International Journal of Digital Crime and Forensics
Volume 13 • Issue 2 • March-April 2021

An Intrusion Detection System Using Modified-Firefly Algorithm in Cloud Environment

Partha Ghosh, Netaji Subhash Engineering College, India
Dipankar Sarkar, Tata Consultancy Service, India
Joy Sharma, Netaji Subhash Engineering College, India
Santanu Phadikar, Maulana Abul Kalam Azad University of Technology, India

ABSTRACT

The present era is being dominated by cloud computing technology which provides services to the users as per demand over the internet. Satisfying the needs of huge people makes the technology prone to activities which come up as a threat. Intrusion detection system (IDS) is an effective method of providing data security to the information stored in the cloud which works by analyzing the network traffic and informs in case of any malicious activities. In order to control high amount of data stored in cloud, data is stored as per relevance leading to distributed computing. To remove redundant data, the authors have implemented data mining process such as feature selection which is used to generate an optimum subset of features from a dataset. In this paper, the proposed IDS provides security working upon the idea of feature selection. The authors have prepared a modified-firefly algorithm which acts as a proficient feature selection method and enables the NSL-KDD dataset to consume less storage space by reducing dimensions as well as less training time with greater classification accuracy.

KEYWORDS

Cloud Computing, Feature Selection, Firefly Algorithm (FA), Intrusion Detection System (IDS), NSL-KDD Dataset, Particle Swarm Optimization (PSO)

INTRODUCTION

Cloud Computing is one of the upcoming technologies which provides software, platforms and infrastructural services as per the requirement of users (Yeboah-Boateng & Essandoh, 2013). Cloud provides service to the people to store and make use of stored materials for their purpose. There is a universal access to the user's data throughout the world using any Internet-ready devices (M. Davis & A. Sedsman, 2010). Cloud computing is basically associated with the Distributed computing. Distributed computing is required in situation where the data is so huge that it cannot be saved in a single storage device, rather data is stored in distributed manner in Cloud Environment. Although with the increasing acceptance of Cloud, the flow of attacks commonly called intrusion to the system is also increasing (Akbar, Dr.K.Nageswara, & Dr.J.A.Chandulal, 2010). Data saved in the Cloud are

DOI: 10.4018/IJDCF.2021030105

highly sensitive and important, so data security is a major concern to protect them from intruder. The malicious attacks affect the properties of storage system such as confidentiality, integrity and availability (A.E. Azzouzi & K.E.E. Kadiri, 2015). IDS is an approach to provide security and shorten the damage of stored data. It provides software and hardware services to put a check on the security, analyses for malicious activity and also produces a report to the management system (Araújo & Abdelouahab, 2012). To detect any kinds of abuse or crime using network, which does not obey the law, IDS is required to build. Network forensic is a method of capturing, storing and analyzing data to find the source of security violator. To detect security violator two types of IDS is there. Host based Intrusion Detection System(HIDS) runs on a particular hosts or machines on the network. HIDS supervises the incoming and outgoing traffic from the machines only and will notify the user or administrator if any unusual action is detected (Partha Ghosh, Ghosh, & Dutta, 2014). Network based Intrusion Detection System(NIDS) are deployed in the crucial points. NIDS watches for abuse of protocols, curious patterns and supervises user actions. It works on the traffic and detects network anomalies (Agrawal & Kamble, 2012). The Network Design should be such that the IDS classifies all types of connections into both normal and abnormal. Artificial Neural Networks can improve the performance and efficiency of IDS (C. Lu, Y. Li, M. Ma & N. Li, 2016). Data mining is the method of exploring patterns from huge data sets involving methods by the combination of machine learning and statistics. As this involves a very large training dataset that increases memory space requirement as well as time. Therefore to reduce the number of features a Feature Selection(FS) algorithm is required which is a Data mining process (P. Ghosh & Mitra, 2015). In this paper, the authors have approached the feature selection using the proposed Modified-Firefly Algorithm (MFA) to reduce the dimension of the dataset. Using the datamining process, after analyzing the network activity and comparing with benchmark signatures network forensics and crimes can be detected. To make the analyzing and detecting process faster and more accurate the author proposed MFA. This Modified Firefly Algorithm is achieved by blending the idea of three different data mining methods. These methods includes PSO, FA and Fuzzy Logics. By exhibiting the results of the experiment it can be concluded that the proposed MFA produces better optimum feature subset than FA. Finally, a number of classifiers have been used namely AdaBoost, Neural Network and Random Forest on this reduced dataset obtained by applying MFA to find the accuracy of the proposed IDS Model. The outcome of these classifiers exhibits improved accuracy.

RELATED WORK

The public Infrastructure-as-a-Service (IaaS) Cloud industry has attained a crucial mass in the last couple of years, with many Cloud Service Providers contributing proficient services. W. Huang et al. discussed the security system provided by public IaaS Cloud and security systems proposed by academia over the same span in their paper (Huang, Ganjali, Kim, Oh, & Lie, 2015). They have also thought deeply and theorized on how industry and academia might handle jointly to solve the crucial security problems in public IaaS Clouds. B. R. Raghunath and S. N. Mahadeo have introduced Network Intrusion Detection System (Raghunath & Mahadeo, 2017). The IDS proposed in that paper utilizes a collection of data mining methods to identify attacks. They presented two specific contributions in their paper, one is the unsupervised anomaly detection techniques that assign scores to the network and the other is the association pattern analysis. H. Liu and R. Setiono defined the datasets largeness in terms of both horizontal largeness and vertical largeness (Liu & Setiono, 1998). S.K. Dash et al. proposed a novel method for classification by tuning the parameters of radial basis function networks using feature selection (Dash, Dash, Dehuri, & Cho, 2013). In this paper, the authors used information gain theory for reducing the features and differential evolution for tuning center and spread of radial basis functions. This approach is validated with a few benchmarking highly skewed and balanced dataset. James Kennedy and Russell Eberhart introduced a new swarm based algorithm known as PSO algorithm (Kennedy & Eberhart, 1995). They tested the algorithm on a number of benchmark

functions for optimization and noted their performance on a number of difficult problems. The unique concept of the above algorithm was found to be the flying potential solutions through the hyperspace to reach towards better solution. Xin-She Yang formulated a new swarm optimization algorithm known as Firefly Algorithm (Yang, 2009). In his work he analyzed the algorithm by comparing with PSO algorithm. Simulating results based on various optimal functions the Firefly Algorithm have displayed improved result in terms of efficiency and convergence in contrast to both PSO and Genetic Algorithm(GA) algorithm. Feature Selection has always been useful in dimensionality reduction of features, when the total dataset is large in terms of both patterns and features. Sankhadeep Chatterjee et al. proposed a modified Cuckoo Search(CS) supported Neural Network(NN-MCS) classifier (Chatterjee, Dey, et al., 2017). In this paper, lévy flight combined with CS has been modified using McCulloch's method of generating stable random numbers. After comparing this method the suggested NN-MCS model shows better result to a greater extent. A. Boucheham and M. Batouche proposed a novel hybrid wrapper/filter feature selection method to label the most descriptive genes for cancer detection (Boucheham & Batouche, 2017). Their proposed model was experimented on nine publicly available cancer DNA microarray datasets which results in selection of potent signatures with better classification accuracy. Hema Banati and Monika Bajaj devised an algorithm to be applied on the medical domain for finding minimal attribute set without compromising with the classification efficiency of the feature subset (Banati & Bajaj, 2011). Experimental results showed that their proposed Firefly Feature Selection algorithm produced better results as compared to other soft-computing techniques. That algorithm consumed less time to find the optimal subset having the same efficiency as that of PSO and Bee. Dharmpal Singh in his paper, proposed a modified form of BAT algorithm based on natural echolocation behaviour of bats to clarify the optimization problems (Singh, 2018). Modified BAT algorithm performed better than other algorithms in term of efficiency and robustness. Farzaneh Hosseini and Marjan Kaedi developed a nature-inspired metaheuristic algorithm named Sun and Leaf Optimization (SLO) (Hosseini & Kaedi, 2018). This algorithm is encouraged by the effect of sunlight on the leaves germination. Leaves grown on the tree are considered as the candidate solutions in the state space. The greener leaves on the direction of sunlight are high quality solutions. The wind effect is also considered here to escape the local optima. A. Gurav et al. proposed heuristic approach with Glowworm that improves the performance by reducing number of iteration for convergence (Gurav, Nair, Gupta, & Valadi, 2015). That feature selection algorithm can be implemented using Glowworm swarm optimization algorithm that improves the predicting power of classification which involves large datasets. Samb et al. in their paper, proposed feature selection that was associated with local search operator to enhance the quality of classifier (Samb, Camara, Ndiaye, Slimani, & Amir Esseghir, 2012). After the dataset has been reduced, a classifier is needed to classify them. J. K. Basu et al. used ANN in Pattern Recognition for classification(Basu, Bhattacharyya, & Kim, 2010). That algorithm has the capability to find complex nonlinear input-output relationships, sequential training process and quick adaptability to data were used. Sankhadeep Chatterjee et al. employed a Multi-objective-GA(MOGA) to train the NN based model. In their paper, the NN has been trained with MOGA to minimize the Root Mean Squared Error(RMSE) and Maximum Error(ME) toward optimizing the weight vector of the NN (Chatterjee, Sarkar, et al., 2017). Dana Balas-Timar et al. suggest that interviews of employee selection process can provide a high level of criterion-related validity when properly designed (Balas-timar, Balas, Breaz, Dey, & Ashour, 2016). In their paper, the authors offered a fuzzy perspective to deal with uncertainties that currently appear in human resources decision making processes, associated in particular with scoring competency based behavioural interviews. Using Apriori algorithm Yi-Chung Hu et al. proposed a data mining technique to explore fuzzy classification rule based (Hu, Chen, & Tzeng, 2003). In their paper they also applied the concept of Genetic Algorithm. For effective classification M. A. Jabbar et al. proposed a new algorithm which combines K-Nearest Neighbor with Genetic Algorithm (Jabbar, Deekshatulu, & Chandra, 2013). Their proposed algorithm produced good results in medical databases. Using the

inference of the above mentioned papers authors have proposed a Modified-Firefly Algorithm for selecting features from NSL-KDD dataset to produce an efficient IDS.

PRELIMINARY STUDY

The excellence of classification of dataset rely upon the approach taken at the time of training of the classifier. The dataset sometimes becomes very large either because of the count of features as well as count of records present in it. This factor increases the cost of processing the data and memory requirement both becomes an overhead. Reducing the dimensionality by selecting a subset from the original feature set are advantageous for the encountered problems (Guyon & Elisseeff, 2003). FS is a method of shortening the dimension of the feature set. FA is a swarm based method that converges quickly and is equally as efficient as another swarm based method. Another easy swarm based method is PSO which has been used in a number of applications. Fuzzy membership is a set of If-Then rules used for fuzzification of certain fitness functions. All the three algorithms that is Firefly, PSO and Fuzzy can be used in feature selection process for obtaining optimal feature subset. In this section the discussion is about the FA, PSO and Fuzzy membership.

Firefly Algorithm (FA)

In 2008 Xin-She Yang proposed Firefly Algorithm at Cambridge University, which was based on the flashing patterns and behavior of fireflies (Fister, Jr Fister, Yang, & Brest, 2013). The fireflies are attracted to one another depending on the light intensity. The changes of attractiveness β with distance r is given by:

$$\beta = \beta_0 e^{-\gamma r^2} \tag{1}$$

where, β_0 is the attractiveness at $r = 0$ and γ is a constant. The Euclidian distance r_{ij} is given by the equation below:

$$r_{ij} = \sqrt{\sum_{k=1}^{d} \left(x_{ik} - x_{jk}\right)^2} \tag{2}$$

The progress of one firefly namely i towards another firefly j is driven by the equation given below:

$$x_i = x_i + \beta_0 e^{-\gamma r^2} \left(x_j - x_i\right) + \alpha \left(rand - \frac{1}{2}\right) \tag{3}$$

where, $\beta_0 e^{-\gamma r^2} \left(x_j - x_i\right)$ representing the attraction between the fireflies i and j. The randomization is done with the randomization parameter α_t and it can be tuned from iteration to iteration.Being a Swarm Intelligence based method it has all the benefits of other such algorithms but with two more advantages which are automatically subdivision and the capability of handling with multimodality. This makes the entire population to get subdivided into groups in which the fireflies can swarm around the local optimum (Yang & He, 2013). Among all this the global maximum is found. FA is both a stochastic and meta-heuristic algorithm (Francisco, Costa, & Rocha, 2014).

Particle Swarm Optimization(PSO)

PSO is a biologically influenced computational search and optimization method developed in 1995 by Eberhart and Kennedy based on the communal behaviors of birds flocking (Rini, Shamsuddin, & Yuhaniz, 2011). The algorithm works as in there are a number of particles say flocking birds that has no leader. They go around searching for food randomly and the bird that is closer to the food guides the others by exchanging information. Other birds change their position according to the instruction and like this they finally reach their food source which is termed as the global optima. Exploring is the ability to identify a good optima. Exploitation is the capability of concentrating the search space to a particular area only. The position of a particle is updated by adding velocity to it using equation given below:

$$x_i\left(t+1\right) = x_i\left(t\right) + v_i\left(t+1\right) \tag{4}$$

whereas the velocity is updated using equation given below.

$$v_i\left(t\right) = v_i\left(t-1\right) + c_1 r_1\left(localbest\left(t\right) - x_i\left(t-1\right)\right) + c_2 r_2(globalbest\left(t\right) - x_i\left(t-1\right)) \tag{5}$$

where c_1, c_2 are the acceleration constants and r_1, r_2 are the randomly generated constants. These parameters are used to tune the convergence rate of the algorithm. As the particle reach local optima very quickly, the swarm converges prematurely to the local minima. This phenomena of PSO has been controlled by presenting a number of controlling parameters. Velocity clamping is one of the methods of controlling the global exploration of the particle. The velocity of a particle in the dimension is clamped at the maximum allowable speed limit j . Velocity is adjusted according to the formula given below:

$$v_{ij} = \begin{cases} v_{ij}\left(t+1\right), & if \quad v_{ij}\left(t+1\right) < v_{max}\left(j\right) \\ v_{max}\left(j\right) & otherwise \end{cases} \tag{6}$$

PSO can be used for feature reduction mechanism but in that case the updating needs to be done in terms of 0 and 1. So, Binary PSO is required in which the particle represents position in binary space along with a classifier for computing the fitness values of the particle in each iteration (Tu, Chuang, Chang, & Yang, 2007). Each particle keeps trail of its coordinates in the problem space and associates them with the best solution it has obtained so far. The PSO algorithm has been used extensively for optimization problems and because of its efficiency if can also be used for feature selection process.

Fuzzy Membership

Fuzzy is a concept where the membership value of an element is measured rather than considering a single label. Mathematically a Fuzzy Set is one in which an object exists in any set with a magnitude that is between 0 and 1.Fuzzy Set Theory is an extension of dual logic or the classical set theory. Taking the contribution of a set of elements towards a particular goal rather than a single influence is the key concept of fuzzy membership. A fuzzy function is a generalization of the concept of classical function (Zimmermann, 2010).In Fuzzy Logic, it represents the degree of truth as an extension of valuation. Degrees of truth are often baffled with probabilities, although they are conceptually different (Kosko, 1990). In a Fuzzy Classifier System the classification is achieved by a number of Fuzzy If=Then rules to represent each feature (Rezaee, Goedhart, Lelieveldt, & Reiber, 1999). Fuzzy

logic is suitable for managing imprecise data. Whenever they are present, Fuzzy logic is used to find the relevant features so that the loses in information from real process could be minimized (Grande, del Rosario Suárez, & Ramó Villar, 2007). Some of the most commonly used Fuzzy Membership Functions are Triangles, Trapezoidal, Gaussian etc. Fuzzy logic can be used in a number of ways especially where normal logic of single labelling doesn't sufficient. These reasons lead the authors to apply Fuzzy Concept in their proposed model to overcome the monotonic distribution.

PROPOSED MODEL

Cloud computing is a computing paradigm, where a huge number of systems are associated in private or public networks, to provide dynamically scalable infrastructure for application, data and file storage. With the advancement of this technology, the cost of computation, application hosting, content storage and delivery is minimized abruptly. The data stored in Cloud Environment is in distributed manner as the data is too large to store in a single storage device. Users often want to store their personal information and secure data in Cloud. Therefore Data Security is a major concern leading to the concern in security of Cloud. With the rapid development of network technology computer crime grows exponentially. Security of computers are badly affected by the criminals. Network security technology is not enough for completely eliminating computer crimes. Some legal punishment and discouragement power should be added with that. So, network forensics means an active defense in network security aspect. IDS is the process of identifying unauthorized use, misuse, and abuse of systems. It is used to check both the authorization of use and the authentication of the usage. It monitors the inbound and outbound packets and alters the host in case of malicious activities. Since the dataset is very large in term of horizontal redundancy, the authors have an overhead of large computation; reducing the accuracy of classification. So Data Mining is introduced to extract or mine the relevant information from the dataset. Feature selection is the method by which the number of features is reduced as a result of which the noise, redundancy of data is reduced and in turn the classification rate increases. The system should also be intelligent enough to automatically learn to improve performance which is the essence of Machine Learning. In this paper a Modified-Firefly algorithm is proposed for choosing the best feature subset. The subset obtained is then sent to be classified using different classifiers to compute the classification accuracy. Requests made by the clients are fulfilled by the Cloud services. So, special security has to be deployed to secure the system from intruders.

Figure 1 shows the proposed model that can maintain safe and fast accomplishment of the task achieved by the Cloud user. In this model, Network Design is such that the Network Intrusion Detection System(NIDS) is placed at the chokepoint of the network.

In this proposed model an IDS has been designed that reduces the dimensionality of features in the dataset by feature selection algorithm. The proposed MFA is applied to extract the relevant feature set from the large dataset. Based on the optimal feature subset obtained, the authors classify the dataset using numerous classification algorithm. The flowchart of our proposed IDS has been shown in Figure 2 and the detail steps of our proposed Modified-Firefly algorithm has been shown in Figure 3. The main objective of authors is to build an efficient IDS, but the largeness of the dataset increases both memory space and processing power of the system therefore an efficient feature selection process has been designed using this proposed Modified-Firefly Algorithm which reduces the number of feature without compromising with the accuracy of classification. In the proposed model the FA has been modified by first introducing two constant c_1 and c_2. Both c_1 and c_2 are known as acceleration constants. The constant β_0 of FA has been replaced. In each iteration the fireflies are randomly attracted towards the global best (gbest) position in the entire population. In this manner the position vector of FA has also been changed since the authors are not only computing the Euclidian distance of x_i and local best (pbest), but also x_i and global best (gbest). The FA has been modified to increase

Figure 1. Intrusion detection system in cloud environment

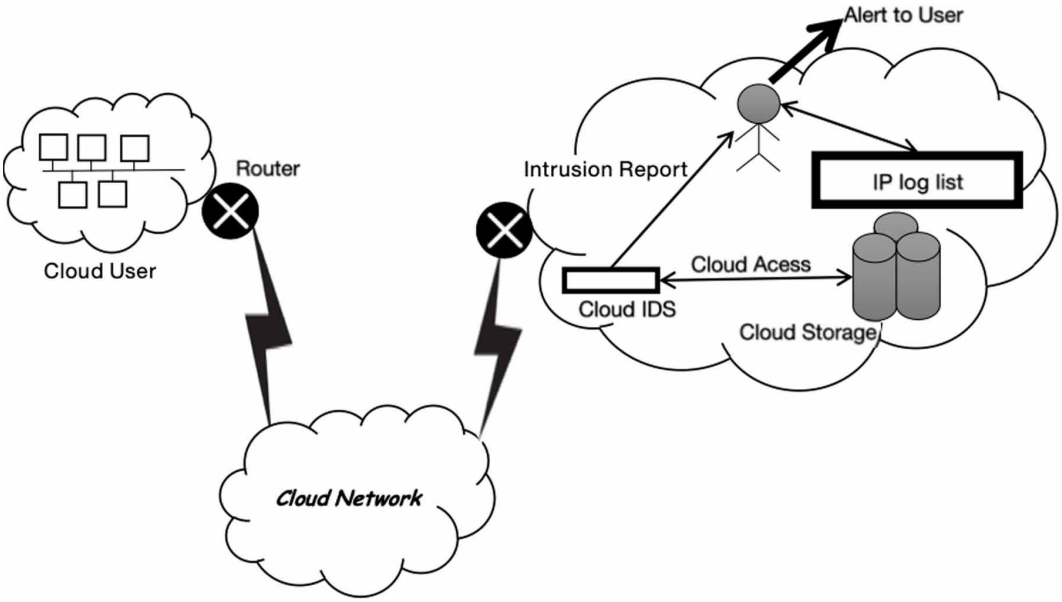


Figure 2. Flowchart of our proposed IDS

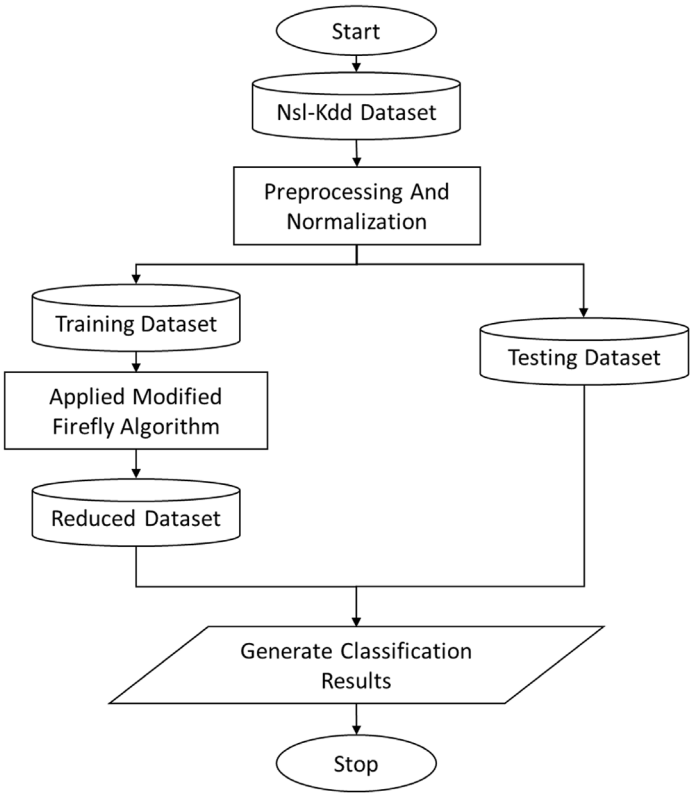
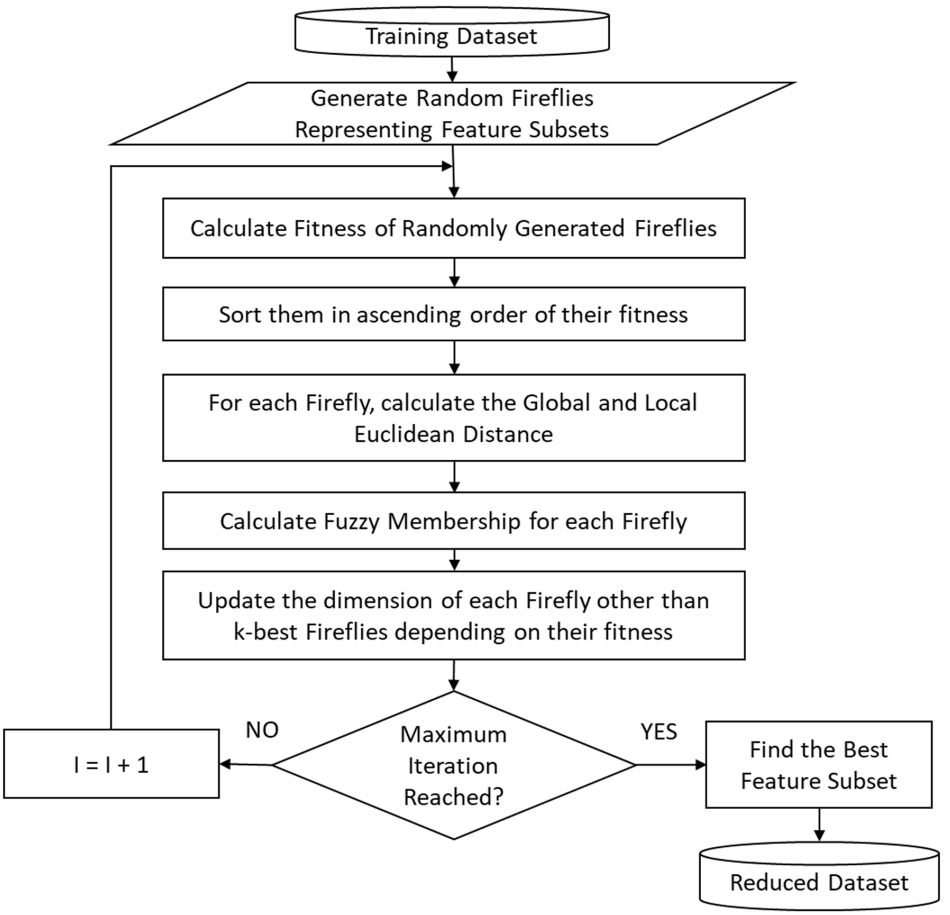


Figure 3. Detail steps of proposed modified-firefly algorithm



the convergence and also enhance its capability for not falling into the local minimum. The distance between and , are found using the equation no. 7 and 8 respectively.

$$r_{px} = \sqrt{\sum_{k=1}^{d} \left(pbest_{ij} - x_{ij} \right)^2} \tag{7}$$

$$r_{gx} = \sqrt{\sum_{k=1}^{d} \left(gbest_{ij} - x_{ij} \right)^2} \tag{8}$$

In Modified-Firefly Algorithm, Fuzzy membership function is introduced in the position vector updation. As in the normal Firefly Algorithm there is a weakness i.e. fireflies move despite of the global optima, which can enhance the convergence time to obtain the global best. To dispose of the weakness of the standard FA and to enhance the collective movement of fireflies, FA is modified in which more than one fireflies can influence movements of others in each iteration. In every iteration the authors take the k-best fireflies, according to their fitness value, as the standard for the movement

of other fireflies in that iteration. In that iteration, the k-best fireflies are kept unchanged and the others are influenced by the k-best. As the movement of fireflies are influenced by more than one fireflies, the contribution of each firefly has been divided among the fireflies. For this purpose, the fuzzy concept comes to place. According to the contribution calculated by the fuzzy concept, the fireflies move towards the k-best. The level of attractiveness of each k-best fireflies is represented as a Fuzzy variable μ. The k-brighter fireflies in each iteration are chosen to be candidates, where k is a user-set parameter. We compute $\mu\left(h\right)$ that is the attractiveness of firefly h, as given below.

$$\mu\left(h\right)=\left(\frac{1}{f\left(h\right)-f\left(p\right)}\right)/\beta \tag{9}$$

Where, h be one of the k-brighter fireflies in each iteration. $f\left(p\right)$ depends on the global optima in each iteration and $f\left(h\right)$ depends on the fitness function of all the k-fireflies. To escape the dependency on the scale of the fitness function, equation no. 10 is used:

$$\beta = f\left(p\right)/l \tag{10}$$

where, l is a user-specified parameter. For a fixed $f\left(h\right)$, the larger the value of l, smaller the attractiveness $\mu\left(h\right)$. The progress of one firefly i towards another firefly j is driven by the modified position vector given by the equation number 11:

$$x_{id}^{t+1}=x_{id}^{r}+\left(c_1*e^{-\gamma r^2}\left(x_{jd}^{t}-x_{id}^{t}\right)+c_1*e^{-\gamma r^2}\left(x_{gd}^{t}-x_{id}^{t}\right)+\sum_{h=1}^{k}\mu\left(h\right)e^{-\gamma r^2}\left(x_{hd}^{t}-x_{id}^{t}\right)\right)*a\left(rand-\frac{1}{2}\right) \tag{11}$$

Once the firefly gets updated on its position, the value of x_{id} may be a real value.But there is a need to transform the real value into binary i.e. 0 and 1 as because the selected features are to be extracted. So a probabilistic rule is needed, based on a hyperbolic tangent sigmoid transfer function, is applied to each dimension of the position vector. The rule is given below.

$$x_{id}=\begin{cases}1, & if\ \ rand<S\left(x_{id}\right)\\ 0, & otherwise\end{cases} \tag{12}$$

where:

$$S\left(x_{id}\right)=\tanh\left(\left|x_{id}\right|\right)=\frac{\exp\left(2*\left|x_{id}\right|\right)-1}{\exp\left(2*\left|x_{id}\right|\right)+1} \tag{13}$$

This way the process continues updating the fireflies and the k-best fireflies is taken out every time. Finally, after the maximum iteration is reached the firefly with the best fitness value is opted as the optimal subset. The experiment with our proposed model showed that it has performed better

than the FA for the work of feature selection in terms of both decreasing the processing time and the memory requirement.

Modified-Firefly Algorithm

```
Input: NSL-KDD Training Dataset.
Output: Best Feature Subset.
Randomly generate initial population of fireflies;
While iteration <=Maximum no. of iteration do
Find the fitness of all fireflies;
     Sort the population based on fitness value;
     for each fireflies other than k-best fireflies do
          for each fireflies do
               Calculate the local and global distance using
equation no. 7 & 8;
               Calculate fuzzy membership using equation no. 9;
               if  I(i)<I(j) then
                    Firefly i is moved towards firefly j using
equation no. 11;
               end
          end
     end
     Generate new population;
     iteration = iteration + 1;
end
```

Best Feature Subset is extracted by the selected features of the best firefly.

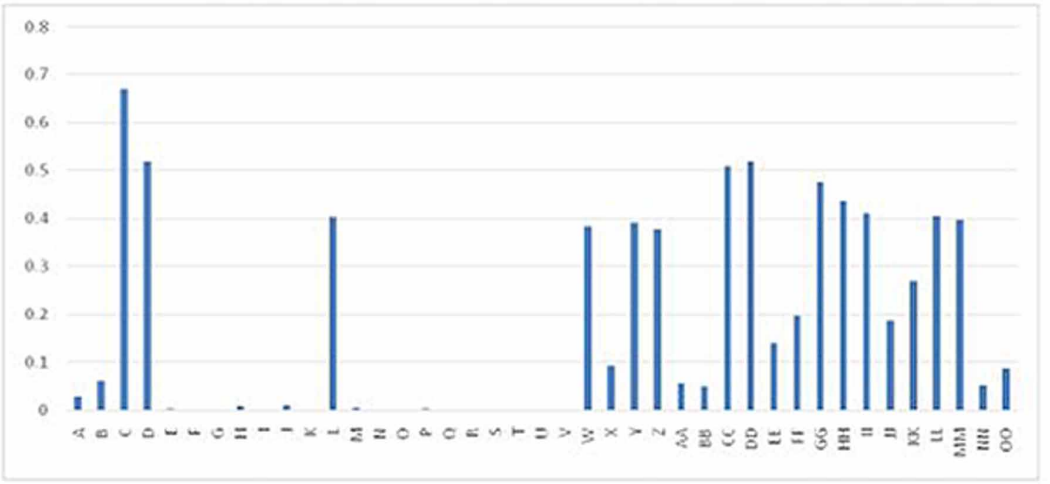
As discussed earlier the algorithm shows the process of the proposed Modified-Firefly approach for Feature Selection. In the first phase, the relevant feature subset is extracted successfully by the proposed method. In the next phase, the obtained feature subset is then passed to the classifiers like Artificial Neural Network, Adaboost and Random Forest to classify the dataset to get better accuracy. Hence, the authors have been successful in incorporating effective memory usage & reduced processing time by reducing the dimension of the dataset.

PERFORMANCE

An efficient IDS is required for detecting all the attacks to secure the Cloud environment. An appropriate training is required to detect all attacks from the normal behavior. The presence of huge number of irrelevant and redundant features in the training dataset cause much more training time and also decreases the classification accuracy. Here in this paper, authors have used NSL-KDD-Train+ dataset for training purpose. There are four components of NSL-KDD dataset namely, KDD Train+, KDD Test+, 20%KDDtraining+ and KDDtest-21.

Our experiment has made use of KDD Train+ and KDD Test+ consisting of 125,973 and 22,544 records respectively. Each of the components consists of 41 features labeled as normal or the specific attack type. KDD Train+ has been used to train the model whereas KDD Test+ has been used to measure the classification accuracy. Here Information Gain (IG) is usedto rank all the features. For predicting the class, Information Gain (IG) measures the amount of information in bits. Figure 4 shows the Information Gain of all the features presents in NSL-KDD Train+ dataset. Here it is found that there are a descent no. of the features that have Information Gain so poor that they do not contribute

Figure 4. Information gain of all the features exist in NSL-KDD-Train+ dataset



in the training at all (IG average = 0.175246). So these features can be ignored for training without compromising with the classification accuracy. This is the main reason for reducing the dataset to produce better accuracy and less processing time. Using the proposed method, a number of irrelevant features are removed from the NSL-KDD-Train+ dataset and a minimized training set is produced. The experiment was performed on a hypervisor using 8 virtual CPU (vCPU), 16 GB of RAM and CentOS 7 Operating System. The results obtained by evaluating the outputs proved to be reliable and successful in differentiating normal and anomalous behaviors.

In order to evaluate NIDS and measure the performance, there are standard metrics that have been used here:

- True Positive(TP)
- True Negative(TN)
- False Positive(FP)
- False Negative(FN)
- True Positive Rate(TPR)
- True Negative Rate(TNR)
- False Positive Rate(FPR)
- False Negative Rate(FNR)
- Accuracy = ((TP+TN)/(TP+FP+FN+TN))*100%

The tables shown above represents the accuracy evaluated by using Neural Network on different number of fireflies, in most of the cases it has been observed that the proposed algorithm perform better accuracy than the others. In most of the cases it is also notable that the True Positive Rate is significantly higher. Figure 5 and Figure 6 shows result of reduced

dataset obtained by 50 fireflies that has been tested on different classifiers like NN, Adaboost, Random Forest. Figure 7 shows reduced dataset obtained with 40 fireflies and tested on Neural Network classifier. Figure 8 shows reduced dataset obtained with 20 fireflies and tested on Neural Network classifier.

The output of the proposed MFA are applied in various classifiers and have obtained higher TPR as shown in the performance matrices. After analyzing the performance matrices it is seen that if the proposed MFA is applied for feature selection, it will provide a potent feature subset. The

Table 1. Classification result by using Neutral Network with 50 fireflies

	NSL-KDD Dataset	Firefly	Modified-Firefly with k=2	Modified-Firefly With k = 3	Modified-Firefly With k = 4
TP	7601	7755	7635	7993	8311
TN	9499	9427	9443	9454	9460
FP	212	284	268	257	251
FN	5232	5078	5198	4840	4522
TPR	0.5923	0.6043	0.59495	0.6228	0.6476
TNR	0.97817	0.97075	0.9724	0.9735	0.9742
FPR	0.02183	0.02925	0.0276	0.0265	0.0258
FNR	0.4077	0.3957	0.40505	0.3772	0.3524
Accuracy(%)	75.8517	76.2154	75.7541	77.3909	78.828

Table 2. Classification result by using Neural Network with 40 fireflies

	NSL-KDD Dataset	Firefly	Modified-Firefly with k=2	Modified-Firefly with k=3	Modified-Firefly with k=4
TP	7601	7393	7807	7899	8488
TN	9499	9411	9422	9425	9459
FP	212	300	289	286	252
FN	5232	5440	5026	4934	4345
TPR	0.5923	0.57609	0.60835	0.61552	0.661419
TNR	0.97817	0.96911	0.97024	0.97054	0.97405
FPR	0.02183	0.03089	0.02976	0.02946	0.02595
FNR	0.4077	0.42391	0.39165	0.38448	0.338581
Accuracy(%)	75.8517	74.5387	76.4239	76.8453	79.6088

performance of the proposed model depicts the authors' success in constructing an efficient IDS by reducing the dimension of the dataset that has been used in a Cloud Environment to provide security effectively with a promising accuracy.

CONCLUSION

As time progresses, the effort to provide security in a Cloud Environment also increases due to the increase of doing computer crime by the users as well as the intruders. The data security of the users is the main concern in this spectra as well as in the Distributed System also. So, a cost effective and accurate system is necessary to provide the security. The largeness of the dataset is one of the main consequences towards developing a cost effective system in term of efficiency without compromising with its accuracy or performance. In this paper the authors have concentrated their work only in feature selection. But it can be extended by reducing the dataset using instance selection methods. If horizontal and vertical reduction techniques are used simultaneously the performance may get better.

Table 3. Classification result by using Neutral Network with 20 fireflies

	NSL-KDD Dataset	Firefly	Modified-Firefly with k=2	Modified-Firefly with k=3	Modified-Firefly with k=4
TP	7601	7791	8091	7545	7876
TN	9499	9432	9428	9486	9470
FP	212	279	283	225	241
FN	5232	5042	4742	5288	4957
TPR	0.5923	0.607106	0.63048	0.58794	0.61373
TNR	0.97817	0.9712697	0.97086	0.97683	0.97518
FPR	0.02183	0.0287303	0.02317	0.02317	0.02482
FNR	0.4077	0.392894	0.36952	0.41206	0.38627
Accuracy(%)	75.8517	76.3973	77.71	75.546	76.943

Figure 5. Accuracy in different values of k using different classifiers (using 50 fireflies)

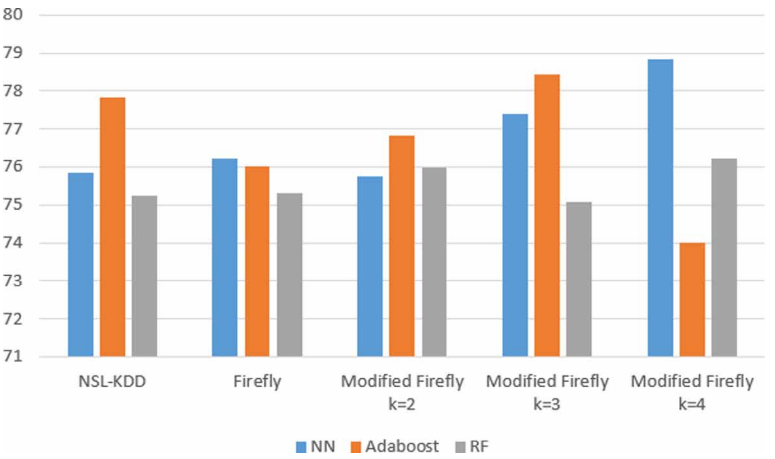


Figure 6. Accuracy in different classifiers using different values of k (using 50 fireflies)

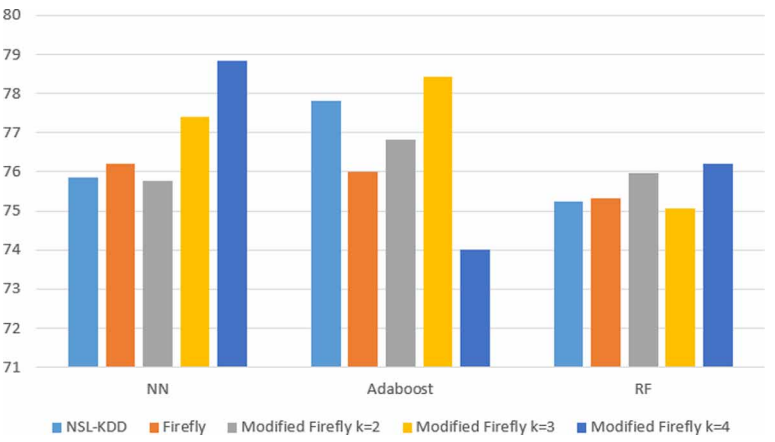


Figure 7. Accuracy in different values of k using NN classifiers (using 40 fireflies)

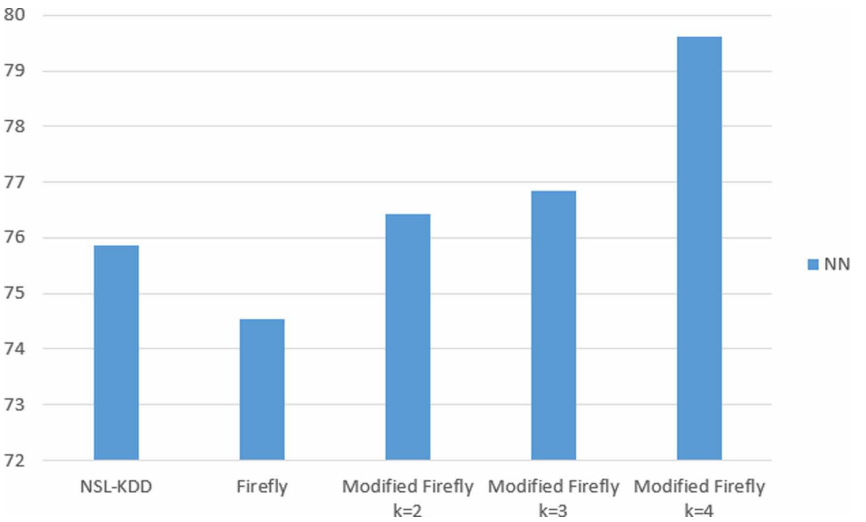
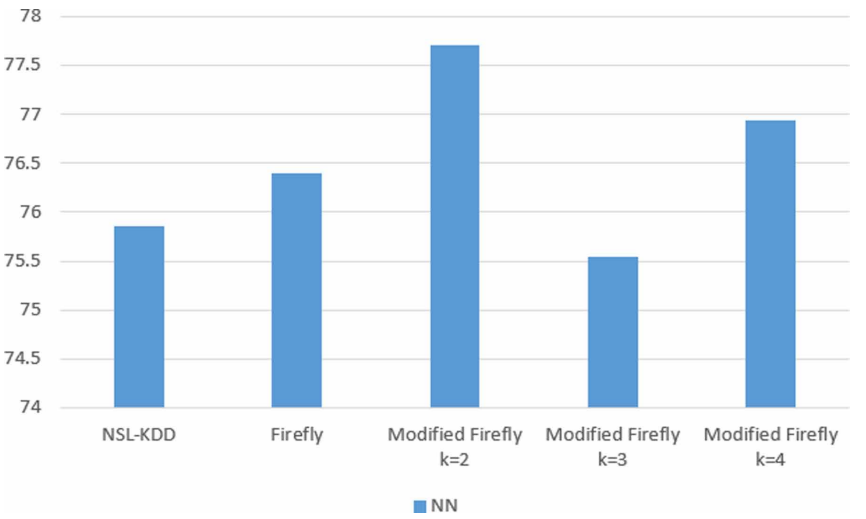


Figure 8. Accuracy in different values of k using NN classifiers (using 20 fireflies)



The Modified-Firefly Algorithm proposed in this paper, has shown better outcomes in accuracy of classification and reduction in the horizontal dimension of the dataset using the Data Mining concept i.e. Feature Selection. The authors have extracted relevant features successfully from the dataset to classify them with the classifiers. The classification results have shown that the accuracy obtained by the classifier is uncompromised. Thus, an effective Cloud Environment can be made to detect attacks by the intruders using the IDS discussed in this paper; leading to more secure system with the improved accuracy and efficiency.

REFERENCES

Agrawal, G., & Kamble, M. (2012). Proposed Multi-Layers Intrusion Detection System (MLIDS) Model. *International Journal of Computer Science and Information Technologies*, *3*(5), 5040–5042.

Akbar, S., Rao, D. K. N., & Chandulal, D. J. A. (2010). Intrusion Detection System Methodologies Based on Data Analysis. *International Journal of Computers and Applications*, *5*(2), 10–20. doi:10.5120/892-1266

Araújo, J. D., & Abdelouahab, Z. (2012). Virtualization in Intrusion Detection Systems : A Study on Different Approaches for Cloud Computing Environments. *IJCSNS International Journal of Computer Science and Network Security*, *12*(11), 9–16.

Azzouzi, A. E., & Kadiri, K. E. E. (2015). Semantic System for Attacks and Intrusions Detection. *International Journal of Digital Crime and Forensics*, *7*(4), 19–32. doi:10.4018/IJDCF.2015100102

Balas-timar, D. V., Balas, V. E., Breaz, A. M., Dey, N., & Ashour, A. S. (2016). Technique for scoring competency based behavioural interviews : a fuzzy. In *9th International Conference on Modern Research In Psychology November 2016* (pp. 105–115). doi:10.5682/9786062805173

Banati, H., & Bajaj, M. (2011). Fire Fly Based Feature Selection Approach. *International Journal of Computer Science Issues*, *8*(4), 473–480.

Basu, J. K., Bhattacharyya, D., & Kim, T. (2010). Use of Artificial Neural Network in Pattern Recognition. *International Journal of Software Engineering and Its Applications*, *4*(2), 23–34.

Boucheham, A., & Batouche, M. (2017). Hybrid Wrapper/Filter Gene Selection Using an Ensemble of Classifiers and PSO Algorithm. *International Journal of Applied Metaheuristic Computing, IGI Global*, *8*(2), 22–37. doi:10.4018/IJAMC.2017040102

Chatterjee, S., Dey, N., Sen, S., Ashour, A. S., Fong, S. J., & Shi, F. (2017). Modified Cuckoo Search based Neural Networks for Forest Types Classification. In *2nd International Conference on Information Technology and Intelligent Trnportation System (ITITS 2017)* (Vol. 296, pp. 490–498). doi:10.3233/978-1-61499-785-6-490

Chatterjee, S., Sarkar, S., Hore, S., Dey, N., Ashour, A. S., Shi, F., & Le, D.-N. (2017). Structural Failure Classification for Reinforced Concrete Buildings Using Trained Neural Network based Multi-objective genetic algorithm. *Structural Engineering and Mechanics*, *63*(4). Advance online publication. doi:10.12989/sem.2017.63.4.000

Dash, S. K., Dash, A. P., Dehuri, S., & Cho, S. (2013). Feature Selection for Designing a Novel Differential Evolution Trained Radial Basis Function Network for Classification. *International Journal of Applied Metaheuristic Computing, IGI Global*, *4*(1), 32–49. doi:10.4018/jamc.2013010103

Davis, M., & Sedsman, A. (2010). Grey Areas - The Legal Dimensions of Cloud Computing. *International Journal of Digital Crime and Forensics*, *2*(1), 30–39. doi:10.4018/jdcf.2010010103

Fister, I., Jr Fister, I., Yang, X.-S., & Brest, J. (2013). A comprehensive review of firefly algorithms. *Swarm and Evolutionary Computation, Elsevier*, *13*, 34–46. doi:10.1016/j.swevo.2013.06.001

Francisco, R. B., Costa, M. F. P., & Rocha, A. M. A. C. (2014). Experiments with firefly algorithm. In ICCSA Springer International Publishing Switzerland, 2014 (pp. 227–236). doi:10.1007/978-3-319-09129-7_17

Ghosh, P., Ghosh, R., & Dutta, R. (2014). An Alternative Model Of Virtualization Based Intrusion Detection System In Cloud Computing. *International Journal of Scientific & Technology Research*, *3*(5), 199–203.

Ghosh, P., & Mitra, R. (2015). Proposed GA-BFSS and logistic regression based intrusion detection system. *Proceedings of the 2015 3rd International Conference on Computer, Communication, Control and Information Technology, C3IT 2015*. doi:10.1109/C3IT.2015.7060117

Grande, J., del Rosario Suárez, M., & Ramó Villar, J. (2007). A Feature Selection Method Using a Fuzzy Mutual Information Measure. Springer-Verlag Berlin Heidelberg. doi:10.1007/978-3-540-74972-1_9

Gurav, A., Nair, V., Gupta, U., & Valadi, J. (2015). Glowworm swarm based informative attribute selection using support vector machines for simultaneous feature selection and classification. In SEMCCO 2015 (pp. 27–37). Springer. doi:10.1007/978-3-319-20294-5_3

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*(3), 1157–1182. doi:10.1016/j.aca.2011.07.027

Hosseini, F., & Kaedi, M. (2018). A Metaheuristic Optimization Algorithm Inspired by the Effect of Sunlight on the Leaf Germination. *International Journal of Applied Metaheuristic Computing, IGI Global*, *9*(1), 40–48. doi:10.4018/IJAMC.2018010103

Hu, Y.-C., Chen, R.-S., & Tzeng, G.-H. (2003). Finding fuzzy classification rules using data mining techniques. *Pattern Recognition Letters, Elsevier*, *24*(1-3), 509–519. doi:10.1016/S0167-8655(02)00273-8

Huang, W., Ganjali, A., Kim, B. H., Oh, S., & Lie, D. (2015). The State of Public Infrastructure-as-a-Service Cloud Security. *ACM Computing Surveys*, *47*(4), 1–31. doi:10.1145/2767181

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. Procedia Technology, 85–94. doi:10.1016/j.protcy.2013.12.340

Kennedy, J., & Eberhart, R. (1995). Particle Swarm Optimization. In *International Conference on Neural Networks* (pp. 1942–1948). IEEE.

Kosko, B. (1990). Fuzziness Vs Probability. *International Journal of General Systems*, *17*(2-3), 211–240. doi:10.1080/03081079008935108

Liu, H., & Setiono, R. (1998). Incremental Feature Selection. *Applied Intelligence*, *9*(3), 217–230. doi:10.1023/A:1008363719778

Lu, C., Li, Y., Ma, M., & Li, N. (2016). A Hybrid NIDS Model Using Artificial Neural Network and D-S Evidence. *International Journal of Digital Crime and Forensics*, *8*(1), 37–50. doi:10.4018/IJDCF.2016010103

Raghunath, B. R., & Mahadeo, S. N. (2017). Network Intrusion Detection System (NIDS). In *First International Conference on Emerging Trends in Engineering and Technology Network* (pp. 1272–1277). doi:10.1109/ICETET.2008.252

Rezaee, R. M., Goedhart, B., Lelieveldt, B. P. F., & Reiber, J. H. C. (1999). Fuzzy feature selection. *Pattern Recognition, Elsevier*, *32*(12), 2011–2019. doi:10.1016/S0031-3203(99)00005-9

Rini, D. P., Shamsuddin, S. M., & Yuhaniz, S. S. (2011). Particle Swarm Optimization : Technique, System and Challenges. *International Journal of Computers and Applications*, *14*(1), 19–27. doi:10.5120/1810-2331

Samb, M. L., Camara, F., Ndiaye, S., Slimani, Y., & Amir Esseghir, M. (2012). A Novel RFE-SVM-based Feature Selection Approach for Classification. *International Journal of Advanced Science and Technology*, *43*, 27–36.

Singh, D. (2018). A Modified Bio Inspired: BAT Algorithm. *International Journal of Applied Metaheuristic Computing, IGI Global*, *9*(1), 60–77. doi:10.4018/IJAMC.2018010105

Tu, C., Chuang, L., Chang, J., & Yang, C. (2007). Feature Selection using PSO-SVM. *IAENG International Journal of Computer Science*, *33*(1).

Yang, X. S. (2009). Firefly algorithms for multimodal optimization. In Springer-Verlag Berlin Heidelberg 2009 (pp. 169–178). doi:10.1007/978-3-642-04944-6_14

Yang, X. S., & He, X. (2013). Firefly Algorithm: Recent Advances and Applications. *International Journal of Swarm Intelligence*, *1*(1), 1–14. doi:10.1504/IJSI.2013.055801

Yeboah-Boateng, E. O., & Essandoh, K. A. (2013). Cloud Computing : The Level of Awareness amongst Small & Medium-sized Enterprises (SMEs) in Developing Economies. *Journal of Emerging Trends in Computing and Information Sciences*, *4*(11), 832–839. doi:10.1007/s10796-013-9450-9

Zimmermann, H. J. (2010). Fuzzy set theory. In Fuzzy Set Theory-and Its Applications (pp. 317–332). doi:10.1002/wics.82

Partha Ghosh is an Assistant Professor of Information Technology at Netaji Subhash Engineering College, Maulana Abul Kalam Azad University of Technology, Kolkata, West Bengal, India. He has done M.Tech. in Computer Science and Engineering from Calcutta University in 2003. He has published more than 15 Research papers in reputed conferences and journals. His research interests include Cloud Computing, Machine Learning, Intrusion Detection System, Computer Networks and Security.

Dipankar Sarkar is a Computer Science and Engineering Graduate from Netaji Subhash Engineering College, MaulanaAbulKalam Azad University of Technology, Kolkata, West Bengal, India. He is currently working at TCS. Areas of interest and experience include Machine Learning and Deep NLP. He has worked on developing intelligent Intrusion Detection systems during his bachelors' studies.

Joy Sharma is a Computer Science and Engineering Graduate from Netaji Subhash Engineering College, MaulanaAbulKalam Azad University of Technology, Kolkata, West Bengal, India. He is currently working in TCS as a Java Developer. His area of interest includes Data Mining, Machine Learning, Artificial Intelligence, Cloud Computing.

Santanu Phadikar is an Associate Professor and Head of the department of Computer Science and Engineering at Maulana Abul Kalam Azad University of Technology, Kolkata, West Bengal, India. He has done M.Tech. in Computer Science and Engineering from Calcutta University in 2003. He pursued his Ph.D. from Indian Institute of Engineering Science and Technology, Shibpur, West Bengal, India in 2013. His research area includes Machine Learning, Intrusion Detection System, Soft Computing and Cloud Computing.