. The proposed System and Mode

This system is process of identifying the abnormal and normal instances that are two phases. The first is the training phase that reduces the irrelevant features. Next phase is detection phase. This proposed system and model are shown in Figure 1.

Since the operations of normal instances are specified and they show expected behavior, we could use the knowledge based (misuse) IDS detection, while unexpected activity (presumably an intrusion would be unusual) is continually designed and progressed and could not be seen as a knowledge based attack, therefore the anomaly IDS detection is performed over novel attacks.

We also report our experimental results over the KDD'99 datasets. The results show that the proposed approach provides better performance compared to the best results from the KDD'99 contest.

V. Experimental Results

In this section, we summarize our experimental results to detect DoS attacks for intrusion detection with over the KDD'99 datasets. Experimental results are presented in terms of the classes that achieved good level of discrimination from others in the training set.

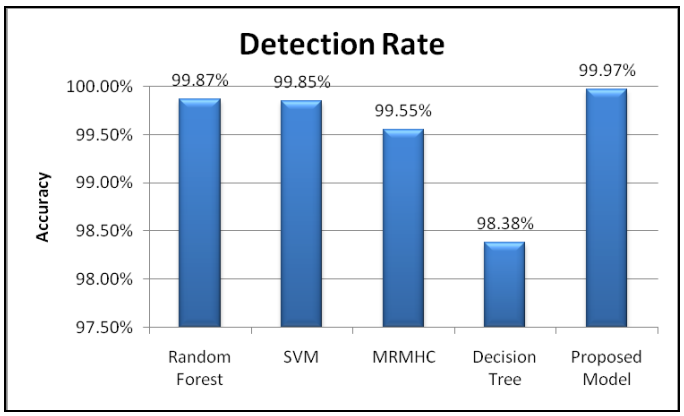

Figure 2. The detection rate between proposed method and other methods.

Firstly, our proposed system will reduced some features in dataset by using Random Forest algorithm at each connection.

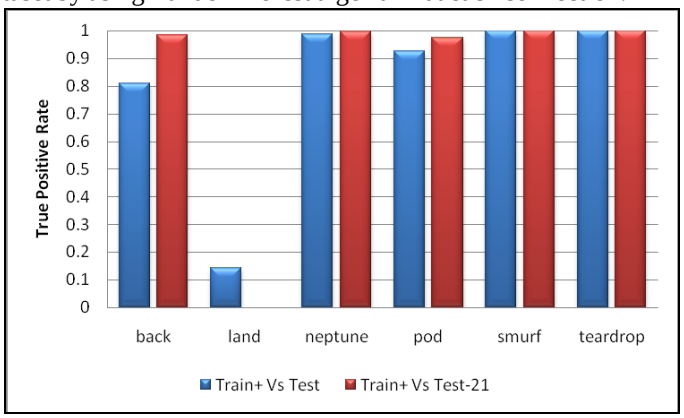

Figure 3. The True Positive Rate on each DoS attack type

So, system will try to detect various anomaly attacks using KDD dataset. The proposed system will reduced in training time and will increase the accuracy of the system's classification. The experimental results will come out by using 10 fold cross validation on train+ atWEKA tool [9].

In the experiments process, the system use 10 trees and reduced features (default 6 in WEKA) to classify. To reduce the features from the dataset, use information gain method with ranker features in WEKA. The accuracy of the system will be increased other systems as shown in Figure 2 and the true positive rate using proposed method on each DoS attack type are shown in figure 3.

According to the results of the Figure 3, we can prove that the true positive rate of our proposed model may higher even exactly to 1. But the same as the others researches, we have a dependency on the datasets. If there is a large amount of appropriate attack connections in a training dataset, we can examine in more details and can be high accuracy. But there is also an increasing of training data size, trees and modeling time.

VI. Conclusion and Future Extension

Recent researches employed decision trees, artificial neural networks and a probabilistic classifier and reported, in terms of detection and false alarm rates, but it was still high false positives rates and irrelevant alerts in detection of DOS attacks.

This paper has presented a comparison of the various data

mining techniques that have been proposed towards the enhancement of anomaly intrusion detection systems. And, we applied the classification methods for classifying the attacks (intrusions) on DARPA dataset. The results showing the performance of the RDF-KNN is better than other classifiers.

Thus, we can extend this experiment for detecting all types of attacks by our proposed model using those two algorithms; the system may expect to get the more accurate and detection rate to detected intrusions. Random Forest will process in the filtering stage and the k-NN will use as a classifier.

Because it can get high accuracy and true positive rate in detecting the DoS attacks than the others, it can use as a novel method to detect the unknown attacks, such as Distributed DoS (DDoS), Bob and others in Intrusion Detection and Prevention Systems (IDPS).

REFERENCES

[1]. K.T.Khaing and T.T.Naing, "Enhanced Feature Ranking and Selection using Recurisive Featue Elemination and k-Nearest Neighbor Algorithms in SVM for IDS", Internaiton Journal of Network and Mobile Technology(IJNMT), No.1, Vol 1. 2010.

[2]. L. Breiman, "Random Forests", Machine Learning 45(1):5–32, 2001.

[3]. V. Marinova-Boncheva, "A Short Survey of Intrusion Detection System" , 2007.

[4]. M. Mahoney and P. Chan, "An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection", Proceeding of Recent Advances in Intrusion Detection (RAID)-2003, Pittsburgh, USA, September 2003.

[5]. KDD'99 datasets, The UCI KDD Archive, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html , Irvine, CA, USA, 1999.

[6]. KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, December 2009.

[7]. Lan Guo, Yan Ma, Bojan Cukic, and Harshinder Singh, "Robust Prediction of Fault-Proneness by Random Forests", Proceedings of the 15th International Symposium on Software Reliability Engineering (ISSRE'04), pp. 417-428, Brittany, France, November 2004.

[8]. Yimin Wu, High-dimensional Pattern Analysis in Multimedia Information Retrieval and Bioinformatics, Doctoral Thesis, State University of New York, January 2004.

[9]. WEKA software, Machine Learning, http://www.cs.waikato.ac.nz/ml/weka/, The University of Waikato, Hamilton, New Zealand.

[10]. MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation, http://www.ll.mit.edu/IST/ideval/,MA, USA.

[11]. J.Zhange and M. Zulkerline, "Network Intrusion Detection using Random Forests",2011.

[12]. J. Zhang and M. Zulkernine, "Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection", Symposium on Network Security and Information Assurance Proc. of the IEEE International Conference on Communications (ICC), 6 pages, Istanbul, Turkey, June 2006.

[13]. S. Thirumuruganathan , "A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm", World Press, May 17, 2010.

[14]. X Wu, V Kumar, J Ross Quinlan, J Ghosh, "Top 10 Data mining Algorithm", Knowledge and Information Systems, Volume 14, Issue 1, pp 1-37 ,2008 – Springer

[15]. Press, W.H., Flnnery B.P., Teukolsky S.A. and Vetterling W.T. .Numerical recipies in C.. Cambridge University Press, Cambridge.

[16]. Chi J., .Entropy based feature evaluation and selection technique., Proc. of 4th Australian Conf. on Neural Networks (ACNN'93), pp. 181-196, 1993.

[17]. Waseem Shahzad, Abdul Rauf Baig, "Compatibility As a Heuristic For Construction Of Rules By Artificial Ants", Journal of Circuits, Systems, and Computers, vol.19, no.1, pp.297-306, Feb 2010.

[18]. Tamas Abraham, "IDDM: Intrusion Detection Using Data Mining Techniques", Technical Report DSTO-GD-0286, DSTO Electronics and Surveillance Research Laboratory, 2001.

[19]. Daniel Barbarra, Julia Couto, Sushil Jajodia, Leonard Popyack, and Ningning Wu, "ADAM: Detecting Intrusions by Data Mining", Proceedings of the 2001 IEEE, Workshop on Information Assurance and Security, NY, USA, June 2001.

[20]. Li-Yeh Chuang, Jung-Chike Li, and Cheng-Hong Yang, "Chaotic Binary Particle Swarm Optimization for Feature Selection using Logistic Map", In Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2008), Hong Kong, pp. 131-136, March 2008.

[21]. D. Kim, H.-N. Nguyen, S.-Y. Ohn, and J. Park. Fusions of GA and SVM for Anomaly Detection in Intrusion Detection System. In Proc. of the 2nd International Symposium on Neural Networks (ISNN'05), Chongqing, China, LNCS, volume 3498, pages 415–420, Springer-Verlag, May 2005.

[22]. J. Park, K. M. Shazzad, and D. Kim. Toward Modeling Lightweight Intrusion Detection System through Correlation-Based Hybrid Feature Selection. In Proc. of the 1st SKLOIS Conf. on Information Security and Cryptology (CISC 2005), Beijing, China, LNCS, volume 3822, pages 279–289, Springer-Verlag, December 2005.

[23]. M. A. Hall. Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. In Proc. of the 17th International Conference on Machine Learning (ICML'00), Stanford, California, USA, pages 359–366, Morgan Kaufmann, June 2000.

[24]. Y. Chen, W.-F. Li, and X.-Q Cheng. Toward Building Lightweight Intrusion Detection System through Modified RMHC and SVM. In Proc. of the 15th IEEE Int. Conf. on Networks (ICON'07), Adelaide, Australia, pages 83–88, IEEE, November 2007.

[25]. S. Lee, D. Kim, Y. Yoon, and J. Park. Quantitative Intrusion Intensity Assessment using Important Feature Selection and Proximity Metrics. In Proc. of the 15th IEEE Pacific Rim Int. Symp. on Dependable Computing (PRDC'09), Shanghai, China, pages 127–134, IEEE, November 2009.

Phyu Thi Htun received the B.E. degree in Information and Technology Engineering from Government Technological College, Thanlyin , Myanmar, in 2006. And she received M.E degree in Information and Technology Engineering from Westen Yangon Technology University,Yangon,Myanmar, in 2010 respectively. She is currently doing research for Ph.D(IT) in University of Technology ,Yadanapon Cyber City , Myanmar since November 2010. Her research interests are in network security, especially intrusion detection systems, and survivability of cloud computing.

Kyaw Thet Khaing received the Master of Computer Technology from University of Computer Studies,Mandalay , Myanmar, in 2004. And he received Ph.D degree in Information Technology from University of Computer Studies,Yangon,Myanmar, in 2010 respectively. His research interests are in network security, especially intrusion detection systems, and survivability of cloud computing.