

## **Genetic-SVM Based Intrusion Detection System for DoS Attacks**

**K. Pradeep Mohan Kumar**

*Research Scholar, CSE PMU Thanjavur, India  
pradeep\_nv2004@yahoo. co. in*

**Dr. M. Aramuthan**

*Associate Professor, IT PKIET Karaikal, India  
aranagai@yahoo. co. in*

**T. Uthra Devi**

*Final B. Tech, CSE PMU Thanjavur, India  
uthradv07@gmail. com*

### **Abstract**

Denial-of-service (DoS) attacks comprised of large numbers of packet streams from different sources on the victim, consuming key resources and rendering it unavailable to authorized users. Hence, there is a need of new techniques and tools to handle this kind of attacks before damaging wide areas. In modern electronic society, IDS have become a necessary component for protecting interconnection of computer resources very effectively. In this paper new hybrid based IDS model, based on Genetic Algorithm (GA) and Support Vector Machine (SVM) for DoS attacks Detection. Attacks are identified by training the SVM classifiers after extracting features from PMU 2014 datasets using Genetic Algorithm. SVM classifier deals with large volume of data, that it easy to detect suspicious behaviours, which takes short time for training and testing process. Genetic-SVM based on wrapper based feature selection which is superior than filter based feature selection. The proposed work was implemented in Mat lab7. 2. The result shows that the proposed hybrid IDS has high detection accuracy (99. 5%) and (0. 5%) of false alarms.

**Keywords**— Intrusion Detection System (IDS), Denial of Service (DoS), GA-SVM

## **I. INTRODUCTION**

An intrusion detection system (IDS) is the process of identifying, blocking and responding to the unauthorized activity as a system administrator to take necessary action. Based on the data collection mechanism IDS can be classified in to three types (i) HIDS, (ii)NIDS, (iii)Hybrid IDS. HIDS resides on a particular host and looks for the indications of attacks on that host system. NIDS is located on a separate system and monitors another system through the network traffic for finding attacks based on rule set. Hybrid based IDS perform both the functionality of Network based and Host based intrusion detection system.

Based on the attack detection techniques IDS can be classified into (i) Anomaly Detection, (ii) Misuse Detection. In anomaly based detection, captured network traffic data is used to compare with predefine normal pattern, if any deviation occur in this pattern considered as attack. On the other hand, misuse detection system, also called as signature based IDS, uses patterns of well known attack pattern to match with captured traffic to find out they Known attack very easily. DoS attacks has become an emerging attacks that creates threat to E-business and Internet providers around the world. So, we need new components like Computational Intelligence (CI) is needed for encountering this kind of threads for building automatic intelligent security system that detect malicious activities and discard it. CI should be provided with learning, adaption, optimization and evolution to deal with new situations, flexible to adapt to the changing environments and goals. CI techniques mainly used for constructing IDS based on artificial intelligence, neural network, fuzzy sets, evolutionary computation, expert system approach, rule based approach, artificial immune systems etc [1]. Existing security systems are evaluated with DARPA 98 and KDDCup99 datasets that affects the performance of system, this dataset only having old pattern of the attacks that degrade the performance of security systems. So, new Hybrid based IDS is the more power full security system to detect the various types of new DoS attacks pattern with higher detection accuracy, reducing false alarms with the help of new dataset.

Hybrid based IDS is examined with large Volume of data set that causes slow training, testing process and low detection rate. So, feature extraction is the challenging task in developing IDS. Generally, the implementation of new hybrid based IDS consists of three phases such as data preprocessing, features selection and classification. The tasks that are carried out in preprocessing phases are (i) Identifies the attributes and their relevant value (ii) Convert original data to numerical data (iii) Perform Data normalization and (iv) Compute redundancy check and handle about null value. Feature selection process is a preprocessing step when constructing IDS, used to reduce the dimensionality of the dataset by removing irrelevant, redundant features and improving the prediction accuracy of the classifier using selected features from the dataset. Classifier module handle large volume of data to finds the conditions of the traffics are either legitimate or malicious attack. Classifier is faced with a problem when it has to generate rules with many attributes or features. Obviously, the time required to generate rules is proportional to the number of features. In addition, irrelevant and redundant features can reduce both the predictive accuracy and comprehensibility of the induced rule and degrade the classifier speed. Thus, selecting

the most relevant features is necessary, this strategy is implemented to simplify the rule set and reduce its computational time while retaining the quality of classifier, as it represents the original features set.

Genetic Algorithm (GA) is an efficient search method based on principles of natural selection and population genetics. It is being effectively applied to problems in business, engineering and science. GA uses randomized operators operating over a population of candidate to generate new points in the search space. GA uses three operators namely selection, crossover and mutation. The selection operator identifies the fittest individuals of the current population to serve as parents of the next generation. Cross over operator combines the second half of the first record with the first half of the second record. Mutation operator randomly changes the bits from '0' to '1' and vice versa. Support Vector Machines are known as maximum-margin classifiers since they find the optimal hyper plane between two classes, defined by a number of support vectors. The feature of the technique is mainly due to the introduction of calculation of pattern weight that allows us to prevent the effects of outliers by permitting a certain amount of misclassification errors. Although this technique was able to provide only linear classification and also handle non-linear problems. The Objective function is used to implicitly map the data points into a higher-dimensional feature space. The rest of this paper is organized as follows. In section II discusses the related works about existing algorithm to detect DoS attacks, In section III illustrates the proposed genetic-SVM based IDS model. Section IV describes the implementation and performance of the proposed algorithm using PMU 2014 dataset. In the last section, deals with conclusion and future work.

## **II. EXISTING WORK**

In the field of DoS and DDoS attacks, many researchers have been done up to now. Many of them consisted of traditional approaches such as firewall, encryption techniques and etc, which didn't met all needs of an intelligence detection system. There upon, researchers attended to the artificial intelligence and data mining techniques. Iftikhar et al [2] introduced feature selection mechanism based on Principal Component Analysis (PCA). However, since this method might ignore some sensitive features, a method was proposed based on Genetic Algorithm and multilayer perceptron (MLP) - The neural network algorithm for mapping input to appropriate output. KDDCup99 was used for dataset. As a result, they selected 12 features among 41 features and claimed that accuracy has improved to 0.99. As mentioned at [3, 4], PCA is not suitable for large dataset and this method is executable just for small dataset. In [5] Singh and Silakari stated that PCA is not proper solution for non-linear dataset, therefore they presented an algorithm based on Generalized Discriminant Analysis (GDA), to generate small size of features and improve classification operation. They asserted that this method is premier than other classification method such as Self-Organizing Map (SOM) and C4.5. KDDCup99 was used for dataset in that research, also 4 different attacks were reviewed: DoS attack, User to Root Attacks, Remote to Local (User) Attacks, and probing. Finally their method accuracy was about 0.98.

Most of the researches in the scope of intrusion detection attack, offer the model by analyzing raw packet data, and processing vast amount of data especially while occurring DoS attack is the main challenge for researchers. For this reason, the idea of attack detection based on statistical data gained from network management protocol was raised. MAID [6] was an intrusion detection system that monitored 27 different SNMP MIB variables and compared the behavior of normal and attack packet. Normal behavior of packet was modeled using probability density function (PDF), and was kept as reference PDF. They compared five similarity metrics by examining algorithm on actual network data and attack. They stated that KST is able to detect more attacks in all situations even at low traffic intensities. D. Dutta and K. Choudhury at [7] claimed that their research was the first intrusion detection system which was integrated Digital Signature of Network Segment (DSNS) with Particle Swarm Optimization (PSO). They also benefited SVM to optimize clustering operation and better centroids selection. PSO [8] is a Swarm Intelligence algorithm, which despite the high Efficiency has low computational complexity.

At [9], J. Yu et al also presented a model based on SNMP and SVM. Unlike previous model that had just introduced a detection model, they proposed two layer architecture. The first layer detected DoS/DDoS attack and the second layer detected these types of attack: TCP-SYNC, UDP and ICMP. Attack type identifying has the advantage of filtering the corresponding packet. Extended architecture of this model was proposed at [10]. Classification and association rule mining that performed by C4. 5 algorithm was operated offline, while getting SNMP MIB variable and detection DoS/DDoS attack was done online. After getting Dataset and generated new packet data, Offline modules extracted model and valuable rule and passed the result to detection module. Function of Getting MIB module was to schedule operation of SNMP pooling. Authors asserted that accuracy value obtained for detection attack was about 99. 13%.

### III. PROPOSED GENETIC-SVM BASED IDS

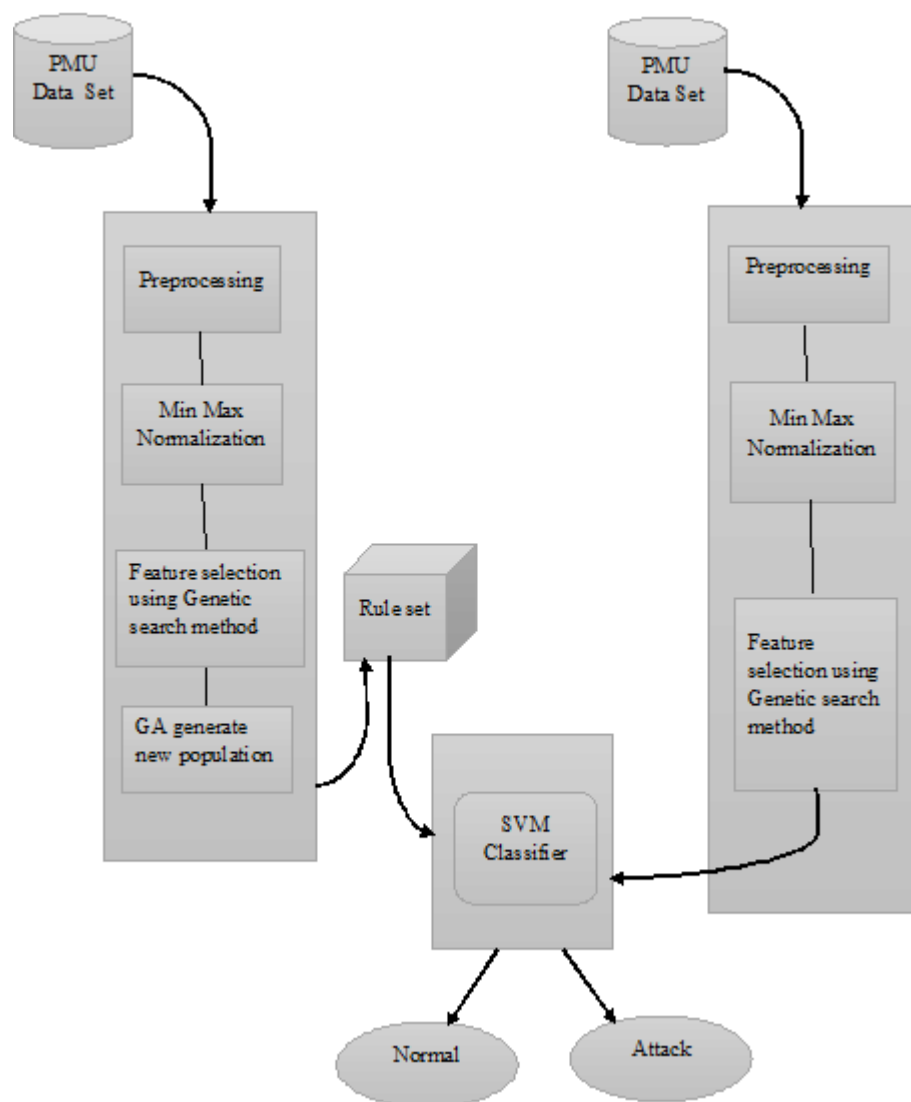
The architecture of the proposed GA-SVM model is shown in Fig. 1. The architecture contains two phases (i) Training phase (ii) Testing phase. In training phase, PMU 2014 dataset undergoes data pre-processing and the pre-processed data is then fed to the feature selection block where feature selection is done using genetic algorithm. The selected features are then given as input to the classifier where DoS attack patterns are classified using SVM. The next stage is the testing stage where the captured traffic is pre processed as in training phase and the identified patterns are matched with the stored DoS patterns in database thereby taking a decision. If any new patterns were found during the analysis of traffic behaviour and if it was found against the legitimate traffic, the new pattern will be captured and updated in the database. The implementation of Genetic-SVM based IDS has three phases which includes

- A. Preprocessing,
- B. Feature Selection and
- C. Classifier.

### A. Data Preprocessing

Data Preprocessing is an important step in the machine learning computing that eliminates out of range values, impossible data combinations, missing values etc. Generally data preprocessing includes learning, normalization, transformation, feature extraction and selection. The output of the data preprocessing is the final training set that extracts knowledge for the testing phase. The following steps are involved in data preprocessing.

- Identifying features and its related values.
- Converting original feature data value in to numerical data value.
- Applying data normalization based on min-max normalization.
- Performing similarity checks and removing null values.



**Fig. 1. Architecture of Genetic-SVM based IDS**

### **B. Feature selection based on Genetic algorithm**

Accuracy of the classifier depends on the selection of optimum feature subset. Feature selection method is mainly used for selecting the subset of features from the original data set. Two feature selection methods namely filter method and wrapper method were already proposed. Filter method is mainly based on the general characteristics of data features without involving machine language. These features are ranked based on certain criteria, where features with highest rank values are selected as optimal. The main advantages of filter method are low computational cost without involving any machine language algorithm for feature selection. Frequently used filter method is the information gain method. Wrapper method is mainly used for feature subset selection from the data set based on objective function and analysis of the performance of feature subset. In this paper, Genetic Algorithm (GA) is used to select optimal feature subset from the dataset. GA reduces the dimensionality of PMU 2014 features from 113 attributes to 12 attributes those are related to the characteristics of DoS attack thereby reducing 88% of the space of features. The twelve attributes that are considered by the GA are 1. Protocol, 2. Source Ip, 3. Dest Ip, 4. src\_byte 5. dst\_bytes, 6. count (No of connections to the same Dest), 7. SYNcount, 8. srv\_count, 9. same\_srvrate, 10. diff\_srvrate, 11. dst\_, host\_same\_src\_port\_rate, 12. FINcount. The existing KDDCup99 dataset contains huge number of redundant records. 10% of the full dataset contains two types of DoS attacks (Smurf and Neptune). These two types constitute over 71% of the testing dataset which completely affects the evaluation of IDS. The steps involved in GA where features are selected from the dataset are presented below in Alg. 1.

- 1) Initialize pre-processed data as population.
- 2) Calculate objective function based on derived rules for DoS attack for each individual pre processed data.
- 3) Select individual solution.
- 4) Perform mating of pair of individuals.
- 5) Perform mutation operation.
- 6) Calculate objective function for newly created population.
- 7) If (6) is satisfied, stop the operation.
- 8) If (6) is not satisfied, repeat from step 3.
- 9) Return the best features from PMU dataset that reflects the properties of DoS attacks.

#### **Alg. 1 Genetic Algorithm based feature selection.**

### **C. SVM Classifier**

GA generates a set of enhanced population of chromosomes, i. e. a group of individuals with different chromosomes. Each individual chromosome consists of twelve different parameters namely

1. Protocol,
2. Source Ip,
3. Dest Ip,

4. src\_byte
5. dst\_bytes,
6. count (No of connections to the same Dest),
7. SYNcount,
8. srv\_count,
9. same\_srvrate,
10. diff\_srvrate,
11. dst\_, host\_same\_src\_port\_rate,
12. FIN count.

The pattern weight of the individual chromosome should be determined properly by using the training dataset to include as many solutions as possible. Calculate the fitness value of each individual in the initial population using Eq. (1) and rank them according to their fitness value. In Eqn (1), X indicates the training dataset and Y indicates the enhanced chromosome subset.

$$F = \sum_{k=1}^n X * Y \quad \dots (1)$$

To calculate the fitness value of an individual or a chromosome, the training record is compared with each gene of the chromosome in the normal population. So each and every record generates different pattern value for different feature values. Similarly training record is compared with each gene of the chromosome in the attack population. So each and every record generates different pattern value for different feature values. Finally we will calculate support vector values for normal pattern and attack pattern with the help of pattern weight for normal and attack population. Now, we will get two SVM values i. e 1 and 0.  $F \geq 1$  indicates normal record and  $F < 1$  indicates attack record. SVM classifies our dataset based on newly generated hyper plane values. Now each and every testing record is compared with each and every gene of the normal and attack population. This generates a pattern weight {0, 1, 2, 3, 4, 5} based on which we will identify whether our testing record belongs to normal or attack model.

## Rule Structure of Dos Attacks in PMU2014 datasets

### Normal Rule Set

```
protocol=tcp, source Ip=172. 20. 62. 33, DestIp=172. 20. 62. 255, 178>src_byte<322,
10>Dst_byte<224, SYNcount=1or2, 1>count<28, 1>srv_count<28, same_srvrate=1,
diff_srvrate=0, 0>dst_, host_same_src_port_rate<1, FIN=1.
```

### Neptune Rule Set

```
protocol=tcp, source Ip= 172. 20. 62. 33, Dest Ip= 172. 20. 62. 255, src_byte= 0,
Dst_byte=0, SYN bit=3 to 160, 3>count<160, 3>srv_count<160, same_srvrate=0,
0>diff_srvrate<1, 0>dst host same src port rate<1, FIN bit=0.
```

### Smurf Rule Set

```
protocol = UDP, source Ip=172. 20. 62. 33, Dest Ip=172. 20. 62. 25, src_byte=
221120, Dst_byte=0, SYN bit=238to512, 238>count<512, 238>srv_count<512,
same_srv rate=1, 0>diff_srv rate<2, 0<dst_, host_same_src_port_rate<2, FIN bit=0.
```

**TABLE 1**

## IV. SIMULATION RESULTS AND DISCUSSIONS

A new dataset for training the models which overcome the limitations of unavailable new attack patterns was developed and efficient training of the models with zero deficiency is achieved. Three types of traffic comprising of both normal and abnormal pattern were generated by using attacking tools XOIC and LOIC which was directed towards the Web server 2012 for about 7 days and all the traffic was captured using Wire shark tool. In total, it contains 20, 00, 000 records and 113 features with 410 MB size. Out of this, Neptune attack has 15, 13, 000 records and Smurf has 4. 00219 records.

The proposed computational intelligence based IDS was implemented in Mat Lab 7. 2. During the evaluation, 100% labelled data of PMU 2014 dataset was used for training and testing the proposed IDS. DoS attacks are difficult to be dealt with because they are very easy to launch, difficult to track and also it is not easy for the victim resources to refuse it. Neptune (Syn Flood), Smurf are the two kinds of DoS attacks in PMU 2014. In the "Smurf" attack, attackers use ICMP echo request packets which is directed to IP broadcast addresses from remote locations to create a denial-of-service attack. Neptune attack describes that each half-open TCP connection made to a machine causes the 'tcpd' server to add a record to the data structure that stores information about all pending connections. This data structure is of finite size, and it can be made to overflow by intentionally creating too many partially-open connections. Neptune attack can be distinguished from normal network traffic by looking for a number of simultaneous SYN packets destined for a particular machine

that are coming from an unreachable host. Based on the above description, the following rule structure was derived from the PMU 2014 dataset and it is given in the Table 1. The proposed computational technique facilitates prompt detection and distinction of possible individual traffic records from crowd. There are 86781 normal and 19, 13, 219 DoS attack traffic in 100 percent labelled PMU 2014 data set. 4, 00, 219 Smurf attacks and 15, 13, 000 Neptune attacks are available in the 100 percent labelled PMU 2014. After removing irrelevant feature and duplicate feature from the data sets, 113 parameters was reduced to 12 parameters where around 88% of duplicated and irrelevant features were removed. Effectiveness of the IDS is evaluated by its ability to make correct predictions of events based on the predictor class as shown in table 3.

In the testing phase, the testing dataset is given to the proposed system, which classifies the input as a normal or attack. The obtained result is then used to compute overall accuracy of the proposed system. The overall accuracy of the proposed system is computed based on the definitions, namely precision, recall and F-measure which are normally used to estimate the rare class prediction. It is advantageous to accomplish a high recall devoid of loss of precision. F-measure is a weighted harmonic mean which evaluates the trade-off between them.

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F- Measure} = \frac{(\beta^2+1)(\text{Precision} \cdot \text{Recall})}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

Where,  $\beta = 1$

$$\text{Overall accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

Where,

*TP* - True positive

*TN* - True negative

*FN* - False negative

*FP* - False positive

These are computed using the confusion matrix in Table 2, and defined as follows:

**TABLE 2**

Active class	Predictor class	
	Normal	Attack
Normal	True Positives (TP)	False Positives (FP)
Attack	False Negatives (FN)	True Negatives (TN)

Where

**True negatives (TN):** indicates the number of normal events are successfully labeled as normal.

**False positives (FP):** refer to the number of normal events being predicted as attacks.

**False negatives (FN):** The number of attack events are incorrectly predicted as normal.

**True positives (TP):** The number of attack events are correctly predicted as attack.

The number of records taken for testing and training phase is shown in Table 3.

**TABLE 3**

Test Data	Training Data	Test Data
Normal	8600	5400
Smurf	4786	400
Neptune	56820	20500

The evaluation metrics are computed for testing dataset in the testing phase and the obtained result for all attacks and normal data are given in Table 4, which is the overall classification performance of the proposed system on PMU 2014 dataset. By analyzing the result, the overall performance of the proposed system is improved significantly and it achieves more than 99.5% accuracy for all types of attacks.

**TABLE 4**

DOS Attacks	Metric	Proposed System
		Accuracy
Normal	Precision	0.999814814
	Recall	0.999814814
	F measure	0.999814813
	Accuracy	0.999923954
Smurf	Precision	1
	Recall	1
	F measure	1
	Accuracy	0.9975062344
Neptune	Precision	0.999951219
	Recall	0.999951219
	F measure	0.999951219
	Accuracy	0.999902443

## V. CONCLUSION

In this paper, new hybrid based computational techniques were proposed for extracting the attack patterns available in the datasets. The result shows that enhanced GA reducing false alarm rate incorporates with SVM classifier. In this model irrelevant and redundant features are not recognized that brings down the processing speed of evaluating the known patterns. An efficient features selection model eliminates dimension of data, reduce redundancy and ambiguity caused by none important attributes. Hence, the performances of the proposed hybrid models are better than existing models. The proposed method performs the classification task and extract the recovered knowledge using GA-SVM. These systems are highly reliable, adequate interpretability and compare with several well known algorithms such as SVM, snort based hybrid system, Teaching Learning based Optimization IDS, Group Teacher Learning based Optimization IDS and Fuzzy logic IDS. The experiment results emphasized that the proposed hybrid models are suitable technique and produced better accuracy compared to the existing model. In future work, the octopus activities will be studied, use as a detection technique to find the patterns of the attacks and evaluate the performance with existing IDS.

## REFERENCES

- [1] J. Ding, *Advances in network management*: CRC press, 2010.
- [2] A. Iftikhar, A. Azween, A. Abdullah, A. Khalid, and H. Muhammad, "Intrusion detection using feature subset selection based on MLP, " *Scientific Research and Essays*, vol. 6, pp. 6804-6810, 2011.

- [3] H. M. Imran, A. Abdullah, M. Hussain, S. Palaniappan, and I. Ahmad, "Intrusion Detection based on Optimum Features Subset and Efficient Dataset Selection, " *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 2, pp. 265-270, 2012.
- [4] K. Delac, M. Grgic, and S. Grgic, "Independent comparative study of PCA, ICA, and LDA on the FERET data set, " *International Journal of Imaging Systems and Technology*, vol. 15, pp. 252-260, 2005.
- [5] S. Singh, S. Silakari, and R. Patel, "An efficient feature reduction technique for intrusion detection system, " in *Machine Learning and Computing, 2009. International Conference on*, 2011.
- [6] J. Li and C. Manikopoulos, "Early statistical anomaly intrusion detection of DOS attacks using MIB traffic parameters, " in *Information Assurance Workshop, 2003. IEEE Systems, Man and Cybernetics Society, 2003*, pp. 53-59.
- [7] D. Dutta and K. Choudhury, "Network Anomaly Detection using PSO-ANN, " *International Journal of Computer Applications*, vol. 77, 2013.
- [8] J. Kennedy, "Particle swarm optimization, " in *Encyclopedia of Machine Learning*, ed: Springer, 2010, pp. 760-766.
- [9] J. Yu, H. Lee, M. -S. Kim, and D. Park, "Traffic flooding attack detection with SNMP MIB using SVM, " *Computer Communications*, vol. 31, pp. 4212-4219, 2008.
- [10] J. Yu, H. Kang, D. Park, H. -C. Bang, and D. W. Kang, "An in-depth analysis on traffic flooding attacks detection and system using data mining techniques, " *Journal of Systems Architecture*, vol. 59, pp. 1005-1012, 2013.