IJICTR
International Journal of Information & Communication Technology Research

Volume 10- Number 1 – Winter 2018 (26-35)

Intrusion Detection System Using SVM as Classifier and GA for Optimizing Feature Vectors

Hossein Gharaee*
ICT Research Institute (ITRC)
Tehran, IRAN
gharaee@itrc.ac.ir

Maryam Fekri
Carleton University
Ottawa, Canada
maryamfekri@cmail.carleton.ca

Hamid Hosseinvand
Shahed University
Tehran, IRAN
hosseinvand@shahed.ac.ir

Received: 27 November, 2017 - Accepted: 18 March, 2018

Abstract—**Nowadays, IDS is an essential technology for defense in depth. Researchers have interested on IDS using data mining and artificial intelligence (AI) techniques as an artful. IDSs can monitor system behavior and network traffic until detect intrusive action. One of the IDS models is anomaly based IDS which trained to distinguish between normal and abnormal traffic. This paper has proposed an anomaly based IDS using GA for optimizing feature vectors and SVM as a classifier. SVM has used as a supervised learning machine that analyses data and recognize patterns, used for classification and regression analysis. After optimization best features for SVM, IDS can detect abnormal traffic more accurate. There is an innovation in fitness function which is formed from TPR, FPR and the number of selected features. The new fitness function reduced the dimension of the data, increased true positive detection and simultaneously decreased false positive detection. In addition, the computation time for training will also have a remarkable reduction. This study proposes a method which can achieve more stable features in comparison with other techniques. The proposed model has been evaluated test with KDD CUP 99 and UNSW-NB15 datasets. Numeric Results and comparison to other models have been presented.**

Keywords- Intrusion Detection System; Genetic; SVM; Feature Selection

I. INTRODUCTION

An intrusion detection system (IDS) has been developed to detect all types of network attacks in available environments. The IDS is placed inside the network to protect and collect network packets promiscuously in the same manner as a network sniffer. The IDS detects malicious network activities by analyzing the collected packets, alarms to system administrator, and blocks attack connections in order to prevent further damages. It also connects to firewall as a fundamental technology for network security [1]. IDS systems have divided in two main models: anomaly based and misuse based. In anomaly based intrusion detection systems try to detect malicious attempt based on deviation from a normal behavior. However in misuse based IDS, there is database of attack signatures that have been used to detect intrusive attempts. On the other hand, anomaly based IDS can detect zero day attacks because it classifies current behavior to normal and abnormal, but misuse based IDS just can detect attacks that happened earlier and their signatures are existent. Anomaly detection techniques are used to identify outliers i.e. events or observations which are not matching with parameters that obtained before. In misuse detection techniques, dataset is comparing with predefined signatures and these signatures are derived from some set of rules to avoid attacks. It is also called as signature-based technique. The main problem with misuse based IDS is that they fail to detect new attacks whose signatures are not present while the anomaly based techniques are adaptive in nature as they can identify novel attacks [2].

The major benefit of anomaly-based detection methods is that they can be very effective at detecting previously unknown threats. For example, suppose a computer becomes infected with

* Corresponding Author

a new type of malware. The malware can consume the computer's processing resources, send large numbers of e-mails, initiate large numbers of network connections, and perform other tasks that may be significantly different from the established behavioral profile for the computer.

Anomaly based IDS must be trained to distinguish between normal and abnormal activity so that can detect attack traffic[3, 4].This can be accomplished in several ways, some researchers proposed some methods to define the normal usage of the system using a mathematical model, and flag any deviation from this as an attack, or IDS models using neural networks, most often with artificial intelligence techniques used for data mining to search the search space to find anomalies[5]. Disadvantage of using such learning algorithms is the huge dimensions of the search space. Another popular evolutionary method is GA (GA) which has high potential of finding the best solution in a search space[4, 6].

GA is a search heuristic that generates useful solutions to optimize search problems. It is one of the powerful algorithms based on evolutionary ideas of natural genetics which generate population of chromosomes as solutions of problem. The most important component of GA is the fitness function which evaluate the chromosomes. A good fitness function would help GA to find closer subset of chromosomes to intended result. Previous proposed IDSs based on GAs use two factor of classification accuracy and numbers of selected features. The major weakness of former models was evaluating the feature chromosomes just based on accuracy or true positive rate of the classification, on the other hand one of vital challenges of IDS models is high false alarm rate as it is not considering in former IDS models. Also In order to decrease the dimension some feature selection techniques have been proposed[7].

Feature selection is a process of selecting subset of relevant features for use in model construction. Feature selection causes simplification of models to make them easier. In this paper we use GA as feature selection technique to find most optimum feature set. Most optimum feature set help us to find anomalies with more accuracy.

Data mining-based classification approaches for intrusion detection have received accolades, the lack of published research in applying rough set based feature selection, enhancing the discriminant function performance in SVM to IDSs seems to be an oversight in intrusion detection. In spite of the earlier works in increasing the performance of IDSs, the overall performance of the IDS certainly needs improvement[8]. Feature selection as a case of study recently has become the important phase of improvement of an IDS as it effects performance of a classifier. Feature selection techniques can powerfully identify a subset of features within a dataset and reduce the number of fields, in order to decrease the time for computation process[9]. In other words, not all of the features are important or related for detecting an intrusion. Therefore some noisy, irrelevant and redundant features can be discarded. Generally, there are two methods for feature selection: filter and wrapper methods[10]. The *filter* method estimate classification performance with indirect assessment and does not depend on classifier performance. In contrast, *wrapper* methods depend on the classifier efficiency, and the evaluation of the selected features is calculated directly from the accuracy of the classifier [11,12].

GA as a *wrapper* based method searches for the best solution which best improve the classifier. *Filter* methods even with the best feature subset do not necessarily guarantee high classification accuracy for any type of dataset. Wrapper methods can reach better accuracy and high classifier performance[13]. So this trend will effect accuracy, needed time for learning, number of samples needed for learning phase, and also false alarm rate of the classifier. However high computational complexity may cause some limitation on their application.

This paper has developed an IDS using GA as an optimizing feature selection. The new feature selection method has the potential to generate optimal feature subset considering challenges of former IDSs. The first novelty is proposing a new fitness function for the GA, in order to decrease the false alarm rate and increasing the true positive rate simultaneously and also minimizing number of features to enhance low learning and computation time. The second novelty is combining GA with Support Vector Machine (SVM) to detect anomalies. The new fitness function of the GA evaluates feature chromosomes considering their effectiveness on True and False Positive rates by using a SVM as a supervised learning classifier. SVM is a good candidate for a classifier, because of its training speed and scalability[14]. The earlier studies have also shown that Least Squares SVM (LSSVM) with RBF kernel has an appropriate performance and has higher detection accuracy against SVM because LSSVM has solved the local optima problem[15]. This article has developed the IDS using LSSVM and GA with the new fitness function and prepare the best optimal feature subsets and detect intrusions with the maximum TPR, minimum FPR and low computation time. This paper also compares such results with outputs of other feature selection techniques. The selected features can also be used by other classifiers in the IDSs.

The rest of the paper is organized as follows: section II the related works on IDS models are reviewed, section III is a basic concepts to GA and SVM, section IV proposes Proposed intrusion detection system based on GA and SVM, V and VI shows analysis of simulation results and conclusions respectively.

II. Related Works

Current IDSs use many techniques. Misuse based IDSs which mentioned before as signature based techniques are still use for some purposes. A combination of genetic fuzzy system and pairwise learning as a misuse IDS has proposed in [16]. The model has a learning stage which uses fuzzy rules to set features and evaluate them and finally classification and detection. Detection rate in all attack type classes have been maximized as it obtains a better separable between normal and attack. However False Alarm Rate has not been considered in the experiments.

Some of these techniques are widely used for anomaly based intrusion detection which are statistical [17], hidden Markov model[18], artificial neural network[19, 20], machine learning [21, 22] and GA[23-24].

There are some methods based on machine learnings which are a better solution for distinguish between normal and abnormal traffic activity[25]. Learning techniques proposed for different intrusion detection problems, can be hardly classified into two categories: supervised and unsupervised[26]. Supervised models start with a labeled dataset to train then classify the data but unsupervised models try to find hidden structure in unlabeled data.

A model with GA proposed as search strategy and logistic regression as learning algorithm to select best subset of features [24]. Another IDS based on GA and SVM has improved the parameters of SVM. This model uses Genetic as an optimization algorithm to maximize the performance of the SVM. Average Detection accuracy rate has been noted as 80.14% by this model [27]. Another article has proposed similar model of IDS using Genetic to optimize SVM's parameters and also as a feature selection. Fitness function which has developed for the GA in the paper has evaluated chromosomes with the maximum accuracy and minimum number of features [28]. Although, in results the false positive rate has not been considered.

A new SVM model in which kernel principal component analysis (KPCA) is combined with GA is proposed as a supervised learning method. A multi-layer SVM classifier is suggested in the paper to detect intrusions and KPCA is used as a feature selection technique to decrease the dimension of feature vectors also reducing the training time. GA is employed to optimize the SVM parameters. In comparison with other detection algorithms, the results show better accuracy, but still has high false positive rate which is a real concern in nowadays intrusion detection models[29]. Thecombination of Genetic and Fuzzy SVM on cloud computing network has been proposed which has improved the detection rate with a feature selection method to 98.51% and has reduces the learning time [30].

A model based on heuristic genetic and neural network is proposed in order to enhance better performance of intrusion detection, in which input features, network structure, and connection weights were all considered jointly in fitness function. This article has enhanced 98.28% and 96.39% accuracy for DOS and PROBE attack and low rate of detection for R2L and U2R 60.32% and 55.17%. The overall FP has mentioned 1.14% which is an average rate in all attack types[31].

Another supervised learning method which is driven from decision tree algorithm and artificial neural network has developed. The flexible neural tree (FNT) model can reduce the number of features [32]. Using 41 features, the best accuracy for the DOS and U2R is given by the FNT model. The decision tree classifier supplied the best accuracy for normal and probe classes, which are a little better than the FNT classifiers.

One of the wrapper based feature selection method is a multi-objective optimization algorithm and also used an unsupervised clustering method based on Growing Hierarchical Self-Organizing Maps (GHSOMSs). SOM has mentioned as one of the most used artificial neural network models for unsupervised learning. The research has selected 25 features for classification and has shown rate of classification accuracy as 99.12 ± 0.61% and the FP rate as 2.24 ± 0.41 %. The paper shows improvement in comparison between IDS models with filter based feature selection and IDSs without feature selection [33].

In addition to such studies, due to high dimensionality of network data, several IDS, in which feature selection is used as a pre-processing phase, have been developed [31]. The feature selection process first remove one input feature from the data; the remaining data set is then prepared for utilizing with classifier. Then, the classifier's performance is compared to that of the original classifier in terms of performance criteria.

A feature selection algorithm based on mutual information and LSSVM (MMIFS) has presented in [34]. Performance of MMIFS has been compared with other mutual information based feature selection approaches and filter based feature selection methods. This research has shown the performance of selected features in terms of TP and FP rate in charts and also has shown first ranked features in each attack class. However, statistical methods just estimate the effect of every feature on detection not extracting the exact result of detection accuracy so they cannot provide a clear and reliable result.

Another intrusion detection model based on Genetic and Neural Network has developed by [35] which has used Genetic for feature selection with a new coding of chromosomes. Each chromosome contains index of each ranked feature instead of binary gens. A new fitness function is developed with new measure to compute the information gain provided by each features subsets. Then the performance of the selected feature subset is tested using Neural Network as a classifier. The fitness function calculate the information needed to classify a given sample by:

$$I(s_1, s_2, \dots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \, log2(\frac{s_i}{s})$$

S is training set samples with their labels. Feature F with values { f_1, f_2, …, f_v } can divide the training set into v subsets { S_1, S_2, …, S_v} where S_j is the subset which has the value f_j for the feature F[35].

Among the statistical approach for feature selection as they are categorized in filter feature selection methods, there are wrapper method approaches which has been proposed and have come to high accuracy in detecting intrusions and also results are accurate and clear to be relied [36].

The model uses GA and a chromosome of features and SVM parameters to optimize both [14]. Experiments on UCI database evaluate the model and shows high accuracy than other prier models in the field of detecting heart disease and cancer. The fitness function allows the chromosome with the maximum classification accuracy and minimum number of features. The fitness function used in the model is as follows:

$$F = W_a \times A + W_f \times (P + (\sum_{i=1}^{nf} C_i \times F_i))^{-1} \quad (1)$$

Which A is the accuracy of classification, F is the value of i^{th} gen in a chromosome which is 0 or 1 indicating i^{th} feature has been selected or not. C is the weight of i^{th} feature, P is a constant value, W_a and W_f are the weight of accuracy and number of selected features.

A detection model of human log intergenic in medical dataset based on Genetic and SVM has developed in [37]. The fitness function is calculating the classification accuracy. Other GA based on detection model with new evaluation function has proposed for optimizing features by considering their confidence rate [38]. In the coding of chromosomes, each gen is related to its confidence rate so the chance to select the feature is based on its confidence rate. The algorithm use classification accuracy and number of selected feature in fitness function:

$$F = f(V) - |V|/|U|$$

V is the confidence rate of each feature.

In the field of detecting intrusions, beside accuracy, false positive detections are also important so we cannot just consider the accuracy in fitness to be the survival.

An ideal IDS system has to detect all attack types. This means to detect each attack correctly and not considering any normal activity as an attack activity which leads to high false alarm rate in an IDS. So in addition to detecting all attack types correctly IDS should also detect all normal activities correctly and not as an attack type. In this paper we review the issue and the proposed model solves the problem.

In this paper, an intrusion detection system is proposed which differs from existing work in many ways. First, the pre-processing technique based on GAs is used which intelligently performs as an optimization. This technique, as a more precise way, is a wrapper based feature selection approach that can calculate true positive rate and false positive rate in order to compare chromosomes directly. Second, in this article, a new Fitness function for the above mentioned GA is offered for the sake of higher accuracy along with low false positive rate with the selected features. Third, the model has combined GA with SVM to classify and detect intrusions.

III. Basic Concepts in GA and SVM

GA is a general adaptive optimization search methodology based on analogy to Darwinian natural selection and genetics in biological systems[14].

According to Darwinian principal of "survival of the fittest", GA works with a fitness function and series of solutions called population and try to reaches the optimal solution by evaluating each individual's fitness. Each chromosome in the population with the higher fitness value has more chance to be kept in the next generation population. Crossover and mutation functions operate on chromosomes and directly impact the fitness values.

A. Genetic operations

During the breeding genetic operations applies on population to reproduce next generation of the population. Crossover, mutation and selection are the three vital operations.

- Selection operator, gives another chance to better chromosomes, by selecting them based on the better fitness value and passing them to the next generation. As the generation pass, members of the population get closer to the best solution.
- Crossover operator, as shown in the Fig1, exchange genes between tow chromosomes. This operation makes children have combination of features from their parents not exactly same as them which leads the algorithm to explore more solutions in the search space.
- Mutation operator, as shown in Fig1, operate on one chromosome and alter a gen, for example in binary GA, gene value may change from 1 to 0.

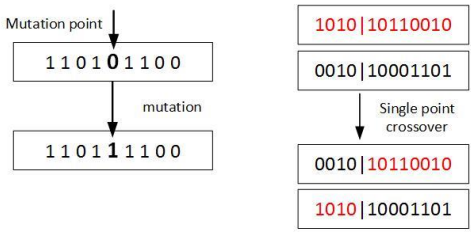

Fig. 1. Mutation and crossover function on feature chromosomes

In this paper we use GA also as a feature selection which can best output a feature chromosome or feature subset with optimal fitness function. Genetic also works with a classifier to help better detection with high performance in order to achieve high accuracy as an IDS system.

B. Support Vector Machine (SVM)

SVM is one of the supervised learning methods which is used for classification. This machine has learning algorithm that can be used as a pattern matching machine for data classification. Let assume that we have series $\{(x_1, c_1), (x_2, c_2), \dots, (x_n . c_n)\}$ and we want to divide them to two classes $c_i = \{-1,1\}$. Linear classification methods, try to separate data with classifier line (that is a liner equation). Svm's Classification method tries to separate data from two classes with finding the best linear equation, in form that be the maximum margin space between two classes. If input data are nonlinear we use kernel for data classification [6]. With regard to function theory: Kernel is a function that is positive and definite and in addition satisfies the mercer condition. By use of kernel, data are mapped to higher dimension space. Some of kernel functions are as follow:

-Polynomial (homogeneous)

$k(x_i, x_j) = (x_i . x_j)^d$

-Polynomial (homogeneous)

$k(x_i, x_j) = (x_i . x_j + 1)^d$

-Gaussian radial basis function

$k(x_i, x_j) = \exp(-\gamma \left|\left|x_i - x_j\right|\right|^2)$, for $\gamma > 0 . \gamma = 1 / 2\sigma^2$

-Hyperbolic tangent

$k(x_i, x_j) = \tanh(k x_i . x_j + c)$, $\kappa > 0$ و c

SVM performance with Gaussian kernel is much better than other kernels in intrusion detection systems because of high dimension of input data [39].

In this paper we use Least Square SVM (LSSVM) that is a regularized reformulation to the standard SVM. A linear equation has to be solved in the optimization stage, which not

only simplifies the process, but also is effective in avoiding local minima in SVM problems. In other words, for binary-class classifications, SVM make an optimal separating hyper plane with the maximum margin between the two classes (positive and negative). It can be formulated as a quadratic programming problem involving inequality constraints whereas the LSSVM involves the equality constraints only. Hence, the solution is obtained by solving a system of linear equations. Experiments shows that LSSVM performance with Radial basis function (RBF) kernel is better than SVM with Gaussian kernel in intrusion detection systems since LSSVM, has improved SVM problems placed in local minima[40,41]

IV. PROPOSED INTRUSION DETECTION SYSTEM BASED ON GA AND SVM

Reviewing the model of [14] in which has used GA and LSSVM to develop a detection system with feature selection in terms of diagnosing cancer and heart disease through medical datasets. The proposed Genetic feature selection SVM (GF-SVM) model has retrieved from the model in [14] but to detect intrusions through network traffic.

System architecture of our proposed GA based feature selection with new fitness function using LSSVM is shown in Fig. 3. The proposed model contains three main steps as it has shown in Fig. 2.

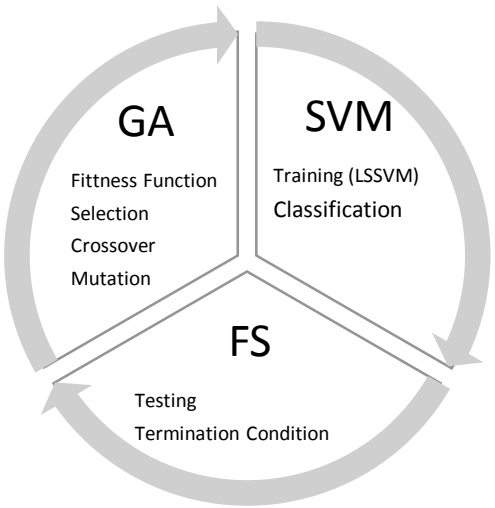

Fig. 2. A representation for the proposed GF-SVM IDS.

1- Feature selection method based on GA: Input is the traffic data then creating feature chromosomes and evaluating each chromosome and then select chromosomes with the highest classification accuracy. Output of this step is the optimized feature chromosomes for each attack type traffic and normal traffic then the selected features will apply on network traffic data. In the next section, we will discuss in more detail and also fig 4 illustrate the process in a flowchart.
2- Training: is the first step of detection which the LSSVM will be trained with the training data.in fact the support vectors will create based on classification of the training traffic. Support vectors should be created precisely in order to reach more accurate detection on the next step. It has been widely discussed earlier in section IV about how support vector will be created.
3- Classification: classifying the traffic data in to two class of normal and anomaly.

A. *Feature selection based on GA*

In this paper we use GA as basis of our proposed feature selection algorithm with feature chromosome. Fig. 3 illustrates the flow chart of the proposed GA feature selection and will continue discussing about each level.

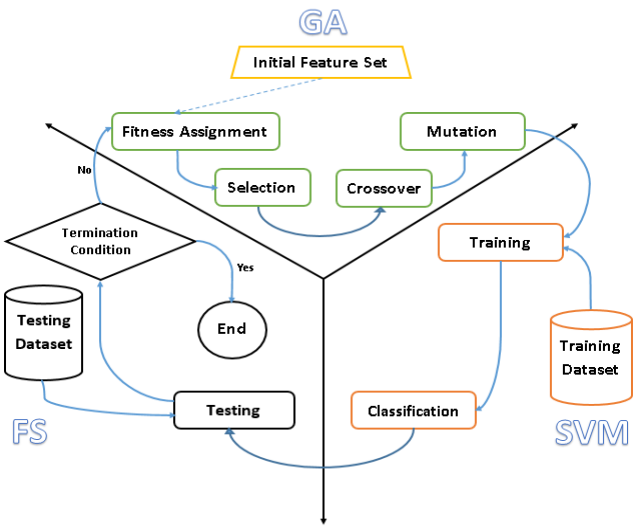

Fig. 3. Feature selection based on Genetic

B. *Chromosome coding*

we code selected feature subset in to a binary code which named F. in the binary coding, 1 indicated that the feature index is selected; 0 indicates that feature with that index is not selected.

TABLE I. CODING OF FEATURE CHROMOSOME

Feature chromosome 1	11010111110…011010

C. *initial population*

Initial population which consist of N parent chromosomes is generated by prior research results in each type of attack. This paper use the crucial features to create chromosomes of the initial population for each attack class. The size of the population should be suitable, if it is too large, then the complexity of the algorithm would be too high and if it is too small then optimal performance of the algorithm is reduced and algorithm may probably get stuck in to local optima solution easily. The size range between 100 and 200 is suitable for the number of records selected from the database in this problem

D. *Train LSSVM classifier*

LSSVM uses in all three steps of the GF-SVM model. In first step LSSVM is for evaluating feature chromosomes and helps the GA to decide about selecting any feature of the 41 features. Classifier trained by training dataset with selected features. Classification accuracy also calculated in order to evaluate how selected features can effect to reach higher accuracy. In step two

and three LSSVM use to finalize the results of the model and to show the effectiveness of the features in detecting intrusions.

E. Proposed fitness function

Fitness function is the basis component of GA to evaluate whether an individual is fit to survive. In further research, researchers use accuracy and number of features as the two important parameters for fitness function and evaluate each feature subset. As the general performance of an IDS depends on how truly detect intrusions and also how not wrongly diagnoses attacks, we assume that important parameters of detecting intrusions in security issues are not only the true positive detections or classification accuracy but also false positive detection are as important as the true positive detections. Because False Positive detection results high false alarm rate, therefore general performance of the IDS will reduce. Previously researchers select the subset which has the high classification accuracy and low number of features but they did not consider the false detections so the feature subset causes high false alarm rate and the performance of the IDS decreases.

The novelty of this paper is proposing a new fitness function which uses 3 parameters named True Positive Rate (TPR), False Positive Rate (FPR) and number of selected features to evaluate each subset of features. A single objective fitness function that consist of 3 goals in to one has designed to solve the multiple criteria problems. The formula is as below:

$$Fittness(S) = W_a * TPR - W_b * FPR + W_c * NumF(S) \quad (2)$$

Where W_a, W_b and W_c are the weight of TP, FP, and weight value for the number of selected features respectively. TP is the rate of the True Positive detections with the selected subset (f) which calculated by:

$$TP\ rate = \frac{TP}{TP + FN} \quad (3)$$

And FP is the rate of the false positive detection with the selected features (f) which calculated by:

$$FP\ rate = \frac{FP}{FP + TN} \quad (4)$$

Generally W_a and W_b can be set from 75 to 100% according to user's requirement. In our study we set W_a to 40% and W_b to 50% and W_c 10% which cause experiments to an optimized result of high TP, low FP with small subset of selected features.

True positive rate of SVM, False positive rate of the SVM, and the number of selected features are used to construct a fitness function. Every chromosome is evaluated by fitness function as (2).

F. Selection and stopping criteria

After evaluating all chromosomes of a population, there is a fitness value for each chromosome and each with the highest fitness value has survived. We use genetic operator which mentioned before, on the high fitness value chromosomes to create the new generation. Termination condition of the algorithm is when the maximum number of iteration reaches. When the ending condition is satisfied, the operation ends; otherwise, we proceed with the next generation operation.

V. ANALYSIS OF SIMULATION RESULTS

The datasets used in these experiments is ''KDD Cup 1999 data and UNSW-NB15[42]", well-known sample traffic datasets. Each record of two datasets that mentioned before is unique with 41 and 47 of continuous and nominal features plus one class label. In this paper, the nominal feature such as protocol (TCP/UDP/ICMP), service type (http/ftp/telnet/y) and TCP status flag (sf/rej/y) have been converted into a numeric feature[43]. The KDD cup 99 data set contains 24 attack types that has been categorized in four groups: Probe, Denial of Service (DOS), User to Root (U2R) and Remote to User (R2L) [43]and UNSW-NB15 data set traffic has categorized in 9 groups: Normal, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms[42].

In experiments there is another parameter which is calculated and is compared with other researcher's results. Equation (5) shows how to calculate accuracy:

$$Accuracy\ (A) = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

A. Results and discussion

Experiments were performed on a Windows platform having configuration Intel core i7 2.3GHz, 8 GB RAM. We have used the open toolbox LS-SVM lab[2] to implement LSSVM and modeling the IDS. The proposed model in this paper first selects the best feature subset in each class in terms of resulting the highest classification accuracy and true positive rate with the lowest false positive rate. Then collects the selected features in each class and gives priority based on the repetition of the features.

Results of the important features in each class type for KDD CUP 99 and UNSW-NB15 dataset has been collected in Table II, that show the number of important features selected by 3 feature selection algorithms and the labels of these important features.

TABLE II. SELECTED FEATURES FOR TWO DATASETS

Result with KDD CUP 99 Dataset			
Class Type	Method	Features Number	Selected Features
Normal	GF-SVM	7	3,10,6,7,33,36,12
	MMIFS	6	5, 23, 3, 6, 35,1
	SVM-SADT	19	1,3,5,6,8,10,11,12,13,22, 23,25,26,27,28,29,32,35,39
DoS	GF-SVM	10	5,3,19,23,4,21,22,24,2,8
	MMIFS	8	2,3,5,6,23,24,36,41
	SVM-SADT	17	2,5,8,9,10,11,12,13,23,25, 26,29,33,5,36,39,41
PROBE	GF-SVM	9	3,35,4,40,36,30,37,34

[1] www.esat.kuleuven.ac.be/sista/lssvmlab

	MMIFS	13	40, 5, 33, 23, 28, 3, 41, 35, 27,32, 12, 24, 28
	SVM-SADT	17	1,2,3,4, 5,6,7,8,23,25,30 ,32,33,35,36,39,41
R2L	GF-SVM	7	5,8,28,29,36,39,40
	MMIFS	15	3, 13, 22, 23, 10, 5, 35, 24, 6, 33, 37, 32, 1, 37, 39
	SVM-SADT	24	1,3,5,6,7,11,12,13,17,18,19,20,21, 22,23,24,25,28,29,32,35,36,37,38,
U2R	GF-SVM	8	3,13,6,14,16,23,32,1
	MMIFS	10	5, 1, 3, 24, 23, 2, 33, 6, 32, 4,14,21
	SVM-SADT	15	13,14,15,16,5,11,10,22, 33,39,34,6,12,32,36
Result with UNSW-NB15 Dataset			
Class Type	Method	Features Number	Selected Features
Normal	GF-SVM	7	1,2,15,18,21,29,31
Fuzzers	GF-SVM	13	2,4,10,14,28,29,31, 41,43,44,45,46,47
Reconnaissance	GF-SVM	14	10,14,19,20,27,30,31, 34,42,43,44,45,46,47
Shellcode	GF-SVM	9	4,10,14,23,37,44,45
DoS	GF-SVM	12	10,13,14,15,17,23,31,42, 43,44,45,47
Exploits	GF-SVM	6	13,14,16,17,31,33
Generic	GF-SVM	9	10,11,19,23,28,31,33,34,46
Analysis	GF-SVM	6	8,10,14,20,24,41
Backdoor	GF-SVM	11	5,10,12,13,14,15,23,41,43,45,47
Worm	GF-SVM	11	1,6,7,8,12,18,22,29,31,42,46

Fig 4–7 show the comparison between the proposed model and MMIFS model [34] in Receiver Operating Characteristic (ROC) curves. The ROC curves illustrate true positive and false positive rate of each step adding features in order of their importance. These results are belong to KDD CUP dataset. For example Fig. 5 shows, the red line for the proposed GF-SVM model is almost near 100% true positive rate and also the value of false positive rate is not going over 2%.

Because there is no related paper that used the UNSW-NB15 dataset for its method testing yet, we just show you the results of our proposed GF-SVM model over new dataset. As I write before, there are some effective parameters in fitness function that have direct effect on detection accuracy. One of these important parameters is W_b; the weight value for FP. Fig 9-12 show you detection accuracy rates for variety W_b values.

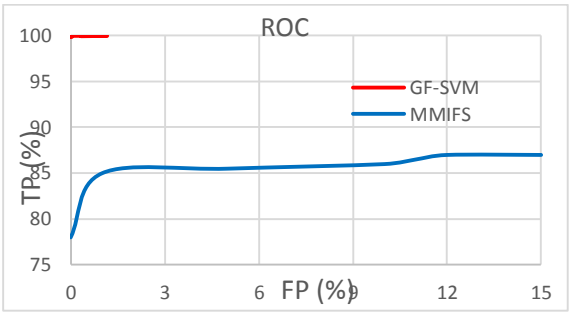

Fig. 4. ROC Curve for the DOS class is shown using two models

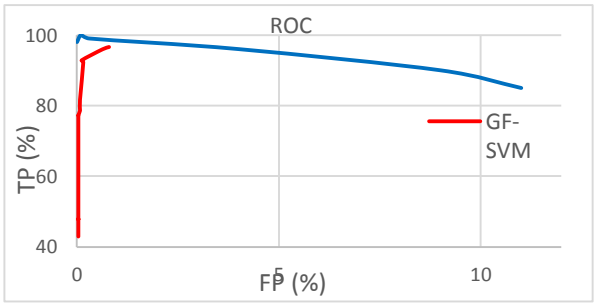

Fig. 5. ROC Curve for the PROB class is shown using two models

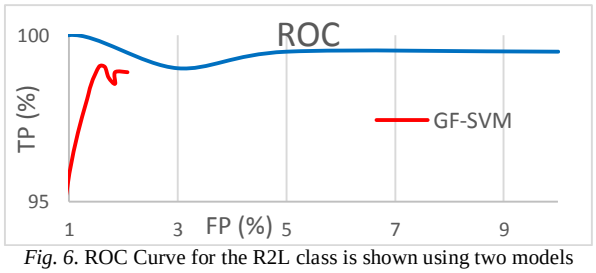

Fig. 6. ROC Curve for the R2L class is shown using two models

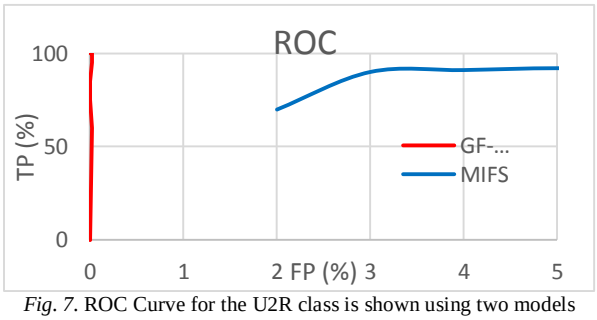

Fig. 7. ROC Curve for the U2R class is shown using two models

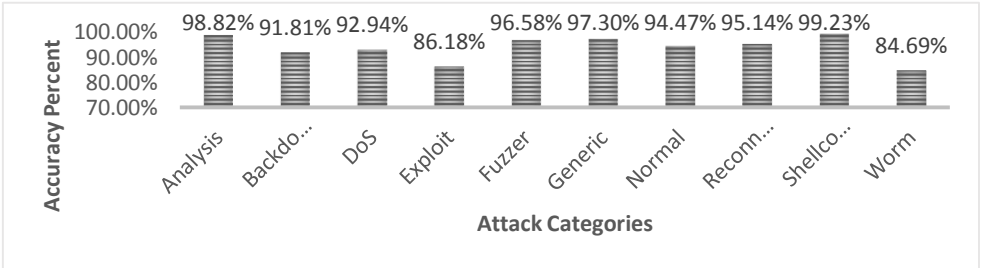

Fig. 8. Detection accuracy rate for W_b = 10

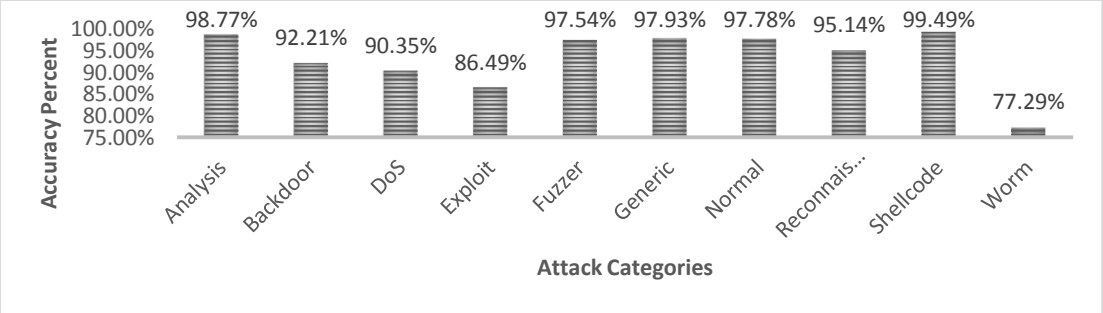

Fig. 9. Detection accuracy rate for W_b = 7

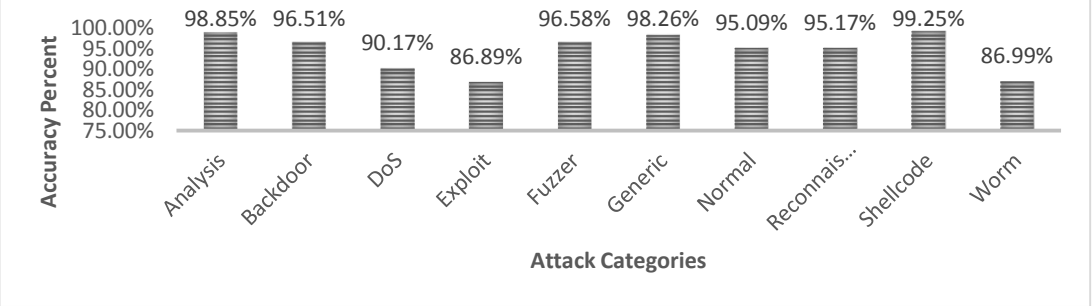

Fig. 10. Detection accuracy rate for W_b = 5

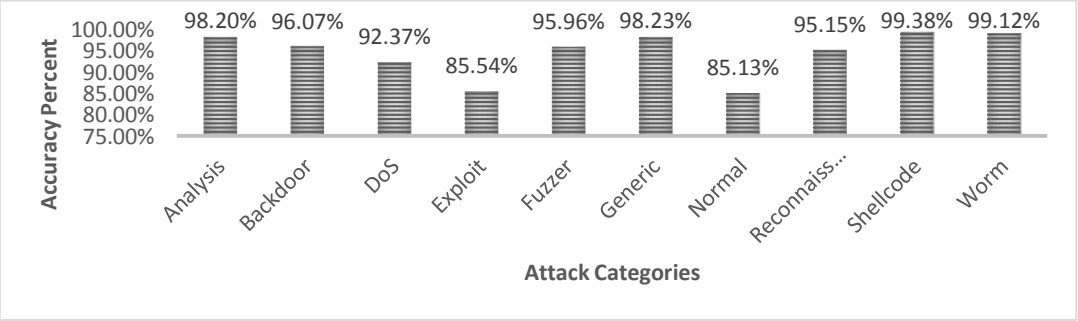

Fig. 11. Detection accuracy rate for W_b = 3

B. *Comparison GF-SVM with other Techniques*

This section shows performance comparison of our proposed GF-SVM model with four other intrusion detection techniques introduced such as MMIF[34], GA-Fuzzy SVM[44], C4.5 [16] and SVM[45]. In order to compare the methods, we applied the MMIFS model in the LSSVM model and obtained the numeric results of the Table III. Other numeric results of the other models are driven exactly from the earlier papers. The results of our model for new UNSW-NB15 dataset are showed in Table III too.

TABLE III. PERFORMANCE

Result with KDD CUP 99 Dataset				
class	Model	Accuracy (%)	TP (%)	FP (%)
Normal	**GF-SVM**	**99.05**	**98.90**	**0.8**
	MMIFS [34]	90.87	98.48	16.81
	GA-Fuzzy SVM [44]	-	98.75	1.25
	C4.5 [16]	99.79	-	-
DOS	**GF-SVM**	**99.95**	**99.97**	**0.06**
	MMIFS [34]	99.86	99.90	0.17
	SVM [45]	98.80	100	24.08
	GA-Fuzzy SVM [44]	-	98.3	2.7
	C4.5 [16]	99.68	-	-
PROBE	**GF-SVM**	**99.06**	**98.2**	**0.079**
	MMIFS [34]	96.99	94	0.039
	SVM [45]	97.31	95.62	24.08
	GA-Fuzzy SVM [44]	-	96.53	1.47
	C4.5 [16]	96.14	-	-
R2L	**GF-SVM**	**98.25**	**90.26**	**0.64**
	MMIFS [34]	94.95	63.42	0.49
	SVM [45]	97.51	65.79	24.08

	GA-Fuzzy SVM [44]	-	85.48	1.45
	C4.5 [16]	85.65	-	-
U2R	**GF-SVM**	**100**	**100**	**0**
	MMIFS [34]	99.91	80	0
	SVM [45]	97.52	52.38	24.08
	GA-Fuzzy SVM [44]	-	89	1.1
	C4.5 [16]	57.69	-	-

Result with UNSW-NB15 Dataset

class	Model	Accuracy (%)	TP (%)	FP (%)
Normal	**GF-SVM**	**97.78**	**99.04**	**0.04**
Fuzzers	**GF-SVM**	**97.54**	**98.69**	**0.03**
Reconnaissance	**GF-SVM**	**95.14**	**93.88**	**0.03**
Shellcode	**GF-SVM**	**99.49**	**100**	**0.09**
DoS	**GF-SVM**	**92.37**	**92.40**	**0.07**
Exploits	**GF-SVM**	**86.49**	**84.40**	**0.10**
Generic	**GF-SVM**	**98.23**	**98.25**	**0.01**
Analysis	**GF-SVM**	**98.85**	**99.84**	**0.12**
Backdoor	**GF-SVM**	**91.81**	**93.13**	**0.24**
Worm	**GF-SVM**	**77.29**	**77.20**	**0.12**

VI. CONCLUSIONS

We presented an intrusion detection approach that took advantage of discriminating properties of GA. We have considered GA and SVM where we introduced a feature selection method and detection method that improved the intrusion detection performance. The feature selection procedure with our new fitness function performed a better selection of feature set, by taking true and false positive rate into account, but this is not enough. According to studies in related work section, nonlinear support vector machines based on RBF kernel as future research will help us to take better result.

The experiments and numeric results developed with KDD CUP 99 and UNSW-NB15 datasets showed efficiency through optimization in classification accuracy, FP, TP rates and ROC curves. The results obtained for GF-SVM model showed 99.05 % detection accuracy for normal traffic class, 99.95% for DOS class, 99.06% for PROBE class, 98.25% for R2L and 100% for U2R in KDD CUP 99. In addition, this paper obtained results equal to 97.45 % for Normal, 96.39 % for Fuzzers, 91.55 % for Reconnaissance, 99.45 % for Shellcode, 91.24 % for DoS, 79.19 % for Exploits and 97.51 % for Generic class in UNSW-NB15 Dataset.

REFERENCES

[1] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," Expert Systems with Applications, vol. 41, no. 4, pp. 1690-1700, 2014.

[2] Chand, Nanak, et al. "A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection." Advances in Computing, Communication, & Automation (ICACCA)(Spring), International Conference on. IEEE, pp. 1-6, 2016.

[3] J. D. a.-V. P. Garcı´a-Teodoro, G. Macia, Ferna´ndez, E.Va´zquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," computers & security 28, 2009.

[4] M. F. Umer, M. Sher, and Y. Bi, "Flow-based intrusion detection: techniques and challenges," Computers & Security, 2017.

[5] M. Moorthy, and S. Sathi yabama, "A study of Intrusion Detection using data mining." pp. 8-15, 2012.

[6] S. Ganapathy, K. Kulothungan, S. Muthurajkumar, M. Vijayalakshmi, P. Yogesh, and A. Kannan, "Intelligent feature selection and classification techniques for intrusion detection in networks: a survey," EURASIP Journal on Wireless Communications and Networking, vol. 2013, no. 1, pp. 271, 2013.

[7] M. S. Abadeh, H. Mohamadi, and J. Habibi, "Design and analysis of genetic fuzzy systems for intrusion detection in computer networks," Expert Systems with Applications, vol. 38, no. 6, pp. 7067-7075, 2011.

[8] Reddy, R. Ravinder, Y. Ramadevi, and KV N. Sunitha. "Effective discriminant function for intrusion detection using SVM." Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on, pp. 1148-1153, IEEE, 2016.

[9] G. Chandrashekar, and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28, 2014.

[10] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, and K. Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," Expert Systems with Applications, vol. 39, no. 1, pp. 424-430, 2012.

[11] A. Fahad, Z. Tari, I. Khalil, I. Habib, and H. Alnuweiri, "Toward an efficient and scalable feature selection approach for internet traffic classification," Computer Networks, vol. 57, no. 9, pp. 2040-2057, 2013.

[12] A. Fahad, Z. Tari, I. Khalil, A. Almalawi, and A. Y. Zomaya, "An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion," Future Generation Computer Systems, vol. 36, pp. 156-169, 2014.

[13] C.-F. Tsai, W. Eberle, and C.-Y. Chu, "GAs in feature and instance selection," Knowledge-Based Systems, vol. 39, pp. 240-247, 2013.

[14] M. Zhao, C. Fu, L. Ji, K. Tang, and M. Zhou, "Feature selection and parameter optimization for support vector machines: A new approach based on GA with feature chromosomes," Expert Systems with Applications, vol. 38, no. 5, pp. 5197-5204, 2011.

[15] L. L. Zhong, Z. Y. Ming, and Z. Y. Bin, "Network intrusion detection method by least squares support vector machine classifier." pp. 295-297.

[16] S. Elhag, A. Fernández, A. Bawakid, S. Alshomrani, and F. Herrera, "On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on Intrusion Detection Systems," Expert Systems with Applications, vol. 42, no. 1, pp. 193-202, 2015.

[17] Motai, Yuichi. "Kernel association for classification and prediction: A survey." IEEE transactions on neural networks and learning systems, vol 26, no 2, pp.208-223, 2015.

[18] Pathak, Priya, et al. "HMM-Based IDS for Attack Detection and Prevention in MANET." Information and Communication Technology for Sustainable Development. Springer, Singapore, pp. 413-421, 2018.

[19] Hodo, Elike, et al. "Threat analysis of IoT networks using artificial neural network intrusion detection system." Networks, Computers and Communications (ISNCC), 2016 International Symposium on. IEEE, pp.1-6, 2016.

[20] Kim, Jihyun, et al. "Long short term memory recurrent neural network classifier for intrusion detection." Platform Technology and Service (PlatCon), 2016 International Conference on. IEEE, pp. 1-5, 2016.

[21] Eberhardt III, John S., Todd A. Radano, and Benjamin E. Peterson. "Application of machine learned Bayesian networks to detection of anomalies in complex systems." U.S. Patent No. 9,349,103. 24 May 2016.

[22] Aljawarneh, Shadi, Monther Aldwairi, and Muneer Bani Yassein. "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient

model." Journal of Computational Science 25, pp. 152-160, 2018.

[23] H. Gharaee, and H. Hosseinvand, "A new feature selection IDS based on GA and SVM." pp. 139-144.

[24] H. Gharaee, M. Fekri, "A New Feature Selection for Intrusion Detection System," Int. J. of Academic Research, Jul. 2015 vol. 7, no. 4 pp. 48-60

[25] C. Khammassi, and S. Krichen, "A GA-LR Wrapper Approach for Feature Selection in Network Intrusion Detection," Computers & Security, 2017.

[26] Y.-F. H. Chih-Fong Tsaia, Chia-Ying Linc, Wei-Yang Lin, "Intrusion detection by machine learning: A review," Expert Systems with Applications , doi:10.1016/j.eswa.05.029, 2009.

[27] P. Sangkatsanee, N. Wattanapongsakorn, and C. Charnsripinyo, "Practical real-time intrusion detection using machine learning approaches," Computer Communications, vol. 34, no. 18, pp. 2227-2235, 2011.

[28] P. M. Mafra, V. Moll, J. da , S. Fraga , and A. Olivo Santin, "Octopus-IIDS: An anomaly based intelligent intrusion detection system."

[29] L. Zhuo, J. Zheng, X. Li, F. Wang, B. Ai, and J. Qian, "A GA based wrapper feature selection method for classification of hyperspectral images using support vector machine." pp. 71471J-71471J.

[30] F. Kuang, W. Xu, and S. Zhang, "A novel hybrid KPCA and SVM with GA model for intrusion detection," Applied Soft Computing, vol. 18, pp. 178-184, 2014.

[31] K. R. A.M.Chandrasekhar, "Intrusion detection technique by using K-means, Fuzzy neural network and SVM classifier," International Conference on Computer Communication and Informatics (ICCCI), 2013.

[32] J. X. Yinhui Lia, Silan Zhanga, Jiakai Yana, Xiaochuan Aib, Kuobin Dai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," Expert Systems with Applications, doi:10.1016/j.eswa.2011.07.032, 2011.

[33] E. d. l. Hoz, E. d. l. Hoz, A. Ortiz, J. Ortega, and A. Martínez-Álvarez, "Feature selection by multi-objective optimisation: Application to network anomaly detection by hierarchical self-organising maps," 2014.

[34] F. Amiri, M. R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," Journal of Network and Computer Applications, vol. 34, no. 4, pp. 1184-1199, 2011.

[35] A. R. K.M. Faraoun, "Data dimensionality reduction based on genetic selection of feature subsets," INFOCOMP Journal of Computer Science, 2007.

[36] J. a. A. S. a. P. S. Annand Kannan and Gerald Q. Maguire, "GA based Feature Selection Algorithm for effective Intrusion Detection in Cloud Networks," 12th International Conference on Data Mining Workshops IEEE, 2012.

[37] Y. Wang, Y. Li, Q. Wang, Y. Lv, S. Wang, X. Chen, X. Yu, W. Jiang, and X. Li, "Computational identification of human long intergenic non-coding RNAs using a GA–SVM algorithm," Gene, vol. 533, no. 1, pp. 94-99, 2014.

[38] J.-M. W. Tarek, M. Hamdani, A.M. Alimi, F. Karray, "Hierarchical GA with new evaluation function and bi-coded representation for the selection of features considering their confidence rate," Applied Soft Computing, 2011.

[39] S. Due, and X. Du, Data Mining and Machine Learning in Cybersecurity: Auerbach Publications, 2011.

[40] Z. Y. M. Lin Li Zhong, Zhang Yu Bin, "Netwrok Intrusion Detection Method by Least " Computer Science and Information Technology (ICCSIT), 3rd IEEE International Conference on Page(s): 295 -297 2010.

[41] P. P. M. Gudadhe, "A New Data Mining Based Network Intrusion Detection ", 7th Annual Communication Networks and Services Research Conference, pp. 372-377, 2009.

[42] N. Moustafa, and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." pp. 1-6.

[43] S.-W. Lin, K.-C. Ying, C.-Y. Lee, and Z.-J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," Applied Soft Computing, vol. 12, no. 10, pp. 3285-3290 , 2012.

[44] A. Kannan, G. Q. Maguire, A. Sharma, and P. Schoo, "GA based feature selection algorithm for effective intrusion detection in cloud networks." pp. 416-423.

[45] A. Chandrasekhar, and K. Raghuveer, "Intrusion detection technique by using k-means, fuzzy neural network and SVM classifiers." pp. 1-7.

Hossein Gharaee received B.S. degree in electrical engineering from Khaje Nasir Toosi University, in 1998, M.S. and Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran, in 2000 and 2009 respectively. Since 2009, he has been with the Department of Network Technology in IRAN Telecom Research Center (ITRC). His research interests include general area of VLSI with emphasis on basic logic circuits for low-voltage low-power applications, DSP Algorithm, crypto chip and Intrusion detection and prevention systems.

Maryam Fekri received B.S. degree in Information Technology from Mazandaran University of Science and Technology, in 2012, M.S. degree in Information Securtiy from Tehran University, Iran, in 2015 and currently studying her second M.S. degree in Digital Media in Carleton Unicersity, Canada. Her research interest field is mainly in Data Analysis and Big Data.

Hamid Hosseinvand received B.S. degree in computer engineering from Sepahan Institute of Higher Education, in 2011, M.S. degree in information technology engineering from Shahed University, Tehran, Iran, in 2017. His research interest include network and information security.